Contributions to Signal Detection, Network Analysis, and Clustering

Nicolas Verzelen

Université de Montpellier

École doctorale Information, Structures et Systèmes (I2S)
Spécialité: Biostatistique

Manuscrit pour l'obtention de

l'HABILITATION À DIRIGER DES RECHERCHES

Contributions to Signal Detection, Network Analysis, and Clustering

Contributions à la détection de signal, l'analyse de réseaux et au clustering

Nicolas Verzelen

Rapporteurs:

Prof.	Cristina Butucea	-	ENSAE
Prof.	Alessandro Rinaldo	-	Carnegie Mellon University
Prof.	Harrison Zhou	-	Yale University

Soutenue le 1er Décembre 2022 devant le jury composé de:

Prof.	Cristina Butucea	-	ENSAE	-	Rapporteure
Prof.	Marc Hoffmann	-	Université Paris-Dauphine	-	Examinateur
Prof.	André Mas	-	Université de Montpellier	-	Examinateur
Dir.	Catherine Matias	-	CNRS	-	Présidente du Jury
Prof.	Joseph Salmon	-	Université de Montpellier	-	Examinateur

"Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d'HDR, les valeurs et principes d'intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l'article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d'intégrité scientifique de l'Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche."

Remerciements

Je suis extrêmement reconnaissant à Cristina Butucea, Alessandro Rinaldo et Harrison Zhou d'avoir accepté de prendre le temps de rapporter mon mémoire malgré leur emploi du temps très chargé. C'est un grand honneur pour moi que des scientifiques d'une telle qualité se penchent sur mes travaux. Marc Hoffmann, André Mas, Catherine Matias, et Joseph Salmon me font également le grand plaisir de participer à mon jury. J'admire tout autant leurs contributions dans leurs domaines scientifiques respectifs que leurs grandes qualités humaines.

Les travaux présentés dans ce mémoire sont le fruit de quatorze années de collaborations scientifiques. Aussi, mes premiers remerciements vont à mon directeur de thèse, Pascal, qui, dès le début de mon doctorat, m'a poussé à cultiver les discussions informelles.

Le métier de chercheur nous donne l'opportunité précieuse de choisir les personnes avec qui nous travaillons. Je tiens tout particulièrement à remercier trois collaborateurs et amis de longue date qui ont une grande influence sur ma pratique scientifique. J'admire l'esprit de synthèse et la gentillesse de Christophe. Sa capacité à épurer, simplifier puis reformuler les preuves les plus complexes profite à toute la communauté. La générosité avec laquelle il encadre ses étudiants est un exemple pour moi. Je suis toujours épaté de la façon avec laquelle Ery identifie des thématiques de recherche originales et construit des questions qui allient profondeur et élégance. Le dynamisme, l'efficacité, la vélocité et l'originalité des idées d'Alexandra n'ont d'égal que sa bienveillance. Il me faudrait au moins me détripler pour profiter pleinement de tels collègues.

Je tiens également à remercier Elisabeth, Etienne, Olga, Magalie, et Sasha. Ces dernières années, nos travaux m'ont fait profiter de vos grandes qualité scientifiques et humaines. Je ne me risquerai pas ici à faire la liste des très nombreux collègues de qui j'ai bénéficié de conseils éclairés ou de discussions animées, mais je leur en suis très reconnaissant.

Au sein de l'INRAE, je bénéficie d'un environnement interdisciplinaire très stimulant. Notamment, les rencontres avec anthropologues, écologues, géographes, généticiens sont enrichissantes à tous points de vue. Je pense notamment aux nombreuses discussions avec Christian, François, Mathieu et Vanesse. Participer, à ma modeste échelle, à vos projets impliquant recherche et société civile est une aventure passionnante. Cette aventure est partagée avec d'autres statisticiens de Paris et Montpellier, notamment Pierre, Sophie, Sarah et Isabelle. Même si je suis un peu envieux de leur expertise en applications et en développement *R*, j'apprends beaucoup à leurs côtés.

J'ai la chance d'avoir rejoint MISTEA il y a maintenant treize ans. La diversité des personnalités ainsi que des profils scientifiques y sont une richesse. Merci à Bénédicte, Bertrand, Céline, Isabelle, Meïli, Nadine, Sébastien,.. pour les échanges informels, qu'ils soient scientifiques ou non. Merci également à Véronique, Maria et Fabienne pour tout l'appui apporté pendant ces années.

Enfin, je suis profondément reconnaissant à mes étudiants, Solène, Yann, Emmanuel et Max. Leur soif d'apprendre ainsi que leurs attentes m'amènent à me questionner et à me renouveler dans mes activités de recherche et d'encadrement.

Contents

Resume

Nicolas Verzelen, PhD
Tenured Research associate

INRAE, UMR MISTEA
Bâtiment 21
2, place Pierre Viala
F-34060 Montpellier, France

Born April 15, 1983
French citizen

Nicolas.Verzelen@inrae.fr

Professional Experience

2009–	**Tenured Research associate** in Statistics within INRAE.
2006–2009	**PhD fellow** and Teaching assistant at Université Paris-Sud.

Education

2005–2008	**PhD** in Mathematics at Université Paris-Sud under the supervision of P. Massart. Title: "Gaussian graphical models and Model selection".
2006	"**Agrégation** de Mathématiques"[1]
2004–2005	**Master 2** in Probability and Statistics at Université Paris-Sud.

Teaching experience

2019–2020	**Tutorial on social network analysis** for Anthropologists and ecologists [2 days]
2015–2022	**Graduate course of Machine Learning** at *Université de Montpellier* [21 hours per year before 2018 and 6 hours per year since 2019]
2015	**Phd Course** on Community Detection at *CREST, ENSAE* [10 hours]
2009–2015	**Undergraduate course** at *Institut-Agro.* Statistic course for agronomy students [40 hours per year]

Scientific and Editorial Duties

Co-Editor in Chief for **ESAIM P-S** journal (2021–2024).
Associate Editor for the **Annals of Statistics** (2018–), **Bernoulli** journal (2019–), and **ESAIM P-S** journal (2018–2021)
Referee for Journals in Statistics, Machine-Learning and conferences in Machine Learning
Referee for research projects (ANR, ISF, ESPRC, ERC)
Member (examinateur) of 4 PhD committees.

Administrative Duties

2020– Leader of the probability and statistics team within MISTEA Lab.

[1] French competitive examination for high-school teachers.

2013–	Member of Hiring committees for Maître de Conférences (Assistant Prof.): AgroParisTech (2013); U. Toulouse Capitole (2013); U. Montpellier (2013); Centrales Marseille (2014); U. Toulouse (2020); ENS Paris-Saclay (2017); U. Marseille (2021)
2015– 2024	Panel Member of the **evaluation committee** (CSS) for Mathematicians and Computer Scientists at INRAE.
2020–	Panel Member for INRAE interdisciplinary Program on Ecosystem Services.
2016–2020	Panel Member for INRAE interdisciplinary Program on global food security.

Conference Organization

2017 and 2022	Member of scientific committee for Journées de Statistiques
2018 and 2022	Session organizer at Journée MAS (France)
2017	Session organizer at the IMS World meeting
2012 –	Co-organization of regional or national 2-days workshops *[around one every two years]*.
2010–2018	Organizer of the joint Statistics seminar between Montpellier University and INRAE

Main Funded Research Projects

2022 – 2024	**ASCAI** (500k€) French-German project funded by the ANR and DFG: Bridging sequential and batch unsupervised learning. *(Co-Leader with A. Carpentier)*
2019 – 2022	**EcoNet** project funded by the ANR: Analysis of Ecological Networks. *(Member)*
2018 – 2022	Research Network (GDR) **Resodiv** funded by the CNRS: interdisciplinary network for objects and knowledge exchange analysis. *(Member of the scientific Panel)*
2017– 2020	**COEX** (700k €) project funded by Agropolis fondation: Adaptative Governance for the Coexistence of Crop Diversity Management Systems coexistence. *(Workpackage Leader)*
2012 – 2020	Research Network **MIRES** funded by INRAE and CNRS: interdisciplinary methods for seed exchange analysis. *(Member and co-leader in 2016–2018)*
2013–2019	Infrastructure d'excellence **PHENOME** funded by the ANR: Technology and Methods for high-throughput phenotypic data. *(Member)*
2012–2015	**CALIBRATION** project funded by the ANR: Mathematical Statistics project on data-driven methods. *(Member)*

PhD Student Supervision

Emmanuel Pilliat (40%)	[2020–2023] co-supervised with A. Carpentier (40%) and J. Salmon (20%). Title: *Optimal change-point estimation and segmentation* [P3].
Yann Issartel (50%)	[2017–2020], co-supervised with C. Giraud (50%). Title: *Inference on random graphs* [P1]. Now Post-doc at CREST ENSAE.
Solène Thépaut (50%)	[2016–2019] co-supervised with C. Giraud (50%). Title: *Clustering Problems for synchrony estimation in population ecology* [P2]. Now Research Engineer at NukkAI.

Other Supervisions

Clémence Huck (50%)	[2020] 6 months Master Internship. Institut Agro.
Emmanuel Pilliat (50%)	[2019] 4 months; Master 2. ENS de Lyon
Alexandre Lecestre (100%)	[2018] 6 Months; Master 2. Université Paris-Saclay.
Mathilde Vimont (60%)	[2018] 6 months; Master Internship. Institut Agro.
Kaïs Zitouni (33%)	[2018] 5 months; Master 2. Université Lyon 1.
Yann Issartel (50%)	[2017] 6 Months; Master 2. Université Paris-Saclay.

Scientific Outputs

In the following thesis, I will describe in depth a sample of my works in mathematical statistics. Those represent the main directions I have been interested in during the last ten years: signal detection and functional estimation [A4, A8, A5, A28, A1, P2, A14], network analysis [A21, A16, A9, A20], clustering [A11, A7, A13], and more recently other unsupervised learning problems such as change-point detection [P3, P4] or seriation problems [P1]. In some chapters, I will describe some open problems as well as some conjectures that are part of my research project for the next years.

Preprints

[P1] Christophe Giraud, Yann Issartel, and Nicolas Verzelen. *Localization in 1D non-parametric latent space models from pairwise affinities*. arXiv preprint arXiv:2108.03098. 2021.

[P2] Solène Thépaut and Nicolas Verzelen. *Optimal Estimation of Schatten Norms of a rectangular Matrix*. arXiv preprint arXiv:2111.13551. 2021.

[P3] Emmanuel Pilliat, Alexandra Carpentier, and Nicolas Verzelen. *Optimal multiple change-point detection for high-dimensional data*. arXiv preprint arXiv:2011.07818. 2020.

[P4] Nicolas Verzelen, Magalie Fromont, Matthieu Lerasle, and Patricia Reynaud-Bouret. *Optimal change-point detection and localization*. arXiv preprint arXiv:2010.11470. 2020.

Journal articles

[A1] Etienne Roquain and Nicolas Verzelen. "False discovery rate control with unknown null distribution: is it possible to mimic the oracle?" In: *The Annals of Statistics (to appear)* (2022).

[A2] Ery Arias-Castro, Antoine Channarond, Bruno Pelletier, and Nicolas Verzelen. "On the estimation of latent distances using graph distances". In: *Electronic Journal of Statistics* 15.1 (2021), pp. 722–747.

[A3] Kay Bogerd, Rui M Castro, Remco van der Hofstad, and Nicolas Verzelen. "Detecting a planted community in an inhomogeneous random graph". In: *Bernoulli* 27.2 (2021), pp. 1159–1188.

[A4] Alexandra Carpentier, Sylvain Delattre, Etienne Roquain, and Nicolas Verzelen. "Estimating minimum effect with outlier selection". In: *The Annals of Statistics* 49.1 (2021), pp. 272–294.

[A5] Alexandra Carpentier and Nicolas Verzelen. "Optimal sparsity testing in linear regression model". In: *Bernoulli* 27.2 (2021), pp. 727–750.

[A6] Ery Arias-Castro, Rong Huang, and Nicolas Verzelen. "Detection of sparse positive dependence". In: *Electronic Journal of Statistics* 14.1 (2020), pp. 702–730.

[A7] Florentina Bunea, Christophe Giraud, Xi Luo, Martin Royer, and Nicolas Verzelen. "Model assisted variable clustering: minimax-optimal recovery and algorithms". In: *The Annals of Statistics* 48.1 (2020), pp. 111–137.

[A8] Alexandra Carpentier and Nicolas Verzelen. "Adaptive estimation of the sparsity in the Gaussian vector model". In: *The Annals of Statistics* 47.1 (2019), pp. 93–126.

[A9] Olga Klopp and Nicolas Verzelen. "Optimal graphon estimation in cut distance". In: *Probability Theory and Related Fields* 174.3 (2019), pp. 1033–1090.

[A10] Ery Arias-Castro, Sébastien Bubeck, Gábor Lugosi, and Nicolas Verzelen. "Detecting Markov random fields hidden in white noise". In: *Bernoulli* 24.4B (2018), pp. 3628–3656.

[A11] Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. "Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization". In: *IEEE Transactions on Information Theory* 64.7 (2018), pp. 4872–4894.

[A12] Olivier Collier, Laëtitia Comminges, Alexandre B Tsybakov, and Nicolas Verzelen. "Optimal adaptive estimation of linear functionals under sparsity". In: *The Annals of Statistics* 46.6A (2018), pp. 3130–3150.

[A13] Christophe Giraud and Nicolas Verzelen. "Partial recovery bounds for clustering with the relaxed K-means". In: *Mathematical Statistics and Learning* (2018), pp. 317–374.

[A14] Nicolas Verzelen and Elisabeth Gassiat. "Adaptive estimation of high-dimensional signal-to-noise ratios". In: *Bernoulli* 24.4B (2018), pp. 3683–3710.

[A15] Ery Arias-Castro, Gábor Lugosi, and Nicolas Verzelen. "Detecting a Path of Correlations in a Network". In: *ALEA* 14 (2017), pp. 33–44.

[A16] Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. "Oracle inequalities for network models and sparse graphon estimation". In: *The Annals of Statistics* 45.1 (2017), pp. 316–354.

[A17] Nicolas Verzelen and Ery Arias-Castro. "Detection and feature selection in sparse mixture models". In: *The Annals of Statistics* 45.5 (2017), pp. 1920–1950.

[A18] Camille Charbonnier, Nicolas Verzelen, and Fanny Villers. "A global homogeneity test for high-dimensional linear regression". In: *Electronic journal of statistics* 9.1 (2015), pp. 318–382.

[A19] Mathieu Thomas, Nicolas Verzelen, Pierre Barbillon, Oliver T Coomes, Sophie Caillon, Doyle McKey, Marianne Elias, Eric Garine, Christine Raimond, Edmond Dounias, et al. "A network-based method to detect patterns of local crop biodiversity: validation at the species and infra-species levels". In: *Advances in Ecological Research* 53 (2015), pp. 259–320.

[A20] Nicolas Verzelen and Ery Arias-Castro. "Community detection in sparse random networks". In: *The Annals of Applied Probability* 25.6 (2015), pp. 3465–3510.

[A21] Ery Arias-Castro and Nicolas Verzelen. "Community detection in dense random networks". In: *Ann. Statist.* 42.3 (2014), pp. 940–969. ISSN: 0090-5364.

[A22] Ingrid Vilmus, Martin Ecarnot, Nicolas Verzelen, and Pierre Roumet. "Monitoring Nitrogen Leaf Resorption Kinetics by Near-Infrared Spectroscopy during Grain Filling in Durum Wheat in Different Nitrogen Availability Conditions". In: *Crop Science* 54.1 (2014), pp. 284–296.

[A23] Nadine Hilgert, André Mas, and Nicolas Verzelen. "Minimax adaptive tests for the functional linear model". In: *The Annals of Statistics* 41.2 (2013), pp. 838–869.

[A24] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. "Graph selection with GGMselect". In: *Statistical applications in genetics and molecular biology* 11.3 (2012).

[A25] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. "High-dimensional regression with unknown variance". In: *Statistical Science* 27.4 (2012), pp. 500–518.

[A26] Nicolas Verzelen. "Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons." In: *Electron. J. Stat.* 6 (2012), pp. 38–90.

[A27] Nicolas Verzelen, Wenwen Tao, and Hans-Georg Müller. "Inferring stochastic dynamics from functional data". In: *Biometrika* 99.3 (2012), pp. 533–550.

[A28] Yuri I Ingster, Alexandre B Tsybakov, and Nicolas Verzelen. "Detection boundary in sparse regression". In: *Electronic Journal of Statistics* 4 (2010), pp. 1476–1526.

[A29] Nicolas Verzelen. "Adaptive estimation of covariance matrices via Cholesky decomposition". In: *Electronic Journal of Statistics* 4 (2010), pp. 1113–1150.

[A30] Nicolas Verzelen. "Adaptive estimation of stationary Gaussian fields". In: *Ann. Statist.* 38.3 (2010), pp. 1363–1402. ISSN: 0090-5364.

[A31] Nicolas Verzelen. "Data-driven neighborhood selection of a Gaussian field". In: *Computational statistics & data analysis* 54.5 (2010), pp. 1355–1371.

[A32] Nicolas Verzelen. "High-dimensional Gaussian model selection on a Gaussian design". In: *Ann. Inst. H. Poincaré Probab. Statist.* 46.2 (2010), pp. 480–524.

[A33] Nicolas Verzelen and Fanny Villers. "Goodness-of-fit tests for high-dimensional Gaussian linear models". In: *The Annals of Statistics* 38.2 (2010), pp. 704–752.

[A34] Nicolas Verzelen and Fanny Villers. "Tests for Gaussian graphical models". In: *Comput. Statist. Data Anal.* 53 (2009), pp. 1894–1905.

[A35] Noel Cressie and Nicolas Verzelen. "Conditional-mean least-squares fitting of Gaussian Markov random fields to Gaussian fields". In: *Computational Statistics & Data Analysis* 52.5 (2008), pp. 2794–2807.

[A36] Nicolas Verzelen, Nicolas Picard, and Sylvie Gourlet-Fleury. "Approximating spatial interactions in a model of forest dynamics as a means of understanding spatial patterns". In: *ecological complexity* 3.3 (2006), pp. 209–218.

Miscellanea

[M1] Etienne Roquain and Nicolas Verzelen. *False discovery rate control with unknown null distribution: illustrations on real data sets.* https://github.com/eroquain/empiricalnull/blob/main/vignette.pdf. 2020.

[M2] Ery Arias-Castro and Nicolas Verzelen. "Discussion of "influential features PCA for high dimensional clustering"". In: *The Annals of Statistics* 44.6 (2016).

PhD Thesis

[T1] Nicolas Verzelen. "Gaussian graphical models and Model selection". PhD thesis. Université Paris Sud-Paris XI, 2008.

Chapter 1

Introduction

1.1 Scientific Trajectory.

My PhD thesis [T1] was about graph selection in Gaussian graphical models. Since these graphs are expressed as the supports of the parameter vectors of some linear regression models [156], I became interested in the minimax analysis of sparse high-dimensional linear regression model [A25].

After the completion of my PhD in 2008, this led me to start working on various minimax estimation problems in high-dimensional linear regression models [A26]. At roughly the same time, I was hired as a research scientist in statistics within INRAE[1]. In this context, I started discussing and collaborating with plant scientists, agronomists, and fellow statisticians within my institute. This lead me to several applied works [A22] as well as some collaborations in functional data analysis [A23, A27] that were motivated by the analysis of high-throughput phenotypic analysis, such as growth curves of plants under hydric stress. Since that period, I really enjoy working on toy and stylized statistical models in order to provide a theoretical basis to practical questions in biology such as heritability estimation in genetics [A14] –see Section 2.3.

My recent research interests have been greatly reshaped by two events. First, Ery Arias-Castro invited me to collaborate with him on community detection problems [A20, A21]. He introduced me to the fields of network analysis and clustering that are central in this manuscript. Second, I have been invited to the MIRES consortium. This interdisciplinary group of anthropologists, geneticists, ecologists, statisticians seeks to provide methods for the analysis of seed exchange networks. Although I have few publications on this topic (but see [A19]), I have been devoting a significant part of my scientific activity to this topic and to the related research projects (see my resume) through informal discussions, consulting, teaching, student supervision... Some of the statistical questions within these applied projects pertain to network analysis and more generally to unsupervised learning. Hence, those are along the lines of my theoretical work. As this manuscript is mostly dedicated to my contributions in mathematical statistics, I do not discuss further my interdisciplinary activities.

In the period 2014–2018, I have been very fortunate to have fruitful and rewarding collaborations on three research directions corresponding to Chapters 2–4: Alexandra Carpentier introduced me to the problem of complexity estimation, which led to our joint works on sparsity testing [A5, A8]. Together with Olga Klopp, we also tackled the problem of sparse graphon estimation [A9, A16]. And with Christophe Giraud, we have provided a general analysis of convex relaxations of

[1] INRAE is a National Institute for Agricultural Research and Environment. I belong to the Applied Mathematics and Computer Science Department.

K-means [A13] for both point and graph clustering.

In the last years, I had the pleasure to participate to the supervision of three PhD students Solène Thépaut, Yann Issartel (both jointly with C. Giraud), and Emmanuel Pilliat (jointly with A. Carpentier and J. Salmon). At the same time, I have moved my interests to other unsupervised problems such as change-point detection [P3, P4] or seriation/ranking problems –see Chapter 5.

Some final words. Most of my mathematical works are grounded in minimax theory. Since my PhD, I have been developing a taste for establishing tight minimax lower and upper bounds for statistical and machine learning questions. Beyond this mathematical inclination, I believe that pinpointing such bounds allows one to form an intuition on the relevant quantities and on the important hypotheses for tackling real-world problems.

1.2 Organization of the manuscript.

The manuscript is organized in four chapters which can be read almost independently. Chapter 2 is dedicated to detection and functional estimation problems mostly in the Gaussian sequence model and in the high-dimensional linear regression model. As a common thread in this chapter, I comment the techniques for establishing the minimax lower bounds. The next chapter is dedicated to network analysis[2]. I mostly describe my joint works on community detection and graphon estimation. Chapter 4 deals with clustering problems. Its organization slightly differs from the other chapters. Starting with a description of my own results on K-means, I provide an account of the-state-of-the art in clustering rates for Gaussian mixtures. This allows me to introduce a few open questions and conjectures I am currently interested in. Finally, in Chapter 5, I describe some recent results and research directions in change-point detection and seriation.

Since the purpose of this manuscript is mainly to give an account of my research results in the last ten years, I would like to stress that I do not fully discuss the related literature and that the bibliography is sometimes partial.

1.3 Notation

Throughout this manuscript, c, c' refer to positive universal constants. For two quantities u and v, $u \lesssim v$ means that $u \leq cv$ for some constant $c > 0$. I write $u \asymp v$, when we have both $u \lesssim v$ and $v \lesssim u$. For a vector θ, $\|\theta\|_q$ stands as usual for its l_q norm for $q \in (0, \infty]$. Besides, $\|\theta\|_0$ stands for the number of non-zero components of θ.

Matrices such as $\mathbf{A}$ or $\mathbf{X}$ are usually written in bold format. Their Frobenius and operator norms are respectively denoted $\|.\|_F$ and $\|.\|_{op}$. I introduce other matrix norms along the manuscript.

[2]clustering problems on networks are in fact postponed to Chapter 4.

Chapter 2

Signal Detection and Functional estimation

2.1 Introduction

Estimation in high-dimensional linear regression models has sparked a lot of interests in the last twenty years [190, 98, 38]. It has lead to fundamental contributions such as compressed sensing theory and analyses of Lasso-type procedures. More generally, these new ideas have spread much beyond this specific model and have had a deep impact in the fields of statistics and machine learning.

Estimation of the regression parameter θ^* in the sparse linear regression model is more or less understood since the works of Candès and Tao [47] and Bickel et al. [22]. This chapter is mainly dedicated to related problems where we do not aim at estimating θ^* completely, but we rather seek to have partial information on θ^*. This includes the problem of testing $\theta^* = 0$ (signal detection), testing whether θ^* belongs to some specific class, or more generally of estimating a low-dimensional function $f(\theta^*)$ of θ^* (functional estimation). Example of functional problems include a specific component θ_i^* of θ^* [43, 120], the l_q norm $\|\theta^*\|_q$ of θ^* [105, 143], or the signal-to-noise ratio [A14, 118, 69].

In this chapter, I will mostly focus the discussion on two toy models: the *Gaussian sequence model* and the random design *Gaussian linear regression model*.

Definition 2.1 (Gaussian sequence model)**.** *We observe a response vector $Y \in \mathbb{R}^n$ sampled from the model*

$$Y = \theta^* + \epsilon \ , \tag{2.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ and $\theta^* \in \mathbb{R}^n$ is an unknown vector.

Definition 2.2 (Random design Gaussian linear regression model)**.** *Let the response vector Y and the covariate matrix $\mathbf{X}$ be such that*

$$Y = \mathbf{X}\theta^* + \epsilon \ , \tag{2.2}$$

where the noise ϵ is Gaussian $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, the parameter $\theta^* \in \mathbb{R}^p$ is unknown, and the design matrix $\mathbf{X} \in \mathbb{R}^{\times p}$ is such that the rows are i.i.d. distributed with $X_i \sim \mathcal{N}(0, \mathbf{\Sigma})$ for some covariance matrix $\mathbf{\Sigma}$.

As we want to emphasize the role played by the sparsity of the parameter θ^* in both models, we introduce, for an integer k, the collection $\mathbb{B}_0[k]$ of k-sparse vectors

$$\mathbb{B}_0[k] := \{\theta \in \mathbb{R}^p : \quad \|\theta\|_0 \leq k\} \ . \tag{2.3}$$

In the Gaussian sequence model, we use, with a slight abuse of notation, the same notation $\mathbb{B}_0[k]$ for the corresponding subset of k-sparse vectors of $\mathbb{R}^n$.

The contributions I describe in this chapter come within the following research program: (i) understanding the optimal rate of testing and estimation for important classes of problems. This includes understanding the role played the sparsity of θ^* on the optimal rate. For this purpose, we adopt a minimax framework. (ii) If possible, introducing a polynomial-time procedure achieving the optimal rate; (iii) If possible also, proposing procedures that are adaptive to the unknown sparsity, which means that they achieve the optimal rate (which depends on $\|\theta^*\|_0$) without prior knowledge on $\|\theta^*\|_0$; (iv) investigating the role played by the knowledge of nuisance parameters such as the noise level σ or the covariance $\mathbf{\Sigma}$ of the covariates.

Organization of the chapter. I will first provide a brief introduction to the minimax framework in both test and functional estimation. After this, I will present five (series) of contributions on related problems. The first one dates back to 2010. Together with Y. Ingster and A. Tsybakov [A28], we characterize the minimax separation distance in the fundamental problem of signal detection (that is testing $\theta^* = 0$) in the sparse linear regression model. The results and techniques are now classical. Section 2.3 is based on a joint work [A14] with E. Gassiat on signal-to-noise ratio (SNR) estimation in linear regression model. Aside from the practical motivations in genetics, this problem is interesting because it unveils the key role played by the knowledge of the design distribution on its difficulty. Section 2.4 is dedicated to the twin problems of testing and estimating the sparsity $\|\theta^*\|_0$ in the Gaussian sequence model and linear regression model. It is based on two joint works with A. Carpentier [A5, A8]. Given the pervasive role of sparsity, this is an important problem per se, but it is also connected to questions of adaptation for confidence regions [96]. From a mathematical viewpoint, it is an emblematic testing problem where both the null and alternative hypotheses are composite, and ont wants to quantify how the size of the null hypothesis has an impact on the difficulty of the testing problem. The next section takes its root on Efron's [76, 78, 74, 75] works on large-scale multiple testing problem. Revisiting many classical data sets, Efron advocates that, in many multiple testing problems, the null distribution is often wrongly chosen and needs in fact to be estimated from the data. This leads to procedures that simultaneously estimate parameters of the null and perform multiple tests on the same data. Efron has proposed a Bayesian approach but there was, until now, no frequentist evidence of the feasibility of such a simultaneous hypothesis learning and hypothesis testing problem. Recasting it as a functional estimation problem in Gaussian sequence model, A. Carpentier, S. Delattre, E. Roquain and myself provide a characterization of the regimes where it is or it is feasible both when the alternative hypotheses are one-sided [A4] and two-sided [A1]. Also, we move slightly away from Gaussian sequence model of Definition 2.1 by allowing some components Y_i to have almost arbitrary distributions, in the spirit of Huber's contamination model [112]. These works [A1, A4] combine some ideas of minimax estimation together with the machinery of false-discovery rate control. Finally, the last section is devoted to a joint recent work with my former PhD student S. Thépaut [P2]. Observing a rectangular matrix hidden in some Gaussian noise, we want to estimate its *effective rank* which expresses as a ratio of Schatten norms[1]. We characterize the minimax estimation rate for this functional.

[1]The q-Schatten norm of a matrix is defined as the l_q norm of its singular values

2.1.1 Detection and minimax separation distances

In this report, we mostly rely on the minimax paradigm to asses the optimality of a specific procedure. For testing problems, the usual counterpart of this minimax risk is the notion of minimax separation distance. As this notion is central in a significant part of my works, I take some time to define them here. Most of the material described here can be found in textbooks [198, 188], although the terminology may differ between research papers in statistics or machine learning.

Let us consider an abstract parametric model $\{\mathbb{P}_{\theta^*},\ \theta^* \in \Theta\}$, but one can think of the Gaussian sequence model (Definition 2.1) for concreteness. Suppose that are given two subsets of parameters $\Theta_0, \Theta_1 \subseteq \Theta$ such that $\Theta_0 \cap \Theta_1 = \emptyset$. Based on an observation $Y \sim \mathbb{P}_{\theta^*}$ for some unknown θ^*, we consider a testing problem of the hypothesis $H_0 : \{\theta^* \in \Theta_0\}$ against the alternative $H_1 : \{\theta^* \in \Theta_1\}$.

Risk of a test. Given a test[2] T, one can define its risk $R(T; \Theta_0; \Theta_1)$ by

$$R(T; \Theta_0; \Theta_1) := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[T = 1] + \sup_{\theta \in \Theta_1} \mathbb{P}_\theta[T = 0] \ . \tag{2.4}$$

It corresponds to the sum of the maximum type I and type II error probabilities. Let us pause to discuss this choice of risk measure. As we consider the sum of the two types of errors in (2.4), this definition breaks the asymmetry between the two error terms. In practice, we may prefer considering a as measure of risk the type II error probability $\sup_{\theta \in \Theta_1} \mathbb{P}_\theta[T = 0]$ while restricting our attention to test whose size is less or equal to some fixed given α. In fact, one can readily adapt the minimax testing theory to this error measure, see e.g. [16] in the Gaussian sequence model. Still, we keep working with the risk (2.4) as it is slightly easier to handle.

Minimax risk of a test. Coming back to (2.4), we define the *minimax risk of the testing problem* as the infimum $R^*(\Theta_0; \Theta_1) = \inf_T R(T; \Theta_0; \Theta_1)$ where the infimum is taken over all possible tests. If this minimax risk is close to 0, this entails that exists a test whose type I and II error probabilities are close to zero. Conversely, a random guess test T is a test which samples 0 or 1 independently of the observations Y. By definition, it risk equals one. If the minimax risk is close to one, this entails that the risk of all possible tests is as bad as that of a random guess test. If we work in an asymptotic setting, where Θ_0 and Θ_1 are allowed to vary with n, we say that a (sequence of) tests T_n is *asymptotically powerful* if $R(T_n; \Theta_0; \Theta_1) \to 0$. This also entails that $R^*(\Theta_0; \Theta_1) \to 0$. We say that all tests T_n are *asymptotically powerless* (or equivalently that the two hypotheses Θ_0 and Θ_1 *merge asymptotically*) if $R^*(\Theta_0; \Theta_1) \to 1$.

Minimax separation distance. Unfortunately, the risk measure $R(T; \Theta_0; \Theta_1)$ is not adequate per se is many situations. Consider for instance a signal detection problem in the Gaussian sequence model (Definition 2.2) with $\Theta_0 = \{0\}$ and $\Theta_1 = \mathbb{R}^n \setminus \{0\}$. For any test T, $R(T; \Theta_0; \Theta_1 \setminus \Theta_0)$ is higher than one because Θ_1 contains parameters that are arbitrarily close to zero. Indeed, the test T either suffers from a high type I error probability or from a high type II error probability in the vicinity of Θ_0. Intuitively, the behavior of T in the vicinity of Θ_0 is not that relevant, as the primary objective of the statistician is (i) T accepts the null when $\theta^* \in \Theta_0$ and (ii) rejects the null when θ^* differs enough from Θ_0. For this reason, we may want to quantify the risk of T outside the vicinity of Θ_0. Let us suppose that the parameter set Θ is endowed with a pseudo-distance $d(.,.)$. Given $\theta' \in \Theta$, we define the distance $d(\theta'; \Theta_0) = \inf_{\theta \in \Theta_0} d(\theta', \theta)$ of θ' to Θ_0. Then, for $\rho > 0$, define

[2] Here, a test T is a measurable function that maps the observation Y to $\{0, 1\}$

$\Theta_1[\rho] = \{\theta \in \Theta_1 : d(\theta, \Theta_0) > \rho\}$ the subset of all alternative parameters that lie at a distance a least ρ from the null hypothesis. Then, for a fixed ρ, we define the risk

$$R(T; \Theta_0; \Theta_1; \rho) := \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta,\sigma}[T = 1] + \sup_{\theta \in \Theta_1[\rho]} \mathbb{P}_{\theta,\sigma}[T = 0] \ , \tag{2.5}$$

where we compute the type II error probability over the restricted parameter set $\Theta_1[\rho]$. By definition this risk $R(T; \Theta_0; \Theta_1; \rho)$ is a non-increasing function of ρ. Then, for a fixed $\gamma \in (0, 1)$, the *separation distance* $\rho_\gamma[T; \theta; \Theta_1]$ of the test T is defined by

$$\rho_\gamma[T; \Theta_0; \Theta_1] = \inf\{\rho > 0 : \quad R(T; \Theta_0; \Theta_1; \rho) \leq \gamma\} \ . \tag{2.6}$$

Finally, we define the *minimax separation distance* $\rho^*_\gamma[\Theta_0; \Theta_1] = \inf_T \rho_\gamma[T; \Theta_0; \Theta_1]$ as the infimum of all tests T of the separation distance. Intuitively, $\rho^*_\gamma[\Theta_0; \Theta_1]$ is the smallest distance ρ to the null hypothesis such that there exists a test which deciphers Θ_0 from $\Theta_1[\rho]$ with a high confidence. Alternatively, $\rho^*_\gamma[\Theta_0; \Theta_1]$ quantifies how far the parameter θ^* should be from the null hypothesis so that, under $\mathbb{P}_{\theta^*}$, a suitable test will reject the null with high probability.

To finish with some terminology, when the null hypothesis $\Theta_0 = \{0\}$ is simple, the testing problem is sometimes referred as a *signal detection* problem and the corresponding minimax separation distance is referred as the *detection boundary*. This boundary interprets as the minimal magnitude of the signal to detect its existence.

Some comments on the testing minimax lower bound proof techniques. If the two hypotheses in the testing problems are simple, that is $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, then the minimax risk $R^*(\Theta_0; \Theta_1)$ nicely expresses as the total variation distance between $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_{\theta_1}$. Indeed,

$$R^*(\Theta_0; \Theta_1) = \inf_T \mathbb{P}_{\theta_0}[T = 1] + \mathbb{P}_{\theta_1}[T = 0] = 1 - \sup_T \big[\, \mathbb{P}_{\theta_0}[T = 0] - \mathbb{P}_{\theta_1}[T = 0] \big] = 1 - \| \, \mathbb{P}_{\theta_0} - \mathbb{P}_{\theta_1} \, \|_{TV} \ ,$$

so that lower bounding the minimax risk amounts to upper bounding the total variation distance between $\| \, \mathbb{P}_{\theta_0} - \mathbb{P}_{\theta_1} \, \|_{TV}$. In many situations, the total variation norm is not easy to handle. In turn, one then upper bounds it by a more suitable discrepancy measures such as Hellinger distance, the Kullback-Leibler discrepancy (by Pinsker's inequality), or the χ^2 divergence (by Cauchy-Schwarz inequality).

If the hypotheses are now composite, one can always lower bound the minimax risk by taking two specific parameters $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$. Unfortunately, this does not allow to capture the difficulty of problem when at least one of the two parameter sets is large.

When the alternative hypothesis is *composite*, a standard work-around is to introduce a prior measure μ supported on Θ_1 and to replace the supremum over $\theta \in \Theta_1$ in the definition of the risk by an integral over μ. This leads us to

$$R^*(\Theta_0; \Theta_1) \geq \inf_T \left[\mathbb{P}_{\theta_0}[T = 1] + \int \mathbb{P}_\theta[T = 1] \mu(d\theta) \right] = 1 - \| \, \mathbb{P}_{\theta_0} - \mathbf{P}_1 \|_{TV} \ , \tag{2.7}$$

where we define $\mathbf{P}_1 = \int \mathbb{P}_\theta \, \mu(d\theta)$ the marginal distribution of Y when θ is sampled according to μ. The challenge is then (i) to upper bound the total variation distance between $\mathbb{P}_{\theta_0}$ and the mixture distribution $\mathbf{P}_1$ and (ii) to build a suitable prior measure μ on Θ_1 to maximize this upper bound. Unfortunately, the Kullback-Leibler discrepancy or the Hellinger distance are sometimes difficult to compute when a mixture distribution is involved[3]. In this case, it is sometimes easier to work with

[3]This is for instance the case in the toy problem of signal detection in the Gaussian sequence model

the χ^2 divergence. For any $\theta \in \Theta_1$, define the likelihood ratio $L_\theta = d\,\mathbb{P}_\theta / d\,\mathbb{P}_{\theta_0}$ and the integrated likelihood ratio $L = \int L_\theta \mu(d\theta)$. By definition of the likelihood, we have

$$\begin{aligned} 2\| \mathbb{P}_{\theta_0} - \mathbf{P} \|_{TV} &= \mathbb{E}_{\theta_0}\left[|L-1|\right] \\ &\leq \left[\mathbb{E}_{\theta_0}\left[(L-1)^2\right]\right]^{1/2} = \left(\mathbb{E}_{\theta_0}[L^2]-1\right)^{1/2} , \end{aligned} \tag{2.8}$$

where $\mathbb{E}_{\theta_0}[L^2] = \int \mathbb{E}_{\theta_0}[L_\theta L_{\theta'}]\mu(d\theta)\mu(d\theta')$. In summary, the minimax risk $R^*(\Theta_0;\Theta_1)$ is close to one (and no test performs much better than random guess), provided that, for a suitable prior μ, the second moment $\mathbb{E}_{\theta_0}[L^2]$ is close to one. This approach, coined as the *second moment* method, is really powerful and allowed to recover tight bound the minimax separation distances for many problems since the seminal works of Ingster [114–116]. In particular, the detection boundaries for signal detection described in Section 2.2 proceed from this approach.

Unfortunately, the second moment method may lead to over-optimistic minimax lower bounds for two reasons:

(a) Cauchy-Schwarz inequality in (2.8) is too rough, or equivalently the χ^2 divergence between probability measures can be much higher than their total variation distance. One work-around is to apply Cauchy-Schwarz inequality to *truncated version* of the likelihood in order to control its second moments. Up to our knowledge, this idea has been originally introduced by Ingster [117]. We further discuss this idea and rely on this approach in Section 3.1 for the problem of community detection.

(b) The second moment approach is particularly suited to signal detection problem where the null hypothesis is simple and the alternative hypothesis is composite. For *composite-composite testing* problems, it is still possible to lower bound the minimax risk $R^*[\Theta_0,\Theta_1]$ by picking a single parameter $\theta_0 \in \Theta_0$. Unfortunately, the resulting minimax lower bound cannot account for the size and the complexity of both Θ_0 and Θ_1 and will lead to suboptimal lower bounds. For this purpose, we would like to build two prior measures μ_0 and μ_1 on Θ_0 and Θ_1 with corresponding marginal measure $\mathbf{P}_0 = \int \mathbb{P}_\theta\, \mu_0(d\theta)$ and $\mathbf{P}_1 = \int \mathbb{P}_\theta\, \mu_1(d\theta)$ which would, arguing as in (2.7), lead to $R^*(\Theta_0;\Theta_1) \geq 1 - \|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}$. However, this total variation distance turns out to be much more delicate to upper bound than in (2.7) because it now involves the distance between two mixture distributions. For instance, the likelihood ratio $L = d\mathbf{P}_1/d\mathbf{P}_0$ now involves a ratio of two integrals and its second moment cannot be explicitly computed even in the simple Gaussian sequence model.

The works described in Section 2.3–2.6 all fall within this class of models where we have to rely on composite-composite problems for the minimax lower bound. We briefly discuss the mathematical techniques to deal with this issue in the corresponding sections.

2.1.2 Functional Estimation

.

Still considering an abstract parametric model $\{\mathbb{P}_{\theta^*},\ \theta^* \in \Theta\}$ while having in mind the Gaussian sequence model for illustration purpose, we are also given a function $f : \Theta \to \mathcal{X}$, where $\mathcal{X}$ is typically of much lower dimension than the original parameter size, the prominent example being $\mathcal{X} = \mathbb{R}$. A functional estimation problem is that of estimating $f(\theta^*)$. In the Gaussian sequence model, prominent examples are the linear function $f(\theta) = \sum_i \theta_i$ or the l_q norm $f(\theta) = \|\theta\|_q$.

Estimation of functionals of non-parametric or high-dimensional models has a long history that dates back to 70's and is still vivid – see e.g. [105, 106, 129, 199, 200, 61, 113, 46, 45,

139, 26, 71, 145]. Provided that the functional is smooth enough and under some additional conditions, Koltchinskii and collaborators [132, 133, 130] have introduced a general bias reduction approach that ensures a near parametric rate of convergence. By contrast, estimation of non-smooth functionals or estimation of smooth functional on non-open subsets (such as $\mathbb{B}_0[k]$) is usually dealt with using case-by-case methods. Despite this, some general techniques turn out to be fruitful in many cases: the methods of fuzzy hypotheses for deriving minimax lower bounds, or more recently the polynomial approximation and moment matching techniques [199, 122, 46, 143]. I describe the former below and will explain the latter later in the chapter.

Impossibility results and connection with minimax testing. The method of fuzzy hypotheses provides a simple connection between functional estimation problems and testing problems–see again textbooks such as [188] for more details. For a simple functional where $\mathcal{X} = \mathbb{R}$, we define, for any $a \in \mathbb{R}$, the parameter set $\Theta_a = f^{-1}(\{a\})$. If we want to prove that any estimator $\widehat{f}$ of $f(\theta^*)$ suffers from an error δ, it suffices to prove that there exist a and b with $|b-a| \geq 2\delta$ such that the minimax risk of testing $H_0 : \{\theta^* \in \Theta_a\}$ against $H_1 : \{\theta^* \in \Theta_b\}$ is high. Indeed, let us build a test $\widetilde{T}$ such that $\widetilde{T} = 0$ if $|\widehat{f} - a| \leq \delta$. Then, one deduces

$$\begin{aligned}\sup_{\theta^* \in \Theta_a \cup \Theta_b} \mathbb{P}_{\theta^*}[|\widehat{f} - f(\theta*)| \geq \delta] &\geq \max\left[\sup_{\theta^* \in \Theta_a} \mathbb{P}_{\theta^*}[\widetilde{T} = 1], \sup_{\theta^* \in \Theta_b} \mathbb{P}_{\theta^*}[\widetilde{T} = 1]\right] \\ &\geq \frac{1}{2} R[\widetilde{T}; \Theta_a, \Theta_b] \geq \frac{1}{2} R^*[\Theta_a, \Theta_b] \ .\end{aligned}$$

As a consequence, lower bounding the error for estimating $f(\theta^*)$ expresses as the minimax risk of a testing problem where the composite hypotheses correspond to slices of the function f. It turns out many impossibility results for functional estimation are proved using this approach. For some functional estimation problems involving e.g. non-smooth functionals, one needs to rely on composite-composite hypothesis testing problems as described in the previous subsection to recover the tight minimax rate.

2.2 Signal detection in sparse Linear regression

This section is mainly based on a joint work [A28] with Y. Ingster and A. Tsybakov. We consider the prototypical problem of signal detection in the possibly sparse linear regression model $Y = \mathbf{X}\theta^* + \epsilon$ (see Definition 2.2). More formally, we aim at testing the null hypothesis $H_0 : \{\theta^* = 0\}$ (no signal) against specific alternatives. As we are interested in situations where the regression parameter θ^* is possibly sparse, we consider alternative hypotheses of the form $H_k : \{\theta^* \in \mathbb{B}_0[k] \setminus \{0\}\}$. To define the separation distance as in (2.6), we introduce, for any $\rho > 0$, the parameter set $\mathbb{B}_0[k, \rho] = \{\theta \in \mathbb{B}_0[k] : \ \|\theta\|_2 \geq \rho\}$ of k-sparse vectors whose norm is higher than ρ.

We first focus on the simpler case where the noise level σ is known and the covariates are independent ($\mathbf{\Sigma} = \mathbf{I}_p$ in Definition 2.2). Our aim is then (i) to characterize the minimax separation distance $\rho^*_\gamma[k]$ as a function of (k, p, n, σ) and (ii) to build a detection procedure which achieves this optimal separation, this simultaneously for all k.

Before stating the main result, let us make two observations. First, the naive estimator $\widehat{\theta}$ of θ^* based on the empirical correlations satisfies $\widehat{\theta} = \frac{1}{n}\mathbf{X}^T Y = \frac{1}{n}\mathbf{X}^T\mathbf{X}\theta^* + \frac{1}{n}\mathbf{X}^T\epsilon$ is, in expectation, equal to θ^*. Intuitively, its distribution can be compared to $\mathcal{N}(\theta^*, \frac{1}{n}\sigma^2\mathbf{I}_p)$. In fact, under the null ($\theta^* = 0$), conditionally to ϵ, we even have $\sigma\sqrt{n}\widehat{\theta}/\|\epsilon\|_2 \sim \mathcal{N}(0, \sigma^2\mathbf{I}_p/n)$. In light of this, it may be tempting to apply the same testing methodologies as in the Gaussian sequence model [61, 72, 16].

In the dense case (large k) we may want to reject the null when $\|\widehat{\theta}\|_2^2$ is large. In the sparse case, we may want to apply an Higher-Criticism approach [72], that is rejecting the null, when the number N_t of coordinates such that $|\widehat{\theta}_i| > t$ is unusually high, this for at least one threshold $t > \sigma/\sqrt{n}$. A second observation is that $\|Y\|_2^2$, in expectation, equals $n\|\theta\|_2^2 + n\sigma^2$ so that a unusual value of $\|Y\|_2^2$ may indicate the presence of the signal. It turns out that a combination of these three statistics simultaneously achieves the minimax separation distance $\rho_\gamma^*[k]$, this for a wide collection of sparsities k.

Theorem 2.3 ([A28]). *Consider an asymptotic setting where $p \to \infty$, $n \to \infty$ (with $\log^3(p) = o(n)$) and where the sparsity $k = p^{1-\beta}$ for some $\beta \in (0,1)$.*

- *If $\beta \in (1/2, 1)$ (highly sparse alternative), then $\frac{\rho_\gamma^{*2}[k]}{\sigma^2} \asymp \frac{k\log(p)}{n} \wedge \frac{1}{\sqrt{n}}$*
- *If $\beta \in (0, 1/2]$ (dense alternative), then $\frac{\rho_\gamma^{*2}[k]}{\sigma^2} \asymp \frac{\sqrt{p}}{n} \wedge \frac{1}{\sqrt{n}}$*

In the highly-sparse regime, when $k\log(p) \ll \sqrt{n}$, it is even possible to characterize the tight leading constant in the minimax separation distance instead of simply giving an expression up to a multiplicative constant, as in this theorem. The minimax lower bound is proved using the classical second moment technique described in the previous section, although recovering the tight constant requires a truncated version.

Unknown noise level σ. Among the three statistics that achieve the minimax separation distance, the test based on $\|Y\|_2^2$ crucially requires the knowledge of σ^2. Indeed, the corresponding test rejects for large values of $[\|Y\|_2^2 - n\sigma^2]/\sigma^2$. Hence, mimicking its performances would require to plug an estimator $\widehat{\sigma}$ satisfying $\widehat{\sigma}^2 - \sigma^2 = o_{\mathbb{P}}(n^{-1/2})$ which is unfortunately not feasible in a high-dimensional situation. It turns out that, when σ is unknown, the corresponding square minimax separation distance respectively becomes of the order of $\frac{k\log(p)}{n}$ (for $\beta \in (1/2, 1)$) and $\frac{\sqrt{p}}{n}$ (for $\beta \in (0, 1/2]$) –see [A28]. In comparison to the known σ setting, we observe that the distance $1/\sqrt{n}$ (which was achieved by the statistic $\|Y\|_2^2$) is not achievable anymore. Importantly, these rates $k\log(p)/n$ and $\sqrt{p}/n$ are derived under stronger restrictions on p. In particular, the $k\log(p)/n$ rate is achievable by the higher criticism test only if $k\log(p) = o(n)$. In a related work, [A26] I have been interested in minimax estimation and testing in the so-called ultra-high dimension regime where $k\log(p)$ exceeds n. In particular, it is proved that, in this regime, the minimax separation distance is of the order of $\exp(ck\log(p)/n)$ for some constant $c > 0$[4]. In other words, the minimum signal strength has to be exponentially higher in this $k\log(p) > n$ regime than the usual bound $k\log(p)/n$.

Related literature and later works. The work [6] appeared simultaneously to ours and provides a similar characterization of the detection rate in Theorem 2.3. As an aside, some of the techniques described here turned out to be useful for the construction of adaptive confidence region [164] of θ^*. More recently, Mukherjee and sen [160] have provided a delicate analysis of the risk when the parameter θ^* is in the vicinity of the detection boundary.

Other signal detection problems. Together with different colleagues, I have also worked on related signal detection problems in the functional linear regression model [A23], in Gaussian Markov fields [A10, A15], or in two-sample models [A6].

[4]The result is stated for $k \le p^{1/3}$, but the proof holds almost verbatim for $k = p^{1/2-\delta}$ with $\delta > 0$

2.3 Signal to Noise Ratio estimation

This section is mainly based on a joint work [A14] with E. Gassiat.

Motivation. As in the previous section, consider the random design high-dimensional linear regression model (2.2). Assume that we have detected the existence of the signal ($\theta^* \neq 0$), then the next step would be to estimate the magnitude of the signal. Define the signal-to-noise ratio (SNR) or equivalently the proportion of explained variation by

$$SNR^* := \frac{\mathbb{E}\left[\|\mathbf{X}_1^T\theta^*\|_2^2\right]}{\sigma^2} = \frac{\|\mathbf{\Sigma}^{1/2}\theta^*\|_2^2}{\sigma^2} \quad \text{and} \quad \eta^* := \frac{\mathbb{E}[\|\mathbf{X}_1^T\theta^*\|_2^2]}{\mathrm{Var}\,(Y_1)} = \frac{SNR^*}{1+SNR^*}\,. \tag{2.9}$$

The latter functional η^* interprets as the ratio of variance of the signal to the total amount of variance of Y. For low-dimensional linear regression models ($p \ll n$), the functional η^* is estimated by the coefficient of determination, which is routinely computed in data analyses. Obviously, the (vanilla version of) coefficient of determination cannot be used when $n < p$. Our interest in this functional is mainly motivated by heritability estimation problems in quantitative genetics. In such studies, the response variable is a phenotype measured on n individuals and the predictor matrix $\mathbf{X}$ are genetic markers on each of these individuals. Then, the heritability of a phenotype is quantified by the proportion of explained variation. Usually, the number p of genetic markers greatly exceeds the number n of individuals. To handle this high-dimensional setting, researchers have assumed that the phenotype can be explained by a small number k of markers, which has spurred interests for statistical methods exploiting the sparsity of θ^*. However, in some complex human traits, it appeared that there was a huge gap (which has been called the "dark matter" of the genome) between the genetic variance explained by populations studies and the one obtained by genome wide associations studies (GWAS), see [153], [180] or [102]. To explain this gap, it has been hypothesized that some traits might be "highly polygenic", meaning that the corresponding regression coefficient vector θ^* may not be considered as sparse. As a consequence, sparsity-based methods would be questionable in this situation. In [A14], we have been interested in characterizing the minimax risk for η^* when θ^* belongs to sparsity classes $\mathbb{B}_0[k]$ for k ranging from 1 to p.

Related literature. If the noise level σ in the model (2.2) is known, then the simple estimator $\widehat{\eta} = 1 - n\sigma^2/\|Y\|_2^2$ is easily shown to be $n^{-1/2}$-consistent. For this reason, estimating η^* only makes sense if σ is unknown and we assume it henceforth. Most of the recent literature at the time of [A14] assumed that the covariance matrix $\mathbf{\Sigma}$ of the covariates (see definition (2.1)) was known and was equal to the identity matrix $\mathbf{I}_p$. In that setting, Dicker [69] and Janson et al. [118] introduced suitable U-statistics of the form

$$T = \frac{Y^T\left(\mathbf{X}\mathbf{X}^T - tr(\mathbf{X}\mathbf{X}^T)\mathbf{I}_n/n\right)Y}{n^2}\,. \tag{2.10}$$

In a high-dimensional asymptotic regime where $p/n \to c \in (0,\infty)$, they proved that $T - \|\mathbf{\Sigma}^{1/2}\theta^*\|_2^2 = (\sigma^2 + \|\mathbf{\Sigma}^{1/2}\theta^*\|_2^2)O_P(1/\sqrt{n})$. The striking consequence of those results, is that is possible to estimate η^* in a high-dimensional regime, even when $p \asymp n$, and this without any sparsity assumption on θ^*.

Minimax and adaptive minimax risk for η^*. In [A14], we make three contributions: first, when $\mathbf{\Sigma}$ is known, we characterize the minimax convergence risk for η^* as a function of (n, p, k)

where we recall that $k \in [1, p]$ is the sparsity of θ^*. For instance, if $p \in [n, n^2]$, the minimax risk (in squared error) is of the order of

$$\min\left[\frac{1}{n} + \frac{k^2 \log^2(p)}{n^2}, \frac{p}{n^2}\right] , \tag{2.11}$$

and is achieved by either a sparse estimator based on the square-root Lasso if k is small (sparse regime) or by some transformation of the statistic T is k is large (dense regime). In practice, the sparsity parameter k is unknown. Our second contribution is an adaptive estimator that combines T and the square-root Lasso. This new estimator, simultaneously achieves the minimax risk (2.11) of all sparsity k, up to a multiplicative $\log(p)$ factor. This $\log(p)$ price is then proved to be unavoidable for adaptation to unknown sparsity.

Unknown distribution of the design. The previous results suggest that η^* can be consistently estimated even in the dense setting (e.g. $k = p$) as long as $p \ll n^2$. However, the covariance matrix $\mathbf{\Sigma}$ is certainly not known in many practical situations, such as in heritability distribution. Unfortunately, none of the known dense estimators works in this regime. For instance, for general $\mathbf{\Sigma}$, the U-statistic T [5] converges to $\|\mathbf{\Sigma}\theta^*\|_2^2$ instead of the desired signal level $\|\mathbf{\Sigma}^{1/2}\theta^*\|_2^2$. Our last contribution is to show, that when $\mathbf{\Sigma}$ is unknown (but is still nicely conditioned), it is impossible to estimate consistently the heritability as long as $p \geq n^{1+\kappa}$ for any $\kappa > 0$ arbitrarily small. In comparison to the known $\mathbf{\Sigma}$ case where consistent estimation is possible as long as $p \ll n^2$, here we cannot handle a non-sparse high-dimensional setting.

Intuitively, the proof of the last impossibility result relies on the fact that, even when p is higher than n, one can consistently estimates the functional $\theta^{*T}\mathbf{\Sigma}^q\theta^*$ for $q = 2, 3, 4, \ldots$ by suitable U-statistics. However, this is not the case for the signal level $\theta^{*T}\mathbf{\Sigma}\theta^*$, which corresponds to $q = 1$. Hence, the key idea of the proof is too build two collections of parameters $(\theta^*, \mathbf{\Sigma}) \in \mathcal{B}_1$ and $(\theta^*, \mathbf{\Sigma}) \in \mathcal{B}_2$ such that: (i) the first 'moments' $\theta^{*T}\mathbf{\Sigma}^q\theta^*$ for $q = 2, 3, \ldots, c_\kappa$ are the same in $(\mathcal{B}_1)$ and $(\mathcal{B}_2)$ while (ii) $\theta^{*T}\mathbf{\Sigma}\theta^*$ differs in $(\mathcal{B}_1)$ and in $(\mathcal{B}_2)$. Thanks to (i), we are able to show that is impossible to decipher whether the true parameters belong to $(\mathcal{B}_1)$ or $(\mathcal{B}_2)$. Thanks to (ii), we deduce that an error in the testing problem $\{(\theta^*, \mathbf{\Sigma}) \in \mathcal{B}_1\}$ against $\{(\theta^*, \mathbf{\Sigma}) \in \mathcal{B}_2\}$ implies an error for the estimation of η^*.

Subsequent works and a conjecture. More recently, Kong and Valiant [135] have improved our results for unknown $\mathbf{\Sigma}$ into several directions. On the positive side, they made use of the intuition behind the minimax lower bound to estimate $\theta^{*T}\mathbf{\Sigma}\theta^*$ by a linear combination of unbiased estimators of $\theta^{*T}\mathbf{\Sigma}^q\theta^*$, for $q = 2, \ldots, q^*$. This allows them, for any *fixed* $\zeta \in (0, 1)$, to build an estimator $\widehat{\eta}$ of η^* achieving $|\widehat{\eta} - \eta^*| \leq \zeta$ provided that $n \geq \psi_1(1/\delta)p^{1-1/\log(1/\delta)}$ (under some additional assumptions). On the negative side, they proved that the dependency in $1/\log(1/\delta)$ cannot be improved. Their proof is based on our strategy except that the sets $\mathcal{B}_1$ and $\mathcal{B}_2$ and the corresponding prior distributions are more delicately chosen. Interestingly, Kong and Valiant also extend their method to estimate the SNR in logistic regression.

Whereas the case of known $\mathbf{\Sigma}$ covariance is mostly understood, there remains some open questions for unknown design distribution in a high-dimensional regime $p = n^{1+\kappa}$ for $\kappa > 0$ fixed. One the positive side, a plug-in approach based on the square-root Lasso achieves a square error of the order of

$$\max\left[\frac{1}{n} + \frac{k^2 \log^2(p)}{n^2}, 1\right] . \tag{2.12}$$

[5] When $\mathbf{\Sigma}$ is known, the idea is to whiten the covariance by computing $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{\Sigma}^{-1/2}$ then compute T with $(Y, \tilde{\mathbf{X}})$.

On the negative side, we are only able to prove the corresponding lower bound either for $k \ll \sqrt{p}$ or for $k = p$. For $k \in [\sqrt{p}, p]$, we conjecture that the upper bound (2.12) is nearly tight, but we do not manage yet to prove the matching lower bound.

2.4 Sparsity testing and Estimation

This section is based on two joint works with A. Carpentier [A5, A8]. They fall within the wider research program of estimating/testing the complexity of a parameter. This objective may be tackled for differents reasons. First, complexity estimation allows to assess the relevance of specific parameter estimation approaches. For instance, inferring the smoothness of a function allows to justify the use of regularity-based procedures. Second, the construction of adaptive confidence regions is related to complexity testing problem since the size of an adaptive confidence region should depend on the complexity of the unknown parameter [109]. Finally, in some practical applications, the primary objective is rather to evaluate the complexity of the parameter than the parameter itself. This is for instance the case in some heritability studies where the goal is to decipher whether a trait is multigenic or "highly polygenic" which amounts to inferring whether a high-dimensional regression parameter is sparse or dense [186, 153].

In [A8], we first deal with this problem in the Gaussian sequence model (Definition 2.1) and we consider the twin objectives of (i) estimating the number $\|\theta^*\|_0$ of non-zero components of θ^* and (ii) given some non-negative integer k_0, testing whether $\|\theta^*\|_0 \leq k_0$ or $\|\theta^*\|_0 > k_0$. The former problem is referred as sparsity estimation and the latter as sparsity testing.

Since the functional $\|\theta^*\|_0$ is not even continuous with respect to θ^*, quantifying the performances of an estimator $\widehat{T}$ with respect to an error of the form $|\widehat{T} - \|\theta^*\|_0|$ does not really make sense as the maximum risk of such an estimator is necessarily high, for instance if θ^* has many very small components. As the formalization of the error measures for sparsity estimation is more intricate, I mostly describe here the result for sparsity testing.

Minimax separation distances. Here, I specialize the general presentation of minimax separation distance to the sparsity testing problem. Given a non-negative integer k_0 and a positive integer Δ such that $k_0 + \Delta \leq n$, we consider the problem testing the null hypothesis $H_{k_0} : \{\theta^* \in \mathbb{B}_0[k_0]\}$ against the alternative $\{\theta^* \in \mathbb{B}_0[k_0 + \Delta] \setminus \mathbb{B}_0[k_0]\}$. We quantify the distance of θ^* to the null hypothesis using $d_2(\theta^*, \mathbb{B}_0[k_0]) := \inf_{u \in \mathbb{B}_0[k_0]} \|\theta^* - u\|_2$. Then, for any $\rho > 0$, the corresponding subset of alternative hypotheses where we remove parameters θ^* that are ρ-close to the null is defined by

$$\mathbb{B}_0[k_0 + \Delta, k_0, \rho] := \{\theta \in \mathbb{B}_0[k_0 + \Delta] : d_2(\theta, \mathbb{B}_0[k_0]) \geq \rho\} \ . \tag{2.13}$$

Then, we define as in (2.6) the separation distance $\rho_\gamma(T; k_0, \Delta)$ of a test T as the minimal distance ρ to the null hypothesis such that the test T has a risk smaller than γ. Finally, $\rho^*_\gamma[k_0, \Delta] = \inf_T \rho_\gamma(T; k_0, \Delta)$ stands for the minimax separation distance. In other words, $\rho^*_\gamma[k_0, \Delta]$ is the minimal distance to $\mathbb{B}_0[k_0]$ for a parameter θ^* in $\mathbb{B}_0[k_0 + \Delta]$ to be reliably detected.

Contribution. In [A8], our contribution is threefold:

(a) When σ (the noise level) is known, we provide a tight characterization of the the minimax separation distance $\rho^*_\gamma[k_0, \Delta]$ for all integers k_0 and all $\Delta > 0$– see Table 2.1. Besides, we introduce a procedure which is simultaneously minimax over all alternatives $\{\theta^* \in \mathbb{B}_0[k_0 + \Delta] \setminus \mathbb{B}_0[k_0]\}$ for $\Delta \in [1, n - k_0]$. From Table 2.1, we observe that, for $k_0 \leq \sqrt{n}$, the minimax

Table 2.1: Square minimax separation distances (in the $\asymp_\gamma$ sense) when the noise level σ is known for all $k_0 \in [0, n-1]$ and $\Delta \in [1, n-k_0]$.

k_0	Δ	$\rho^{*2}_\gamma[k_0,\Delta]/\sigma^2$
$k_0 \le \sqrt{n}$	$1 \le \Delta \le \sqrt{n}$	$\Delta \log\left(1+\frac{\sqrt{n}}{\Delta}\right)$
	$\sqrt{n} < \Delta \le n-k_0$	$\sqrt{n}$
$k_0 > \sqrt{n}$	$1 \le \Delta \le \sqrt{n^{1/2}k_0}$	$\Delta \log\left(1+\frac{k_0}{\Delta}\right)$
	$\sqrt{n^{1/2}k_0} \le \Delta \le k_0$	$\Delta \frac{\log^2\left(1+\frac{k_0}{\Delta}\right)}{\log\left(1+\frac{k_0}{\sqrt{n}}\right)}$
	$k_0 \le \Delta \le n-k_0$	$\frac{k_0}{\log\left(1+\frac{k_0}{\sqrt{n}}\right)}$

separation distance $\rho^*_\gamma[k_0,\Delta]$ is similar to the detection one $\rho^*_\gamma[0,\Delta]$, hence the size of the null hypothesis does not have an impact on the difficulty of the problem. For larger k_0 and for large Δ, $\rho^*_\gamma[k_0,\Delta]$ is much larger than its counterpart $\rho^*_\gamma[0,\Delta]$ and is almost proportional to the size k_0.

(b) In the more realistic setting where the noise level σ is unknown, the minimax separation distance $\rho^*_{\gamma,\mathrm{var}}[k_0,\Delta]$ is established and minimax adaptive tests are exhibited. Interestingly, it is proved that the sparsity testing problem under unknown noise level is no more difficult than under known noise level for small Δ. For large Δ, the knowledge of σ plays an important role.

(c) We reformulate the sparsity estimation problem as a multiple testing problem where we simultaneously consider all nested hypotheses H_q for $q \in [0,n]$. Introducing a multiple testing procedure which is simultaneously optimal over all q, we derive an estimator $\widehat{k}$ which is less than or equal to $\|\theta^*\|_0$ with high probability and is also closest to $\|\theta^*\|_0$ in a minimax sense. Interestingly, this property is valid uniformly for all possible $\theta^* \in \mathbb{R}^n$ and avoid us to rely on any particular assumption on the parameter. More generally, this perspective also provides a general roadmap to handle the problem of complexity estimation using simultaneous separation distances.

Some aspects of the proof techniques. As explained in Section 2.1.1, the second moment method is particularly suited when we consider a simple null hypothesis test as in signal detection. To prove a minimax lower bound that depends on both k_0 and Δ (as in the regime $k_0 \ge \sqrt{n}$ in Table 2.1), we build two prior distribution μ_0 and μ_1 on $\mathbb{B}_0[k_0]$ and $\mathbb{B}_0[k_0+\Delta]$ respectively and we need to show that the total variation distance between the corresponding marginal distributions $\mathbf{P}_0 = \int \mathbb{P}_\theta \, \mu_0(d\theta)$ and $\mathbf{P}_1 = \int \mathbb{P}_\theta \, \mu_1(d\theta)$ is small. In a seminal work, Lepski et al. [143] have shown that it is possible to control the distance between $\mathbf{P}_0$ and $\mathbf{P}_1$ it terms of the moments of μ_0 and μ_1. More precisely, if we write $m_q(\mu) = \int \theta^q \mu(d\theta)$, then the total variation distance between $\mathbf{P}_0$ and $\mathbf{P}_1$ is small provided that (i) μ_0 and μ_1 have a bounded support and (ii) $m_q(\mu_0) = m_q(\mu_1)$ for all $q = 1, \dots, c\log(n)$ –see also [46] for a related explanation. Then, the challenge is to build suitable measures μ_0 and μ_1 that satisfy (i) and (ii) while being respectively supported on $\mathbb{B}_0[k_0]$ and $\{\theta \in \mathbb{B}_0[k_0+\Delta] : \ d_2(\theta, \mathbb{B}_0[k_0] \ge \rho\}$. This approach is sometimes referred as the moment matching method and turns out to be fruitful in many modern problems [199, 200].

Conversely, our optimal tests rely on some integrals of the empirical characteristic function of Y which, at least in expectation, approximate the non-continuous function $f(\theta) = \sum_{i=1}^{n} \mathbf{1}_{\theta_i \neq 0}$. This differs from typical estimators of non-smooth functionals which are rather based on the so-called polynomial approximation techniques (e.g. [106, 46, 143]). The latter technique amounts to build a best polynomial approximation of the non-smooth function and plug unbiased estimators of the moments into this polynomial.

One possible weakness of our work is that it is not robust against departures from the Gaussian distribution. In fact, the minimax separation distances in Table 2.1 only hold for Gaussian noise. When the noise distribution is unknown (but say sub-Gaussian), our statistics based on the empirical characteristic function of the data are not trustable. We conjecture that the minimax distances should differ by polylogarithmic factors in this setting.

Extension to linear regression models. In the subsequent work [A5], we have considered the sparsity testing problem for the high-dimensional linear regression model (Definition 2.2). We have provided a tight characterization of the separations distances when the covariates are independent ($\mathbf{\Sigma} = \mathbf{I}_p$ in Definition 2.2). For general and unknown covariance matrix $\mathbf{\Sigma}$, we have some partial results. Unfortunately, getting a tight characterization (with the correct logarithm) in all regimes seems out of reach with our current techniques.

2.5 Multiple testing with unknown distribution

This section is mainly based on two joint works [A1, A4] with A. Carpentier, S. Delattre, and E. Roquain. While these contributions are mainly motivated by multiple testing considerations, a significant part of the challenge amounts to estimating some functional in a sparse Gaussian sequence model and to evaluate to what extent the estimation error of these functionals perturbs the testing problem.

Toy multiple testing model. In large-scale data analysis, the practitioner routinely faces the problem of simultaneously testing a large number n of null hypotheses. In the last decades, an impressive amount of multiple testing procedures have been developed –see, e.g., [70]. Theoretically-founded control of the amount of false rejections are provided notably by controlling the false discovery rate (FDR), that is, the average proportion of errors among the rejections, as done by the famous Benjamini-Hochberg procedure (BH), introduced by [19]. A prototypical multiple testing model can be recast in the Gaussian sequence model. Suppose that we observe independent random Y_i, with $i = 1, \dots, n$. For each $i = 1, \dots, n$, we consider the testing problem.

$$H_{0,i} : \{Y_i \sim \mathcal{N}(0, \sigma^2)\} \qquad \text{against} \qquad H_{1,i} : \{Y_i \sim \mathcal{N}(\theta_i^*, \sigma^2) \text{ with } \theta_i^* \neq 0\} . \tag{2.14}$$

Then, the multiple-testing problem is exactly equivalent to variable selection, that is selecting the support $S^* = \{i, \theta_i^* \neq 0\}$ of θ^* in the Gaussian sequence model. Importantly, works in variable selection [40, 41] and in multiple testing (e.g. [170]) differ with respect to the loss function under consideration, the multiple testing literature being more focused on FDR or FWER[6]. A slight variation of (2.14) can also be considered to handle distribution-free alternatives.

$$H_{0,i} : \{Y_i \sim \mathcal{N}(0, \sigma^2)\} \qquad \text{against} \qquad H_{1,i} : \{Y_i \nsim \mathcal{N}(0, \sigma^2)\} . \tag{2.15}$$

It is well known [19] that the celebrated Benjamini-Hochberg procedure controls the FDR in (2.15) and exhibits optimal power in (2.14).

[6] family-wise error rate

Efron's model for multiple testing problems with unknown null distributions. However, the two testing models (2.14,2.15) as well as the FDR controlling procedures developed in the multiple testing literature rely on the fact that the null distribution of the test statistics is known – it corresponds to $\mathcal{N}(0, \sigma^2)$ above. This is in plain contrast with common practice, where the null distribution is often *mis-specified*. This phenomenon, pointed out in a series of pioneering papers by Efron [76, 78, 75, 77] and studied further in [183, 181, 12, 175, 176] is illustrated in Figure 2.1 for four classical datasets. As one can see, the theoretical null distribution $\mathcal{N}(0, 1)$ does not faithfully describe the overall behavior of the measurements. As a result, using this theoretical null distribution into a standard multiple testing procedure (e.g., BH) can lead to an important resurgence of false discoveries. Markedly, this effect is sometimes more severe than simply ignoring the multiplicity of the tests (see [M1]), and thus the benefit of using a multiple testing correction can be lost. One hypothesized reason for this gap between the theoretical and empirical null distributions is data often come from raw measurements that have been "cleaned" via many sophisticated normalization processes in which case the practitioner has no prior belief in the null distribution. Hence, the null distribution is implicitly defined as the "background noise" of the measurements and searching signal in the data boils down to make some assumption on this background (typically Gaussian) and find outliers, defined as items that significantly deviate from the background. This occurs for instance in astrophysics datasets [182, 157, 184].

To address these issues, Efron popularized the concept of *empirical null distribution*, that is, of a null distribution estimated from the data, in the works [76, 78, 75, 77, 79] notably through the *two-group mixture model* and the *local fdr* method. Therein, an important message is that a significant improvement can already be obtained by replacing the theoretical null $\mathcal{N}(0, 1)$ by a Gaussian $\mathcal{N}(\theta, \sigma^2)$ with unspecified scaling parameters θ and σ. This type of techniques is widely used nowadays, mostly in genomics [4, 121, 205, 62] but also in other applied fields, such as neuro-imaging, see, e.g., [140]. However, when available, the theoretical FDR controlling properties often rely on stringent assumptions on the underlying mixture model (parameters fixed with n, specific alternatives and existence of suitable parameter estimators), which are not met in general.

In [A4] and [A1], we introduce frequentist counterparts of Efron's model, which are more in line with (2.14) and (2.15). The situations are quite contrasted depending on whether we focus on two-sided hypotheses or on one-sided alternatives.

2.5.1 One-side tests and minimum effect estimation [A4]

Two one-sided contamination models. In the one-sided case where observations Y_i under the alternative hypothesis tend to take higher values than under the null, defining an identifiable model is much easier than in the general problem. In fact, the counterpart of (2.14) corresponds to a model

$$Y_i = \mu^* + \gamma_i^* + \epsilon_i, \quad 1 \leq i \leq n \ , \tag{2.16}$$

where the ϵ_i's are i.i.d. $\mathcal{N}(0, \sigma^2)$ distributed, whereas $\mu^* \in \mathbb{R}$ and $\gamma^* \in \mathbb{R}_+^n$ are unknown. Upon defining the mean vector $\theta^* = \mu^* + \gamma^*$, we come back to the Gaussian sequence model, where $\mu^* = \min_i \theta_i^*$ is the minimum coordinate of the vector θ^*. Equipped with (2.16), the multiple-testing hypotheses writes as

$$\underline{H}_{0,i} : \{\theta_i^* = \mu^*\} \quad \text{against} \quad \underline{H}_{1,i} : \{\theta_i^* > \mu^*\} \ .$$

Observe that the null distribution corresponding $\underline{H}_{0,i}$ is unknown as the minimum θ_i^* is unknown. Within this model, the number of alternative hypotheses (also called the number of contaminated

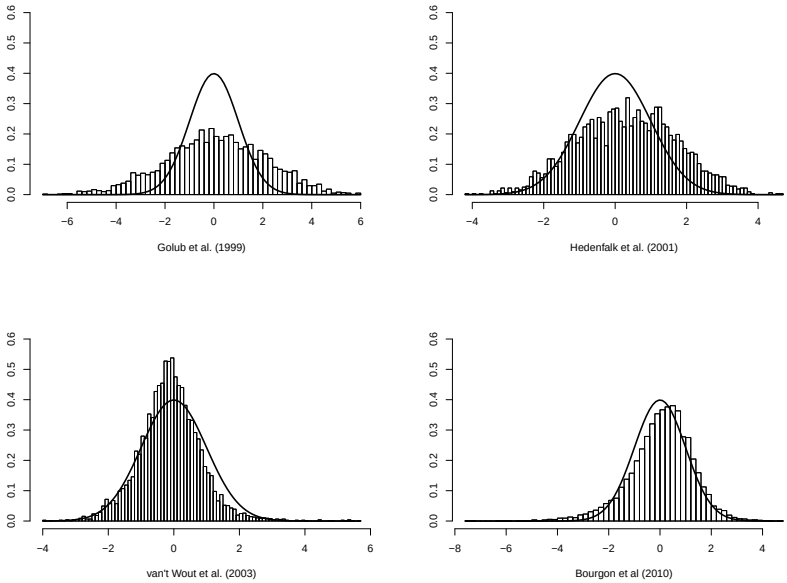

Figure 2.1: Histograms of the test statistics (rescaled to be all marginally standard Gaussian), for three datasets presented by Efron: [103] (top-left); [107] (top-right); [197] (bottom-left); and [31] (bottom-right). The solid curve is the standard Gaussian density. Pictures reproducible from the vignette [M1].

data) equals the number of non-zero entries of γ^*, that is its sparsity. For this purpose, we consider, for an integer $k \in [0, n-1]$,

$$\mathcal{M}_k = \left\{ \gamma^* \in \mathbb{R}_+^n \ : \ \sum_{i=1}^n \mathbf{1}_{\{\gamma_i^* > 0\}} \leq k \right\} \ . \tag{2.17}$$

In what follows, we refer to the model (2.16) as *Gaussian One-Sided Gaussian Contamination (gOSC) model*. We focus on the following related twin problems

(i) Estimating the null distribution, that is estimating the minimum functional $f(\theta^*) = \min_i \theta_i^* = \mu^*$ in the model (2.16). In that respect, the alternative hypotheses play the role of a nuisance parameter. More precisely, we want to characterize the minimax convergence risk

$$\mathcal{R}[k, n] = \inf_{\widehat{\mu}} \sup_{(\mu^*, \gamma^*) \in \mathbb{R} \times \mathcal{M}_k} \mathbb{E}_{\mu^*, \gamma^*} \left[|\widehat{\mu} - \mu^*| \right] \ , \tag{2.18}$$

as well as exhibiting optimal procedures.

(ii) Plugging such a suitable estimator $\widehat{\mu}$ of the null distribution into a BH procedure and evaluating its performances.

Furthermore, we consider a more general *One-Sided Contamination (OSC) model*, where the hypotheses are of the form $\underline{H}'_{0,i} : \{Y_i \sim \mathcal{N}(\mu^*, \sigma^2)\}$ against $\underline{H}'_{1,i} : \{Y_i \succ \mathcal{N}(\mu^*, \sigma)\}$. Here, $\succ$ stands for the stochastic domination. In that setting, $\mu^* = \min_i \mathbb{E}[Y_i]$. As we defined the minimax estimation risk $\mathcal{R}[k, n]$ for (gOSC), we define its counterpart $\overline{\mathcal{R}}[k, n]$ in the OSC model.

Minimax estimation of μ^*. First, we characterize these two minimax risks by deriving matching (up to numerical constants) lower and upper bounds, this uniformly over all numbers k of contaminated data. The results are summarized in Table 2.2 below. For $k \leq \sqrt{n}$, the rate is parametric in both models. For $k \in (\sqrt{n}, n/2)$, one-sided contaminations lead to some $\sqrt{\log(k^2/n)}$ gain over the risk k/n we could have expected as k is the number of nuisance parameters. Assuming that the contaminations are Gaussian leads to an additional logarithmic gain. For $k \in [n/2, n-1]$, the minimax risks are more intricate, but the optimal rate still converges to 0 even as the proportion $\frac{n-k}{n}$ of non-contaminated samples goes to 0 slowly enough. In other words, it is still possible to estimate consistently the minimum value μ^* (or equivalently the null) even if a really tiny proportion of the observations follows the null hypothesis. Let us make a few comments on the mathematical techniques. The minimax lower bound for gOSC is based on moment matching techniques as in the previous section, whereas the one for OSC relies on techniques pertaining to robust estimation. The main challenge is to introduce an optimal test for gOSC. In a nutshell, we first compute the empirical Laplace transform of the data that we plug into a large collection of Chebychev's polynomials.

	General bound	$1 \leq k \leq 2\sqrt{n}$	$2\sqrt{n} \leq k \leq n/2$	$n/2 \leq k \leq n-1$
$\overline{\mathcal{R}}[k,n]$	$\frac{\log\left(\frac{n}{n-k}\right)}{\log^{1/2}(1+\frac{k^2}{n})}$	$n^{-1/2}$	$\frac{k/n}{\log^{1/2}(k^2/n)}$	$\frac{\log\left(\frac{n}{n-k}\right)}{\log^{1/2} n}$
$\mathcal{R}[k,n]$	$\frac{\log^2\left(1+\sqrt{\frac{k}{n-k}}\right)}{\log^{3/2}\left(1+(\frac{k}{\sqrt{n}})^{2/3}\right)}$	$n^{-1/2}$	$\frac{k/n}{\log^{3/2}(k^2/n)}$	$\frac{\log^2\left(\frac{n}{n-k}\right)}{\log^{3/2} n}$

Table 2.2: Minimax estimation risks of μ^*.

Application to multiple testing with unknown parameters. Let us now come back to Efron's model. In [A4], we show that some minor modification of the quantile-based estimators $\widehat{\mu}$ and $\widehat{\sigma}$[7] introduced for OSC model, can be used to estimate the null distribution and then to rescale the p-values. We can then be suitably combined with classical multiples testing procedures such as Benjamini-Hochberg (BH) procedure. As long as the proportion of true alternative hypotheses is bounded away from 1 (say it is smaller than 0.9), then the empirical procedure enjoys the same nice properties as the oracle BH procedure that would know the parameters in advance: the false discovery rate (FDR) and the true discovery proportion (TDP) are similar for both procedures.

2.5.2 General alternatives and FDR control under unknown null distribution [A1]

In [A1], we consider a more general two-sided multiple testing model with unknown null distribution. More formally, we observe a random vector $Y = (Y_i)_{1 \leq i \leq n}$ in $\mathbb{R}^n$ whose distribution is denoted by $P = \otimes_{i=1}^n P_i$. For identifiability purpose, we assume that more than half of the P_i's follow the same (null) distribution Gaussian distribution while the others are "contaminated" and can be arbitrary.

[7] we also consider the case of unknown variance in [A4]

Hence, the null distribution parameters (μ^*, σ^*) are defined as the unique parameters such that

$$n_0(P) = \sum_{i=1}^{n} \mathbf{1}_{P_i = \mathcal{N}(\mu^*, \sigma^{*2})} > n/2 \ .$$

Our testing problem is then a counterpart of (2.15) where the null distribution is unknown. This provides a general and simple setting to address the following question:

When the null distribution is unknown, is it possible to build a procedure that both controls the FDR at the nominal level and has a power asymptotically mimicking the oracle?

Here, what we call the oracle procedure corresponds to the classical Benjamini-Hochberg (BH) procedure the statistician would have carried out if an oracle had given them the true values μ^* and σ^*.

On the feasibility side, estimating μ^* and σ falls into the classical problem of estimating the mean and the variance in a variation of Huber's contamination model and is well understood. For instance, the empirical median $\widehat{\mu}$ is well known to be optimal. Similarly, $\widehat{\sigma}$ is to be estimated by a combination of empirical quantiles. Here, the challenge is to study how the estimation error of $(\widehat{\mu}, \widehat{\sigma})$ propagates when using the rescaled p-values $\hat{p}_i$ based on these parameters instead of the oracle p-values p_i^* one would compute if the null was known in advance. The main results are summarized in the next theorem.

Theorem 2.4. *Consider an asymptotic setting where n goes to infinity. Let k_n be an upper bound of the number of true discoveries (aka contaminations)*

(i) for a sparsity parameter $k_n \gg n/\log(n)$, there exists no sequence of procedures asymptotically mimicking the oracle.

(ii) for a sparsity parameter $k_n \ll n/\log(n)$, the sequence of plug-in BH procedures $(BH_\alpha(\widehat{\mu}, \widehat{\sigma}))_{\alpha \in (0,1)}$ is asymptotically mimicking the oracle, for the scaling $(\widehat{\mu}, \widehat{\sigma})$ given by standard robust estimators.

Contrary to one-sided alternatives, it is now only possible to build a procedure that performs a well as the oracle only if the proportion of true discoveries (contaminations) is smaller than $1/\log(n)$. When the proportion of contaminations is higher than $1/\log(n)$, then the multiple testing procedure either suffers from a high FDR or is much more conservative than the oracle BH procedure. As an extension, we also consider in [A1] general null distributions (beyon the Gaussian case) that are known up to a location parameter μ^*.

2.6 Schatten norm estimation of rectangular matrices

To conclude this chapter, we move away from the Gaussian sequence and the high-dimensional linear regression models to matrix problems. The following material is based on the joint work [P2] with my PhD student Solène Thépaut. Consider the signal plus noise model. We observe a matrix $\mathbf{Y}$ such that

$$\mathbf{Y} = \mathbf{A} + \mathbf{E} \ , \tag{2.19}$$

Here, $\mathbf{A}$ stands for the $p \times q$ unobserved signal matrix and $\mathbf{E}$ is a $p \times q$ noise matrix with independent entries following a standard normal distribution. Without loss of generality, we assume that $p \geq q$. Data of this form arise in many statistical problems such as network analysis (Chapter 3)

or clustering (Chapter 4). In the previous sections, we emphasized the key role of sparsity for estimating or testing a vector θ^*. Although entry-wise sparsity is relevant in some matrix problems, low-rank properties on $\mathbf{A}$ are underlying many statistical procedures. For instance, PCA or more generally singular value thresholding methods (see e.g. [51, 95]) are proved to estimate well low-rank matrices $\mathbf{A}$. Besides, when the matrix $\mathbf{Y}$ is sampled from a Gaussian mixture model with K groups or from Stochastic block model with K groups, then the rank of $\mathbf{A}$ is at most K.

In light of the ubiquity of the low-rank assumption, we focus our attention on estimating specific quantities of $\mathbf{A}$ that are related to its rank. On the one hand, checking the rank or the effective rank of $\mathbf{A}$ allows to assess the relevance of low-rank based procedures. On the other hand, evaluating the rank (or the effective rank) of $\mathbf{A}$ may also be an objective per se to characterize the complexity of the signal matrix $\mathbf{A}$. For the specific case of stochastic block models (defined in Chapter 4), there are significant works that aim at estimating/testing the number of groups– see e.g. [52, 141]. More generally, estimating the rank of a noisy matrix can help selecting the number of components in PCA [58, 125].

In [P2], we do not aim at estimating/testing the exact rank of $\mathbf{A}$, but we rather focus on some measures of effective ranks. Before introducing these quantities and explaining their interest, we need to define the Schatten norm of a matrix. Given a $p \times q$ matrix $\mathbf{A}$, we write $\kappa_1(\mathbf{A}) \geq \kappa_2(\mathbf{A}) \ldots \geq \kappa_q(\mathbf{A}) \geq 0$ for its ordered sequence of singular values. For any $s \geq 1$, the s-Schatten norm of $\mathbf{A}$ is defined as the l_s norm of its sequence of singular values, that is

$$\|\mathbf{A}\|_s^s = \sum_{i=1}^{q} \kappa_i^s(\mathbf{A}) \ . \tag{2.20}$$

For $s = \infty$, we define $\|\mathbf{A}\|_\infty = \kappa_1(\mathbf{A})$ as the operator norm of $\mathbf{A}$. Other classical examples involve the Frobenius norm ($\|\mathbf{A}\|_2$) or the trace norm ($\|\mathbf{A}\|_1$). Let us briefly explain how these Schatten norms are related to effective rank measures.

Effective Rank of a matrix. Since the rank of a matrix is very sensitive to small perturbations, it is difficult to estimate it from $\mathbf{Y}$. Furthermore, a full rank matrix $\mathbf{A}$ may have only few large singular values together with many small singular values. For such a matrix, the rank is poorly informative of the structure of $\mathbf{A}$. As an alternative, various notions of effective ranks have been introduced. In particular, some of these notions of effective rank are at the heart of high-dimensional or infinite-dimensional probabilistic results. For instance, in [131], Koltchinskii and Lounici consider, for a non-negative symmetric matrix $\mathbf{\Sigma}$, the measure $\mathrm{tr}[\mathbf{\Sigma}]/\|\mathbf{\Sigma}\|_\infty = \|\mathbf{\Sigma}\|_1/\|\mathbf{\Sigma}\|_\infty$. For a rectangular matrix $\mathbf{A}$, one can think of two extensions of this index, depending on whether we work directly with the singular values of $\mathbf{A}$ or with the singular values of the square matrix $\mathbf{A}^T\mathbf{A}$.

$$\mathrm{ER}_{1,\infty}(\mathbf{A}) = \frac{\|\mathbf{A}\|_1}{\|\mathbf{A}\|_\infty} \quad ; \qquad \mathrm{ER}_{2,\infty}(\mathbf{A}) = \mathrm{ER}_{1,\infty}(\mathbf{A}^T\mathbf{A}) = \frac{\|\mathbf{A}\|_2^2}{\|\mathbf{A}\|_\infty^2} \ . \tag{2.21}$$

More generally, if we borrow the formalism of diversity measures in ecology [126], we can introduce, for any positive $s > 0$ different from 1, the Hill's effective number [126] of singular values, which interprets as a Renyi entropy of the sequence of singular values

$$\mathrm{ER}_{1,s}(\mathbf{A}) = \left(\frac{\|\mathbf{A}\|_s}{\|\mathbf{A}\|_1}\right)^{s/(1-s)} ; \qquad \mathrm{ER}_{2,s}(\mathbf{A}) = \mathrm{ER}_{1,s}(\mathbf{A}^T\mathbf{A}) = \left(\frac{\|\mathbf{A}\|_{2s}}{\|\mathbf{A}\|_2}\right)^{2s/(1-s)} . \tag{2.22}$$

When all the non-zero singular values of $\mathbf{A}$ are equal, all these effective rank indices are equal to the true rank of $\mathbf{A}$. However, these measures differ in the way they treat heterogeneous values

for the singular values. In short, smaller values of s in $\mathrm{ER}_{1,s}(\mathbf{A})$ and $\mathrm{ER}_{2,s}(\mathbf{A})$ are more prone to take into account smaller singular values in the effective rank. See [126] and references therein for further discussions.

Our results. As a warm-up, we consider the Frobenius norm $\|\mathbf{A}\|_2$ and we prove that the simple quadratic estimator $(\|\mathbf{Y}\|_2^2 - pq)_+^{1/2}$, where $x_+ = \max(x, 0)$ achieves the optimal risk $(pq)^{1/4}$. Interestingly, this risk $(pq)^{1/4}$ cannot be improved by any estimator even if the matrix $\mathbf{A}$ is additionally known to be of rank at most one. Then, we establish that a non-linear transformation of $\kappa_1(\mathbf{Y})$ estimates $\|\mathbf{A}\|_\infty = \kappa_1(\mathbf{A})$ with the same optimal error $(pq)^{1/4}$.

Regarding general even norms $\|\mathbf{A}\|_{2k}$ where k is any integer, we first remark that $\|\mathbf{A}\|_{2k}^{2k} = tr[(\mathbf{A}^T\mathbf{A})^{2k}]$ is a polynomial with respect to the entries of $\mathbf{A}$. This allows us to build an unbiased estimator U_k of $\|\mathbf{A}\|_{2k}^{2k}$ based on Hermite polynomials. Relying on the invariance of $\|\mathbf{A}\|_{2k}^{2k}$ by left and right orthogonal transformations, we establish that this estimator has a simple expression as an algebraic combination of monomials of the form $tr[(\mathbf{YY})^l]$ so that the estimator U_k can be efficiently computed. One of our main result is a general variance upper bound for U_k, which allows us to prove that the estimator $(U_k)_+^{1/(2k)}$ achieves the optimal risk $(pq)^{1/4}$ uniformly over all matrices $\mathbf{A}$.

Regarding general norms $\|\mathbf{A}\|_s$ where $s \geq 1$ is not an even integer, we first exhibit a simple plug-in estimator based on a linear transformation of the empirical singular values $(\kappa_i(\mathbf{Y}))$ that achieves a much higher error of the order $q^{1/s}(pq)^{1/4}$ (compare with $(pq)^{1/4}$ for even Schatten norms). Our second main result is a minimax lower bound of order $\frac{q^{1/s}}{\log^s(q)}(pq)^{1/4}$ stating that it is impossible to estimate non-even Schatten norms at a much faster rate. Using polynomial approximation techniques that approximate $\|\mathbf{A}\|_s$ by a linear combination of even Schatten norms $\|\mathbf{A}\|_{2k}^{2k}$, we are able to close this logarithmic gap between our upper and lower minimax bounds.

Finally, we extend our analysis to a signal plus noise model with a general subGaussian noise distribution. Quite surprisingly, we establish that the convergence rates of our estimators remain almost unchanged, despite the fact that the definition of U_k heavily depends on the sequence of moments of the normal distribution. However, the analysis turns out to be much more involved in comparison to Gaussian case. Finally, we are able to come back to the initial problem of effective rank estimation and to construct suitable optimal estimators.

In this work [P2], we characterized the minimax risk for Schatten norms and effective ranks estimation. We argued why estimating the effective rank of a matrix can be more appealing in some situations than estimating its ranks. Still, testing hypotheses of the form $\{\mathrm{Rank}(\mathbf{A}) \leq k\}$ against $\{\mathrm{Rank}(\mathbf{A}) > k\}$ remains a really interesting problem for the future.

Open Problem 2.1. *Establishing the minimax separation distances for rank test as we did for the sparsity testing problem [A8].*

Chapter 3

Network Analysis

This chapter is dedicated to the statistical analysis of network data although some of the contributions specifically pertaining to clustering are postponed to chapter 4. We focus our attention to two contrasted problems: community detection [A3, A20, A21] and graphon estimation [A9, A16].

In both cases, we observe the adjacency matrix $\mathbf{A}$ of an undirected random graph with n nodes. Hence, $\mathbf{A} \in \{0,1\}^{n\times n}$ is a symmetric matrix whose terms on the diagonal are all equal to zero. The problem of community detection is that of testing whether the network is homogeneous (in some sense to be defined) or if there exists a set of nodes that are unusually connected. With the formalism of the previous chapter, this interprets as a signal detection problem in a random graph. Detecting the existence of a signal is, in some way, the least ambitious objective that we can aim for. In [A3, A20, A21], we precisely pinpoint the minimum expected density of the community (a.k.a. the minimum signal strength) which makes it detectable by a suitable test. Whereas for dense graphs, the mathematical tools and the desired bounds do not differ much from comparable problems in Gaussian noise, the analysis of sparse graphs requires specific probabilistic tools.

Moving away from community detection, we also consider the problem of estimating the distribution of $\mathbf{A}$ only assuming that this distribution is invariant by any permutation of the nodes. This corresponds to assuming that $\mathbf{A}$ is sampled from a graphon [150]. Our goal is therefore to estimate this graphon[1]. In [A16], we characterize the minimax estimation of rate of sparse graphons with respect to the δ_2-distance, which interprets as a counterpart of the Frobenius distance for matrices. In [A9], we consider graphon estimation with respect to the cut distance, which is more grounded in graph theory. This second contribution is rather appealing from a graph-theoretic and probabilistic point of view, as trivial estimators turn out to be optimal.

3.1 Community Detection

This section is mostly dedicated to two joint contributions with E. Arias-Castro [A20, A21]. Observing an $n \times n$ adjacency matrix, we consider a toy testing problem where we want to decipher whether $\mathbf{A}$ is sampled from an Erdös-Renyi random graph with connection probability p_0 or if, for some subset $S \subset [n]$ of size m, the connection probability between nodes in S equals $p_1 > p_0$, in which case S is referred as the community. We focus on situations where the community size m, if it exists, is much smaller than n. The specific instance of this problem where $p_0 = 1/2$ and $p_1 = 1$ is a randomized version of the planted clique problem and is central in complexity theory [7]. Besides, this average planted clique problem has been recently used for establishing computational

[1] The definition of graphon is postponed to Section 3.2

lower bounds for statistical problems – see the seminal work [20] on sparse PCA. Here, we consider the more general version of community detection where p_0 and p_1 are arbitrary.

More generally, our problem fits within the more general field of model testing for random graphs. This encompasses the problem of detecting a latent geometry [37] or testing the number of groups of a stochastic block models [91, 14, 25, 141] and its generalization [123]. The settings and mathematical tools in our works significantly differ from the latter references mainly because we focus on a regime where the size m of the community is much smaller than n. Nevertheless, there exist some connections between our problem and that of recovering the communities in a stochastic block models with a large $K = n/m$ number of communities [33, 54].

Let us define our problem with the minimax formalism of the previous chapter. We consider an asymptotic framework where $m, n \to \infty$, and p_0, p_1 may also change. Given a test T, we consider as before the maximum risk

$$R_n(T) = \mathbb{P}_0(T = 1) + \max_{|S|=m} \mathbb{P}_S(T = 0) \ , \tag{3.1}$$

where $\mathbb{P}_0$ is the distribution under the null and $\mathbb{P}_S$ is the distribution under the alternative where S indexes the community. We say that a sequence of tests (T_n) for a sequence of problems $(\mathbf{A}_n)$ is asymptotically powerful (resp. powerless) if $R_n(T_n) \to 0$ (resp. $\to 1$). Practically speaking, a sequence of tests is asymptotically powerless if it does not perform substantially better than any guessing that ignores the adjacency matrix $\mathbf{A}$. In [A20, A21], we establish the fundamental statistical (information theoretic) difficulty of detecting a community in a network by providing the conditions on (n, m, p_0, p_1) under which there exist asymptotically powerful tests.

3.1.1 Dense regime

We first focus on the quasi-normal regime where mp_0 is either bounded away from zero, or tends to zero slowly, specifically,

$$\log\left(1 \vee \frac{1}{mp_0}\right) = o\left[\log\left(\frac{n}{m}\right)\right] \ . \tag{3.2}$$

It encompasses the case where p_0 is constant. Because of its importance in describing the tails of the binomial distribution, the relative entropy or Kullback-Leibler divergence of $\mathrm{Bern}(p_1)$ to $\mathrm{Bern}(p_0)$ — appears our results. $H_{p_0}(p_1) = p_1 \log(\frac{p_1}{p_0}) + (1 - p_1) \log(\frac{1-p_1}{1-p_0})$

Theorem 3.1. *Assuming that $m \gg \log(n)$ and (3.2) hold, all tests are asymptotically powerless if*

$$\frac{p_1 - p_0}{\sqrt{p_0}} \frac{m^2}{n} \to 0, \quad \textit{and} \quad \limsup \frac{mH_{p_0}(p_1)}{2\log(n/m)} < 1. \tag{3.3}$$

Conversely, if any of the two conditions is not satisfied, then it is possible to build a powerful test. The test is the combination of the two natural tests that arise in the related problem of submatrix detection [39] and much of the work in that field [6, 44, A28]. First, the total degree test rejects for large values of the total number of edges in the graph $A^* := \sum_{1 \le i < j \le n} \mathbf{A}_{i,j}$. Through Chebyshev inequality, one easily show that the test based on A^* is asymptotically powerful if $\frac{p_1 - p_0}{\sqrt{p_0}} \frac{m^2}{n} \to \infty$. Second, we consider the scan test that rejects for large values of $A^*_m := \max_{|S|=m} A_S$ where $A_S := \sum_{i,j \in S, i<j} \mathbf{A}_{i,j}$ is the number of edges in the graph induced by S. By Chernoff's bound together with an union bound, we derive that this test is asymptotically powerful if $\liminf \frac{mH_{p_0}(p_1)}{2\log(n/m)} > 1$.

Unknown p_0 and adaptation to unknown m. We also consider the situation, common in practice, where p_0 is unknown. We derive the corresponding lower bound in this situation and design a combination of two tests that achieve this bound. While the scan test can be easily adapted to unknown by plugging a suitable estimator of p_0, this is not the case the total degree test. For this reason, we have to work with the second moment of the degree distribution to craft a so-called *degree-variance test*.

On computational-statistical trade-offs. In the setting where $m \gg n^{2/3}$ for known p_0, and $m \gg n^{3/4}$ for unknown p_0, this detection boundary is achieved by the total degree test and the degree variance test, respectively, which can be computed in polynomial-time. Otherwise, there is a large discrepancy between the information theoretic detection boundary, achieved by the scan test, and what polynomial tests are shown to achieve, which is not surprising since average planted clique is a specific instance of our problem. Subsequently to our work, there has been an effort towards establishing computational lower bounds for community detection by a reduction to the planted clique regime ($p_0 = 1/2$, $p_1 = 1$) – see [104] and recent series of works of Brennan and Bresler [33, 36].

On the proof arguments. Following the general strategy for proving impossibility results in test problems (see Chapter 2), we may be tempted to use a second moment strategy. For $N = \binom{n}{m}$, define the mixture distribution $\mathbb{Q} = N^{-1} \sum_{S:|S|=m} \mathbb{P}_S$. Write the likelihood $L = N^{-1} \sum_{S:|S|=m} L_S$ where $L_S = d\,\mathbb{P}_S / d\,\mathbb{P}_0$. We need to prove that that the total variation distance, or equivalently, $\mathbb{E}_0[|L-1|]$ goes to zero. If we use, as it is often case, Cauchy-Schwarz inequality and bound the second moment $\mathbb{E}_0[(L-1)^2]$, this will unfortunately lead to a loose conditions for impossibility of detection. For this reason, we have to rely on a adaptive truncated second moment. This argument was introduced, up to our knowledge, by Ingster in the 90's (see also [39]) and we consider an adaptive version here. The idea is to introduce, for each S, and event Γ_S and to work with the truncated likelihood $\tilde{L} = N^{-1} \sum_{S:|S|=m} \tilde{L}_S$ where $\tilde{L}_S = L_S \mathbf{1}_{\overline{\Gamma_S}}$, where $\overline{\Gamma_S}$ is the complementary event of Γ_S. Then, since $\tilde{L} \le L$, simple algebra and the definition of L_S leads to

$$\mathbb{E}_0[|L-1|] \le \mathbb{E}_0[L-\tilde{L}] + \mathbb{E}_0^{1/2}[(1-\tilde{L})^2] = N^{-1} \sum_{S:|S|=m} \mathbb{P}_S(\Gamma_S) + \Big[\mathbb{E}_0[\tilde{L}^2] - 1 + \frac{2}{N} \sum_{S:|S|=m} \mathbb{P}_S(\Gamma_S) \Big]^{1/2}.$$

With this bound in mind, it suffices to choose an event Γ_S with small probability so that $\mathbb{P}_S(\Gamma_S) = o(1)$ (uniformly in S) while enforcing the second truncated moment $\mathbb{E}_0[\tilde{L}^2]$ to be close to one. In this relatively dense situation, we choose a complementary event $\overline{\Gamma}_S = \cap_{\ell=1}^m \sup_{T \subset S,\ |T|=l} A_T \le \zeta(\ell)$ for some suitable function ζ. In other words, we remove the event of small probability, where a subgraph induced by some $T \subset S$ has a unusually high density.

3.1.2 Sparse regime

In a subsequent work [A20] with E. Arias-Castro, we focus on the *sparse* regime where $p_0 \le \frac{1}{m} \left(\frac{m}{n}\right)^{c_0}$ for some constant $c_0 > 0$. Here, sparse implies that $mp_0 \le 1$. In those sparse regimes, a combination of the scan test and the total degree test is still asymptotically powerful when the condition (3.3) is not satisfied. However, it turns out that it is possible to detect the presence of the subgraph below the threshold (3.3). Informally, the main reason for this is that sparse Erdos-Renyi random graphs are much less homogeneous that their counterparts, so that it becomes possible to detect a subtle signature of the signal by looking at the geometry of the graph.

In these regimes, it is helpful to parametrize our problem with $\lambda_0 = np_0$ and $\lambda_1 = mp_1$ and $\kappa \in (0,1)$ such that $m = n^\kappa$. Note that these quantities λ_0, λ_1, and κ may vary with n. We illustrate our results in two emblematic regimes. They can be summarized as follows.

Regime 1: $\lambda_0 = (n/m)^\alpha$ with fixed $0 < \alpha < 1$. Compared to the setting in our previous work [A21], the total degree test remains a contender, scanning over subsets of size exactly m as in A^*_m does not seem to be optimal anymore, all the more so when p_0 is small. Instead, we scan over subsets of a wider range of sizes, using

$$A^\ddagger_n = \sup_{\ell=m/u_n}^{m} \frac{A^*_k}{k} \,, \tag{3.4}$$

where $u_n = \log\log(n/m)$. We call this the broad scan test. In analogy with the dense case [A21], we find that a combination of the total degree test and the broad scan test based on (3.4) is asymptotically optimal in the following sense. When $\kappa > \frac{1+\alpha}{2+\alpha}$, the total degree test is asymptotically powerful when $\lambda_1 \gg \frac{n^{(1+\alpha)/2}}{m^{1+\alpha}}$ and the two hypotheses merge asymptotically when $\lambda_1 \ll \frac{n^{(1+\alpha)/2}}{m^{1+\alpha}}$. When $\kappa < \frac{1+\alpha}{2+\alpha}$, that is for smaller m, there exists a sequence of increasing functions ψ_m such that the broad scan test is asymptotically powerful when $\liminf(1-\alpha)\psi_m(\lambda_1) > 1$ and the hypotheses merge asymptotically when $\limsup(1-\alpha)\psi_m(\lambda_1) < 1$. In summary, we establish the existence of a sharp threshold $\psi_m(.)$ for detection[2] in this sparse regime.

Regime 2: $\lambda_0 > 0$ and $\lambda_1 > 0$ are fixed. The poissonian regime where λ_0 and λ_1 are assumed fixed is depicted on Figure 3.1. When $\lambda_1 > 1$, the broad scan test is asymptotically powerful. When $\lambda_0 > e$ and $\lambda_1 < 1$, no test is able to fully separate the hypotheses. In fact, for any fixed (λ_0, λ_1) a test based on the number of triangles has some nontrivial power (depending on (λ_0, λ_1)), implying that the two hypotheses do not completely merge in this case or more precisely that the total variation distance between $\mathbb{P}_0$ and $\mathbb{Q}$ (as defined in the previous subsection) is bounded away from 0 and 1. The case where $\lambda_0 < e$ is not completely settled. No test is able to fully separate the hypotheses if $\lambda_1 < \sqrt{\lambda_0/e}$. The largest connected component test is optimal up to a constant when $\lambda_0 < 1$ and a test based on counting subtrees of a certain size bridges the gap in constants for $1 \le \lambda_0 < e$, but not completely. When λ_0 is bounded from above and $\lambda_1 = o(1)$, the two hypotheses merge asymptotically.

On the proof techniques. In comparison to the dense case, both the feasibility and the impossibility results are more challenging. In particular, the second regime (λ_0 and λ_1 fixed) requires to control the behavior of multi-type poissonian branching processes to understand the geometry of the connected components and the number of subtrees of the graph under the alternative hypothesis. Along these lines, the corresponding events Γ_S in the the truncated second moment have to be carefully crafted in terms of the number of large subtrees inside the subgraph induced by S.

Open Problem 3.1. *Establish the tight detection threshold in Figure 3.1 in the regime $\lambda_0 < e$ and $\lambda_1 < 1$.*

3.1.3 Extension to inhomogeneous random graphs

While the above statistical problem is elegant, one may opt that the real-world networks are much less homogeneous than a typical realization of an Erdos-Renyi random graph. In particular, the

[2] Although the exact form of this threshold $\psi_m(.)$ is implicit, we show that detection occurs at $\lambda_1 \asymp (1-\alpha)^{-1}$.

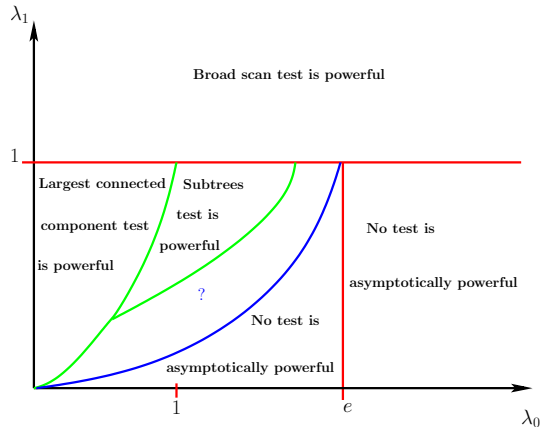

Figure 3.1: Detection diagram in the poissonian asymptotic where λ_0 and λ_1 are fixed and $m = n^\kappa$ with $0 < \kappa < 1/2$.

degree distribution of nodes in real networks is far from that of binomial distribution and may exhibit for instance power laws [163].

In a recent joint work [A3] with K. Bogerd, R. Castro, and R. van der Hofstad, we partially extend our results to the more general problem of detecting a small community in an already inhomogeneous random graph. In a size n inhomogeneous random graph, we are given a symmetric matrix $\mathbf{p} = (p_{ij}) \in [0,1]^{n\times n}$ and each edge is sampled independently with probability p_{ij}, that is $\mathbb{P}_0[\mathbf{A}_{ij} = 1] = p_{ij}$. Here we consider the scenario where, under the null hypothesis, the edge probabilities p_{ij} have a so-called rank-1 structure. That is, we assume that each vertex $i \in [n]$ is assigned a weight $\theta_i \in (0,1)$ and that the edge probabilities are given by $p_{ij} = \theta_i\theta_j$. This is probably one of the simplest models for inhomogeneous random graphs possible. Note that this model is very similar to the degree-corrected stochastic block model [206], except that, under the null, there is only only single group. Furthermore, there are strong connections between this null model and the configuration model [163]. Under the alternative, there exists a subset $S \subset [n]$ of size m such that the connection probability between two nodes of S is higher by a factor $\rho_S > 1$, that is $\mathbb{P}_S[\mathbf{A}_{ij} = 1] = \rho_S\theta_i\theta_j$ if $i, j \in S$ where $\mathbb{P}_S[A_{ij} = 1] = \theta_i\theta_j$ if $i \notin S$ or $j \notin S$.

The detection problem now amounts to deciphering whether an adjacency matrix $\mathbf{A}$ has been sampled according to $\mathbb{P}_0$ or to $\mathbb{P}_S$ for some $S \subset [n]$, without knowing in advance the weights θ_i. Provided that (i) the community size m is small enough and (ii) the heterogeneity $\max_i \theta_i / \min_i \theta_i$ is mild, we are able to identify the threshold ρ_S^* at which it is possible to distinguish both hypotheses. The corresponding optimal test is a variation of the scan test that accounts for the heterogeneity of the connection probabilities.

3.2 Graphon Estimation

Whereas the previous section focused on parametric random graph model with few parameters, we move to non-parametric models of random graphs. This section is based on two joint works [A9, A16] with O. Klopp and A. Tsybakov. In the 2000's, Lovász and many of his coauthors (see e.g. [150, 151]) have introduced graphons as the limit of a graph sequences. Quite interestingly, Aldous-Hoover-Kallenberg's representation theorem [150, 66] implies that any random graph distribution

can be represented by a graphon, through the use of W-random graphs defined below. For those reasons, one may interpret graphons as an universal class for random graph distribution.

There are several equivalent ways of defining graphons [150]. Here, we consider a graphon as a symmetric measurable functions $W : [0,1]^2 \to [0,1]$. Henceforth, we write $\mathcal{W}$ for the collection of all possible graphons. Given a graphon W_0, a *W-random graph* with n nodes is sampled as follows: first, one samples $\xi_1, \ldots, \xi_n$ are unobserved (latent) i.i.d. random variables uniformly distributed on $[0,1]$. Those represent the hidden labels of those n nodes. Then, given those labels, we build the $n \times n$ symmetric probability matrix $\mathbf{\Theta}_0$ by $(\mathbf{\Theta}_0)_{ij} = W_0(\zeta_i, \zeta_j)$ if $i \neq j$ and $(\mathbf{\Theta}_0)_{ii} = 0$. Finally, we sample the symmetric adjacency matrix $\mathbf{A}$: the observations $\mathbf{A}_{ij}$ for $1 \leq j < i \leq n$ are assumed to be independent Bernoulli random variables with success probabilities $(\mathbf{\Theta}_0)_{ij}$.

For a fixed graphon W_0, the expected number of edges of the corresponding W-random graph is of the order of $0.5n^2 \int W_0(x,y)dxdy \asymp n^2$. To better handle sparse networks, it has been proposed to modify this model by the introduction of a sparsity parameter ρ_n. In this sparse W-random graph model, the probability matrix $\mathbf{\Theta}_0$ is now built as

$$(\mathbf{\Theta}_0)_{ij} = \rho_n W_0(\xi_i, \xi_j) \quad \text{for } i \neq j \tag{3.5}$$

where $\rho_n > 0$ is the scale parameter that can be interpreted as the expected proportion of non-zero edges. Then, the adjacency matrix $\mathbf{A}$ is sampled as above as an inhomogeneous random graph with probability matrix $\mathbf{\Theta}_0$. This sparse W-random graph model was considered in [201, 196, 24, 23] among others. We can summarize the generating process of graph as follows

$$W_0 \to \mathbf{\Theta}_0 \to \mathbf{A} \ . \tag{3.6}$$

In [A16], we have been interested in the two related problems of inferring the probability matrix $\mathbf{\Theta}_0$ and the graphon W_0 from a single observation of the adjacency matrix $\mathbf{A}$. We first describe estimation in Frobenius/δ_2 norm, before moving to a weaker distance, called the cut distance.

3.2.1 Sparse Graphon estimation

Estimating the probability matrix $\mathbf{\Theta}_0$. First, we study optimal rates of estimation of the probability matrix $\mathbf{\Theta}_0$ under the Frobenius norm from a sample $\mathbf{A}$. We estimate $\mathbf{\Theta}_0$ by a block-constant matrix and we focus on deriving oracle inequalities with optimal rates. Estimating $\mathbf{\Theta}_0$ by a $K \times K$ block constant matrix is equivalent to fitting a stochastic block model with K classes. Estimation of $\mathbf{\Theta}_0$ has already been considered by [51, 201] but convergence rates obtained there are far from being optimal. Gao et al. [90] have established the minimax estimation rates for $\mathbf{\Theta}_0$ on classes of block constant matrices and on classes of smooth matrices. Their analysis is restricted to the dense case corresponding to $\rho_n = 1$ when dealing with model (3.5).

For a fixed integer K and a threshold ρ_n, let $\widehat{\mathbf{\Theta}}^{\rho_n}$ be the least-square estimator of $\mathbf{\Theta}_0$ among K-block constants matrices whose maximum entry is smaller than ρ_n. In the following result, we consider the matrix $\mathbf{\Theta}_0$ as fixed (or equivalently we work conditionally to $\zeta_1, \ldots, \zeta_n$)

Theorem 3.2 ([A16]). *If $\max_{ij} |(\mathbf{\Theta}_0)_{ij}| \leq \rho_n$, then the restricted least-square estimator $\widehat{\mathbf{\Theta}}^{\rho_n}$ satisfies*

$$\mathbb{E}\left[\frac{1}{n^2}\|\widehat{\mathbf{\Theta}}^{\rho_n} - \mathbf{\Theta}_0\|_F^2\right] \lesssim \frac{1}{n^2}\|\mathbf{\Theta}_0 - \mathbf{\Theta}_*\|_F^2 + \rho_n\left(\frac{\log(K)}{n} + \frac{K^2}{n^2}\right) \ , \tag{3.7}$$

where $\mathbf{\Theta}_*$ is the best approximation of $\mathbf{\Theta}_0$ by a K-block constants

Conversely, (3.7) is shown to be minimax optimal when $\mathbf{\Theta}_0$ is a K-block constant matrix. In particular, we recover the right dependency with respect to both the number K of blocks and the scale parameter ρ_n. As a corollary of (3.7), we recover minimax non-parametric rates for estimating $\mathbf{\Theta}_0$ if, up to a permutation of its rows and columns, $\mathbf{\Theta}_0$ is smooth. See also [89] for related and almost simultaneous results.

Graphon estimation in δ_2 distance. In the W-random graph model, the ultimate objective is to estimate the graphon function W_0 rather than the probability matrix $\mathbf{\Theta}_0$. Unfortunately, contrary to $\mathbf{\Theta}_0$, the graphon W_0 is not identifiable. We recall the following result [150, Sect.10]. Two graphons U and W in $\mathcal{W}$ are called weakly isomorphic if there exist measure preserving maps ϕ, ψ: $[0,1] \to [0,1]$ such that $U^\phi = W^\psi$ almost everywhere. Here $U^\phi(x,y) = U(\phi(x),\phi(y))$. Two graphons U and W define the same probability distribution if and only if they are weakly isomorphic. Hence, we could only hope to estimate W in the corresponding quotient space $\widetilde{\mathcal{W}}$. Then, the counterpart of the l_2 distance in the corresponding quotient space $\widetilde{\mathcal{W}}$ is defined [150, Ch.8,13] as

$$\delta_2^2(W,W') := \inf_{\tau\in\mathcal{M}} \int\int_{(0,1)^2} |W(\tau(x),\tau(y)) - W'(x,y)|^2 \mathrm{d}x\mathrm{d}y \;, \tag{3.8}$$

where $\mathcal{M}$ is the set of all measure-preserving bijections $\tau : [0,1] \to [0,1]$.

Given a symmetric $n \times n$ matrix $\mathbf{\Theta}$, we can easily transform it as a graphon function $W_{\mathbf{\Theta}}$ by $W_{\mathbf{\Theta}}(x,y) = \mathbf{\Theta}_{\lceil nx\rceil,\lceil ny\rceil}$ for any $(x,y) \in (0,1)$. From a practical perspective, if we have built a suitable estimator $\widehat{\mathbf{\Theta}}$ of $\mathbf{\Theta}_0$, it is natural to estimate W_0 by $W_{\widehat{\mathbf{\Theta}}/\rho_n}$, since we have no additional information of W_0 other than the fact that $\mathbf{\Theta}_0$ has been sampled according to (3.5).

As an important estimation class, consider the collection $\mathcal{W}[K]$ of k-step functions that is the subset of graphons $W \in \mathcal{W}$ such that, for some $\mathbf{Q} \in \mathbb{R}^{K\times K}_{\mathrm{sym}}$ and some $\phi : [0,1] \to [K]$, $W(x,y) = \mathbf{Q}_{\phi(x),\phi(y)}$ for all $x,y \in [0,1]$. In fact, $\mathcal{W}[k]$ stands for the collections of all graphons corresponding to a stochastic block model with at most K groups. The following proposition controls the risk of our estimator on the class of k-steps graphons.

Proposition 3.3 ([A16])**.** *Assume that the graphon $W_0 \in \mathcal{W}[K]$. The graphon $W_{\widehat{\mathbf{\Theta}}^{\rho_n}/\rho_n}$ estimated the restricted least squares estimator with K groups satisfies*

$$\mathbb{E}\left[\delta_2^2\left(W_{\widehat{\mathbf{\Theta}}^{\rho_n}/\rho_n}, W_0\right)\right] \lesssim \left[\frac{1}{\rho_n}\left(\frac{K^2}{n^2} + \frac{\log(K)}{n}\right) + \sqrt{\frac{K}{n}}\right]. \tag{3.9}$$

Comparing (3.9) with (3.7), we observe an additional error term of the order of $\sqrt{K/n}$, called the agnostic error. The error term would arise even if we had observed perfectly the matrix $\mathbf{\Theta}_0$. Intuitively, it comes from the fact that $\mathbf{\Theta}_0$ in (3.5) is a sample from W_0 where we do not observe the design $\xi_1,\dots,\xi_n$. Although this term seems unavoidable, the main technical challenge is to prove a matching minimax lower bound (3.9). Indeed, the definition of the δ_2 distance involves an infimum over all measure-preserving bijection τ, so that it is delicate to show that a collection of graphons are $\sqrt{K/n}$-distant in δ_2^2 distance. See Lemma 8 in [A16] for the minimax lower bound matching the risk bound (3.9).

As a closing remark, we point out that the least-square estimators considered in [A16] suffer from an exponential complexity. Polynomial-time procedures such as those based on singular value thresholding unfortunately achieve slower convergence rates –see e.g. [A9]. While a computational-statistical tradeoff is not formally proved, this is somewhat expected given the conjectured computational lower bounds for clustering stochastic-block models with a large number of groups [54].

3.2.2 Graphon estimation in cut distance

In the previous subsection, we were interested in estimating the probability matrix $\mathbf{\Theta}_0$ in Frobenius norm and the graphon function W_0 in the corresponding δ_2-distance. While these two loss functions are sensible for matrices and bivariate functions, they do not reflect particular structural similarities between the graphs. For this purpose, Frieze and Kannan [86] have introduced another notion of distance, called the *cut distance*. The cut norm of a matrix $\mathbf{B} = (\mathbf{B}_{ij}) \in \mathbb{R}^{n\times n}$ is defined by

$$\|\mathbf{B}\|_\square = \frac{1}{n^2} \max_{S,T\subset[n]} \Big| \sum_{i\in S, j\in T} \mathbf{B}_{ij} \Big|.$$

In other words, $\|\mathbf{B}\|_\square$ corresponds (up to to a renormalization) to the maximal sum of entries over all submatrices of $\mathbf{B}$. Then, the cut distance $d_\square(G, G')$ between two graphs G and G' defined on the same set of nodes and with adjacency matrices $\mathbf{A}$ and $\mathbf{A}'$ is defined as the cut norm $\|\mathbf{A} - \mathbf{A}'\|_\square$. Denoting $e_G(S, T)$ the number of edge between nodes in S and T in the graph G, the cut distance $d(G, G')$ is the supremum over all S, T of $|e_G(S, T) - e_{G'}(S, T)|/n^2$. In other words, $d_\square(G, G')$ is small if the restrictions of G and G' to all subsets S, T have similar edge densities. Similarly, we define the cut norm of a graphon $W \in \mathcal{W}$ by $\|W\|_\square = \sup_{S,T\subset[0,1]} |\int_{S\times T} W(x, y)\mathrm{d}x\mathrm{d}y|$, where the supremum is taken over all measurable subsets S and T. Since the graphons W are not identifiable, we consider the metric induced by $\|\cdot\|_\square$ on the quotient space $\widetilde{\mathcal{W}}^+$ defined by

$$\delta_\square(W_1, W_2) = \inf_{\tau\in\mathcal{M}} \|W_1 - W_2^\tau\|_\square \,, \tag{3.10}$$

where we take the infimum in the set $\mathcal{M}$ of all measure-preserving bijections $\tau : [0, 1] \to [0, 1]$ and $W^\tau(x, y) = W(\tau(x), \tau(y))$.

The cut distance is also a cornerstone in the graph limit theory introduced by Lovász and Szegedy [151] and further developed in, e.g. [29, 30]. In particular, this theory states that graphons can be interpreted as limits (with respect to $\delta_\square$) of graph sequences. Besides, convergence in $\delta_\square$ is equivalent to other structural properties such as the convergence of all homomorphisms numbers. Given a simple graph F with q nodes and a graphon W_0, the homomorphisms number $t(F, W_0)$ is the probability that a size q subgraph of W-random graph (3.5) contains the edge set F. As a consequence, the homomorphisms numbers $t(F, W_0)$ and $t(F, W_0')$ are close when the expected number of subgraphs F for a size n random graph G sampled from W_0 is close to that of a size n random graph sampled from W_0'. It has been established that convergence in the cut distance is equivalent to convergence of homomorphism numbers for all simple graphs F (see Theorem 11.5 in [150] for more details). Hence, estimating well the graphon W_0 in the cut distance allows to estimate well the number of small patterns induced by W_0.

The metric $\delta_\square$ is dominated by the previous metric δ_2 so that convergence in cut metric is weaker than convergence in δ_2. One of the striking consequences of the celebrated Szemerédy's regularity Lemma [185] states that an adjacency matrix sampled from a W-random graph model converges to the true graphon W_0 in cut distance, this at an *uniform* rate over all graphons. More precisely, fix a graphon W_0 in $\mathcal{W}$ and let $\mathbf{A}$ denote the adjacency matrix of a size n random graph sampled from W_0 (as in (3.6) but with $\rho_n = 1$). With high probability, the empirical graphon $W_\mathbf{A}$ built from the adjacency matrix $\mathbf{A}$ satisfies $\delta_\square[W_0, W_\mathbf{A}] \lesssim 1/\sqrt{\log(n)}$, this uniformly over all possible W_0.

Our contribution. In [A9], we investigate to what extent the $\delta_\square$ convergence rate can be improved for specific classes of graphon such a K-steps graphons and how to craft an optimal estimator. From a practical perspective, the problem turns out to be a trivial one, as we establish that

the empirical graphon $W_{\mathbf{A}}$ turns out to be minimax optimal simultaneously over all classes $\mathcal{W}[K]$. Recall that $\mathcal{W}[K]$ is the collection of graphons corresponding to stochastic block models with at most K groups. In practice, it could be disappointing that the raw data are already optimal with respect to the cut distance, whereas they perform really badly with respect to the δ_2 distance. This is why we also prove that a singular value thresholding estimator is still optimal with respect to the cut metric $\delta_\square$ while achieving the best known rate in δ_2-distance in the class of polynomial-time estimators. Our main result is a characterization of the minimax convergence rate over the class $\mathcal{W}[K]$.

Theorem 3.4 ([A9])**.** *For any $K \geq 2$, we have*

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[K]} \mathbb{E}_{W_0} \left[\delta_\square \left(\widehat{W}, W_0 \right) \right] \asymp \min \left(\sqrt{\frac{K}{n \log(K)}}, \frac{1}{\sqrt{\log(n)}} \right) . \tag{3.11}$$

The rate (3.11) is achieved by the empirical graphon $W_{\mathbf{A}}$. In some way, the purpose of [A9] is twofold: first, we provide a matching minimax lower bound for the universal $1/\sqrt{\log(n)}$ rate. Second, we prove how graphons with a simpler structure, that is belonging to $\mathcal{W}[k]$, are to be estimated at as faster rate than the universal convergence rate.

From a technical perspective, the tools needed for deriving optimal cut distance rates differ from those used for the δ_2-distance. The proof relies among other things on a careful application of Szemerédi's regularity lemma to distorted versions of the graphon, together with Khintchine's inequality.

Discussion and open problems. Graphon estimation is arguably an interesting and challenging problem. Unfortunately, there are some important limitations from a practical perspective, the main one being that a suitable estimator $\widehat{W}$ of a graphon W_0 is possibly hard to interpret. To illustrate this issue, we move to a slightly different definition of graphons. A graphon is now defined by a triplet (Ω, μ, W), where (Ω, μ) is a probability space and $W : \Omega \times \Omega \to [0,1]$ is a measurable bivariate function. In the previous definition (Ω, μ) was restricted to be the unit segment $[0,1]$ endowed with the Lebesgue measure λ. It turns out than any such graphon (Ω, μ, W) is weakly-isomorphic to some $([0,1], \lambda, W')$ for some W'–see [150]. Since graphons are only identifiable up to a weak isomorphism, it was therefore not restrictive to focus our attention to representatives with the latent space $[0,1]$ as we did in this chapter (and as done in most of the literature in graphon estimation). However, this general latent space perspective on graphons raises an important problem. Even if the graphon W is a simple function on a latent space Ω, it is possible that all the equivalent graphons W' on $[0,1]$ are really erratic. As a consequence, a good estimation of the graphon (e.g. with respect to the δ_2 metric) on $[0,1]$ is possibly not insightful to understand the random graph model. As a simple example, consider a random geometric graph on $[0,1]^d$ with $d \geq 2$. While there exists a simple graphon representation on the latent space $[0,1]^d$, any representation on $[0,1]$ is irregular. This suggests that a most important question in graphon estimation is to find some informative latent space (Ω, μ) to represent the graph. Second, this raises the question of defining suitable non-parametric class for graphons. In classical non-parametric estimation on $[0,1]$, smoothness classes (e.g. Hölder's class) are particularly suited. For this reason, optimal estimation rates have been studied for such classes in [90] and in our own work [A9]. However, this class does not capture simple random graph models with latent space dimension higher than 1. In summary, the most important challenges to make the graphon estimation framework applicable seem the lack of a suitable approximation theory and the problem of finding a suitable representation.

Chapter 4

Clustering

4.1 Introduction

The problem of clustering is that of grouping similar "objects". Depending on the context, these objects can be points in a metric space, nodes of a graph,... Clustering serves many purposes in data analysis, including data visualization, data compression, dimension reduction, ...

This diversity of motivations has spurred long and vivid research streams at the crossroads of statistics and theoretical computer science. In this chapter, I only focus on a single perspective, which is referred as the "hidden partition"[1] problem. Informally, this perspective postulates that there exists an unknown true underlying partition $G^* = (G_1^*, \dots, G_K^*)$ of these n objects in K groups. The data-set $\mathbf{X}$ is assumed to have been sampled according to a distribution $\mathbb{P}_{G^*}$. Then, the goal is to recover this hidden partition G^* from $\mathbf{X}$. Within this formalism, we can summarize the data generating process and the statistical objectives as follows

$$G^* \underset{\mathbb{P}_{G^*}}{\to} \mathbf{X} \to \widehat{G} \ .$$

Given a given partition $G^* = (G_1^*, \dots, G_K^*)$ and $a \in [n]$, $k^*(a)$ stands for index of the group of the a-th object. This viewpoint includes some of the most popular clustering probabilistic models such as Gaussian mixture models (GMM) or Stochastic block models (SBM) defined below.

Definition 4.1 ((conditional) Gaussian Mixture Model (GMM))**.** *Let $\mu_1, \dots, \mu_K \in \mathbb{R}^p$ and $\mathbf{\Sigma}_1, \dots, \mathbf{\Sigma}_K \in \mathbb{R}^{p \times p}$ be K vectors and K covariance matrices. Fix a partition G^* with K groups. Then, the data matrix $\mathbf{X} = (X_1, \dots, X_n)^T \in \mathbb{R}^{n \times p}$ has independent rows that satisfy $X_a \sim \mathcal{N}(\mu_{k^*(a)}, \mathbf{\Sigma}_{k^*(a)})$ for all $a = 1, \dots, n$.*

The above model differs from the usual definition of Gaussian mixtures because the partition G^* is considered as fixed, whereas it is usually assumed that the partition G^* has been sampled according to some expected proportions $\pi^* = (\pi_1^*, \dots, \pi_K^*)$. One may interpret Definition 4.1 as a specific instance of the classical Gaussian mixture model where we work conditionally on on the hidden labels.

The problem of recovering the partition G^* from $\mathbf{X}$ is identifiable if all the $(\mu_k, \mathbf{\Sigma}_k)$'s are distinct for $k = 1, \dots, K$. In this chapter, we focus our attention on settings where the μ_k's are all distinct and we want to build upon the differences in the means to recover the partition.

Definition 4.2 ((conditional) Stochastic Block Models (SBM))**.** *Let $\mathbf{Q} \in [0,1]^{K \times K}_{\text{sym}}$ denote a symmetric matrix of probabilities. Fix a partition G^* with K groups. The data matrix $\mathbf{X} \in \{0,1\}^{n \times n}$*

[1] also called planted partition.

corresponds to the adjacency matrix of an undirected simple graph, that is $\mathbf{X}$ *is symmetric and its diagonal is zero. The* $\mathbf{X}_{a,b}$*'s are independent and* $\mathbf{X}_{a,b} \sim \mathcal{B}(\mathbf{Q}_{k^*(a),k^*(b)})$ *for* $1 \leq a < b \leq n$.

As previously, this slightly differs from the usual definition of SBM [110], as the partition G^* is considered as fixed and not sampled according to some proportions π^*. When the matrix $\mathbf{Q} = (\alpha-\beta)\mathbf{I}_K+\beta\mathbf{J}_K$ where $\mathbf{J}_K$ is the constant matrix with 1's and $0 < \beta < \alpha < 1$, this corresponds to the so-called affiliation model where nodes within the same group are connected with a higher probability α, whereas the connection probability β between nodes in distinct groups is lower. In the sequel, we say more generally that the model is assortative if the diagonal terms of $\mathbf{Q}$ are larger than non-diagonal terms, so that the nodes in the same group tend to be more connected than nodes in a different group. If we pick $\beta > \alpha$ in the previous matrix $\mathbf{Q}$, then the corresponding graph will be nearly K-partite, in the sense that there are more edges between groups than within groups. In fact, it is possible to sample various forms of random graphs by playing with with the matrix $\mathbf{Q}$. For general $\mathbf{Q}$, two nodes in the same group share the property that at least, in expectation, they are similarly connected to all the other nodes. In turns out that the problem of recovering G^* from $\mathbf{X}$ is identifiable if and only if the all the rows of $\mathbf{Q}$ are distinct.

In both these models, the general objective is to estimate from the raw data $\mathbf{X}$ a partition $\widehat{G}$ which is as close as possible to the true partition G^*. If possible, the corresponding procedure should run in polynomial time with respect to the size (K, n, p) of the problem. For both the Gaussian mixture models and the stochastic block models, this problem has attracted a lot of interest. Many different procedures have been studied including spectral clustering [149, 142], semi-definite programs [158, 5], Loyd's algorithms [152] or more generally iterative [93] algorithms. For SBMs, more specific procedures tailored for sparse graphs such as [28, 3] have also been proposed –see the survey [2].

Organization of the chapter. I will first describe in depth a joint work [A13] with C. Giraud, where we study the behavior of a convex relaxation of the K-means algorithm both for general GMMs and SBMs. Interestingly, this fairly general method achieves nearly optimal clustering errors in almost all regimes. Then, Section 4.3 is more specifically dedicated to the case where the number of K of groups is large. This section differs from the rest of the manuscript as I mostly describe open problems and conjectures, although a joint work [A11] with J. Banks, C. Moore, J. Xu, and R. Vershynin is mentioned. Section 4.4 is dedicated to the slightly different problem of variable clustering that we addressed with F. Bunea, C. Giraud, X. Luo, and M. Royer [A7]. Finally, I discuss an older joint work with E. Arias-Castro on sparse clustering [A17].

In this chapter, my hope is to provide a clear picture of the state-of-the-art in GMM clustering. I find this field especially interesting because (i) it shares deep connections with other statistical problems (e.g. density estimation in GMM), (ii) it provides some intuitions on other related clustering problems (e.g. block Ising models), and (iii) there exist many open problems when additional structure is added (e.g. sparsity).

4.2 Analysis of relaxed K-means for GMM and SBM

This section is mainly based on [A13].

4.2.1 K-means and relaxed K-means

When the objects we want to cluster correspond to vectors in a Euclidean space, one of the most standard clustering approach is based on the minimization of the K-means criterion [147]. Writing $X_a \in \mathbb{R}^p$ for the object $a \in [n]$, the K-means criterion of a partition $G = (G_1, \dots, G_k)$ of $[n]$ is defined as

$$\mathrm{Crit}(G) = \sum_{k=1}^{K} \sum_{a \in G_k} \left\| X_a - \frac{1}{|G_k|} \sum_{b \in G_k} X_b \right\|_2^2, \tag{4.1}$$

where $\|.\|_2$ is the Euclidean norm. The criterion (4.1) quantifies the dispersion of each group. Hence, a smaller value of the criterion indicates that the partition is more homogeneous. A K-means procedure then aims at finding a partition $\widehat{G}$ that minimizes, at least locally, the K-means criterion (4.1). Unfortunately, this minimization problem is NP-hard [10].

In practice, one often resorts to iterative minimization procedures such as Lloyd's algorithm [147] and its variants [8], but those are only proved to converge to a local minimum of (4.1), unless the initialization is close enough to the global one. As an alternative, Peng and Wei [168] have suggested to relax the K-means criterion to a Semi-Definite Program (SDP) followed by a rounding step. The resulting program is provably solvable in polynomial time.

Let us describe this convex criterion. We denote $\mathbf{X} \in \mathbb{R}^{n \times p}$, the $n \times p$ matrix whose a-th row is given by X_a (as in Definition 4.1). Any partition G of $[n]$ can be encoded by a so-called partnership matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ such that $\mathbf{B}_{ab} = 0$ if and only if a and b belong to distinct groups of G and $\mathbf{B}_{ab} = 1/|G_k|$ if a and b are in the same group G_k. For a fixed number K of groups, the collection of all partnership matrices when G spans all possible partitions may be described as

$$\mathcal{P} = \left\{ \mathbf{B} \in \mathbb{R}^{n \times n} : \text{symmetric},\ \mathbf{B}^2 = \mathbf{B},\ \mathrm{tr}(\mathbf{B}) = K,\ \mathbf{B}1 = 1,\ \mathbf{B} \geq 0 \right\} .$$

Here, $\mathbf{B} \geq 0$ means that all entries of $\mathbf{B}$ are nonnegative and 1 is the constant vector whose coordinates are all equal to one. Peng and Wei [168] have established that minimizing the criterion (4.1) turns out to be equivalent to maximizing $\langle \mathbf{X}\mathbf{X}^T, \mathbf{B} \rangle$ over the space $\mathcal{P}$ of partnership matrices.

The constrain set $\mathcal{P}$ is non-convex which is expected as K-means is NP-hard [10]. Still, we are in good position to convexify the criterion now that K-means is expressed as a linear maximization problem. Indeed, Peng and Wei [168] suggest to drop the condition $\mathbf{B}^2 = \mathbf{B}$ in the set $\mathcal{P}$

$$\mathcal{C} = \left\{ \mathbf{B} \in \mathbb{R}^{n \times n} :\ \text{Positive Semi Definite},\ \mathrm{tr}(\mathbf{B}) = K,\ \mathbf{B}1 = 1,\ \mathbf{B} \geq 0 \right\}.$$

Hence, the corresponding semi-definite program (SDP) is defined as

$$\widehat{\mathbf{B}} \in \arg\max_{B \in \mathcal{C}} \langle \mathbf{X}\mathbf{X}^T, B \rangle. \tag{4.2}$$

$\widehat{\mathbf{B}}$ does not necessarily correspond to a partition. An additional step is therefore needed to round $\widehat{\mathbf{B}}$ into a proper partnership matrix. If $\widehat{\mathbf{B}}$ is close enough to the matrix $\mathbf{B}^*$ corresponding to the true partition G^*, then the rows of $\widehat{\mathbf{B}}$ corresponding to the same groups in G^* should be similar. This is why we choose the final rounding step to be done by applying a clustering algorithm to the rows of $\widehat{\mathbf{B}}$. For technical reasons, we resort here to an approximate K-medoids[2] on the rows of $\widehat{\mathbf{B}}$ (as in [82]) which can be performed efficiently [49]. This two-step procedure is referred henceforth as relaxed K-means.

[2] K-medoid is the counterpart of K-means where the square Euclidean norm in (4.1) is replaced by the Euclidean norm

One may interpret the exact K-means minimization problem as the maximum likelihood estimator of G^* in a Gaussian mixture model (Definition 4.1), where all the covariance matrices are identical and proportional to the identity matrix. Still, the K-means criterion and its relaxation do not pertain to a particular hidden partition model and are applied in various contexts. The purpose of our work [A13] is to show that the relaxed K-means is highly versatile and lead to near optimal clustering performances for two different emblematic models: Gaussian mixture models and stochastic block models.

4.2.2 Clustering Gaussian mixtures

In this subsection, we consider general Gaussian mixture models as defined in Definition 4.1. All the results can be safely extended to subGaussian mixtures – see [A13].

Our goal is to quantify the ability of a clustering procedure to recover the hidden partition G^*. One may aim at establishing that $\hat{G} = G^*$ with high probability, that is at *exactly* recovering G^*. Unfortunately, this objective is sometimes too demanding when the groups do not differ that much. For this reason, we introduce the following loss function

$$err(\widehat{G}, G^*) = \min_{\pi \in \mathcal{S}_K} \frac{1}{2n} \sum_{k=1}^{K} \left| G_k^* \triangle \widehat{G}_{\pi(k)} \right| , \tag{4.3}$$

where Δ stands for the symmetric difference and $\mathcal{S}_K$ is the collection of permutation of $[K]$. Hence, $err(\widehat{G}, G^*)$ quantifies the proportion of misclassified objects in $\hat{G}$. To simplify the exposition, we assume throughout that the true partition is approximately balanced, that is $\min_k |G_k^*| \asymp \max_k |G_k^*| \asymp n/K$ – see [A13] for a general treatment. In such a case, if we pick $\widehat{G}$ uniformly at random, then its loss is of the order of $(K-1)/K$. We say that a procedure achieves *approximate recovery* when the loss $err(\widehat{G}, G^*)$ is strictly smaller than this quantity in the sense that $err(\widehat{G}, G^*) \leq \alpha(K-1)/K$ for some fixed $\alpha \in [0, 1)$. We are interested in finding the *minimal* separation condition between the groups to be able to achieve *approximate recovery* and when approximate recovery is possible to achieve *the smallest possible error*.

For this purpose, we need to introduce some sort of signal-to-noise ratio s^2 of the clustering problem. Intuitively, the larger the Euclidean distance between two centers $\Delta_{jk} = \|\mu_k - \mu_j\|_2$, the more easily we can recover the partition. The difficulty of the clustering problem also depends on the covariance matrices $\mathbf{\Sigma}_k$'s. We introduce two quantities to quantify the noise level: the maximum operator and Frobenius norms $\sigma^2 = \max_k \|\mathbf{\Sigma}_k\|_{op}$ and $\nu^2 = \max_k \|\mathbf{\Sigma}_k\|_F$.

Actually, as shown in Theorem 4.3 below, the misclassification error of the relaxed K-means decreases exponentially fast with the signal-to-noise ratio

$$s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{n\Delta^4}{K\nu^4}. \tag{4.4}$$

To ease the interpretation, one may consider the spherical case $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_K = \sigma^2 \mathbf{I}_p$, in which case s^2 simplifies as $s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{n\Delta^4}{Kp\sigma^4}$. We explain below the intuition behind (4.4).

4.2.2.1 Equal trace case

In the following theorem, we assume that $tr[\mathbf{\Sigma}_1] = tr[\mathbf{\Sigma}_2] = \ldots = tr[\mathbf{\Sigma}_K]$.

Theorem 4.3 (Equal Trace case [A13])**.** *If the signal-to-noise ratio s^2 satisfies $s^2 \gtrsim K$, then, with high-probability, the proportion of misclassified points by the relaxed K-means estimator (4.2) satisfies $err(\widehat{G}, G^*) \leq e^{-c's^2}$.*

A few comments are in order. Theorem 4.3 ensures partial recovery as soon as $s^2 \gtrsim K$. If we introduce the effective ranks R_Σ as the ratio $R_\Sigma = \frac{\nu^4}{\sigma^4} = \frac{\max_{k=1,\dots,K} |\Sigma_k|_F^2}{\max_{k=1,\dots,K} |\Sigma_k|_{op}^2}$, then Theorem 4.3 guaranties approximate recovery as soon as $s^2 \gtrsim K$, or equivalently

$$\frac{\Delta^2}{\sigma^2} \gtrsim \left(1 \vee \sqrt{\frac{R_\Sigma}{n}}\right) K \ , \tag{4.5}$$

and then the misclassification error is upper bounded by $e^{-c's^2} \leq e^{-cK}$ with high probability. In [A13], we advocate why, at least in the spherical case $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_K = \sigma^2 \mathbf{I}_p$, the exponential rate $e^{-c's^2}$ is essentially optimal. Recall that $s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{n\Delta^4}{Kp\sigma^4}$. Since the clustering problem only makes sense if $\Delta \gtrsim \sigma$ (otherwise even the Bayes classifier is not able to achieve approximate recovery), then we have $s^2 = \frac{\Delta^2}{\sigma^2}$ in the low-dimensional setting $n \geq Kp$ and $s^2 = \frac{n\Delta^4}{Kp\sigma^4}$ in the high-dimensional setting $n \leq Kp$. In the former case, the rate $e^{-c'\Delta^2/\sigma^2}$ turns out to correspond to the miss-classification error of the oracle procedure that knows the mean parameters before hand. In the high-dimensional case, the rate $e^{-c'n\Delta^4/(Kp\sigma^4)}$ corresponds to the optimal classification error in the arguably simpler case of supervised classification, where the statistician is also given the hidden labels of the n observations and has to to classify a new observation. As a consequence, the misclassification error e^{-cs^2} cannot be improved. This has been later formalized in a proper minimax lower bound by [161] in the specific case $K = 2$. When the number of groups K is considered as as constant, the signal condition $s^2 \gtrsim K$ cannot be improved as $s^2 = o(1)$ would lead to a trivial error.

We temporarily leave aside the discussion of the possible sub-optimality of the condition $s^2 \gtrsim K$ when the number of components K is large. This aspect of the problem remains partially ill-understood and certainly involves computational-statistical trade-offs [A11]. This will be the topic of Section 4.3.

Comparison with the literature. Lu and Zhou [152] provide exponential missclassification error for the Lloyd algorithm under the requirement $\frac{\Delta^2}{\sigma^2} \gtrsim K^2 \left(1 \vee \frac{pK}{n}\right)$ which is stronger than (4.5) in a high-dimensional setting. In the low-dimensional setting, Lu and Zhou are able to recover the optimal asymptotic constant inside the exponential whereas the constant c' in Theorem 4.3 is suboptimal. See also [161] for more recent results on Lloyd's algorithm that allow to also handle the high-dimensional setting. To the best of our knowledge, our result was the first of this kind for an SDP. In an independent and simultaneous work, Fei and Chen [83] have derived a similar in spirit result in the very precise setting where the groups are of equal size, but their signal requirement is again stronger than (4.5), especially in a high-dimensional setting.

Now assume that the common covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_1 = \ldots \mathbf{\Sigma}_k$ is not spherical. In that case, relaxed K-means still achieves the exponential error $e^{-c's^2}$, but it seems possible to achieve faster rates, at least for large sample size. Indeed, when the parameters μ_k's and $\mathbf{\Sigma}$ are known, the error of the Bayes classifier will depend on the Mahalanobis distance $d_\Sigma^2(\mu_k, \mu_l) = (\mu_k - \mu_l)^T \mathbf{\Sigma}^{-1} (\mu_k - \mu_l)$ rather than $\Delta^2/\sigma^2 = \min_{k \neq l} \|\mu_k - \mu_l\|^2 / \|\mathbf{\Sigma}\|_{op}$. Hence, the optimal rate of decay should involve d_Σ instead of Δ^2/σ^2, at least when the sample size is large. This is an active and stimulating direction of research [53, 64]. Theorem 4.3 entails that relaxed K-means is able to cope with non-spherical covariances, but K-means is certainly not able to build upon the geometry of covariance as in the Mahalanobis distance.

4.2.2.2 Unequal trace case

In the previous subsection, we assumed that all the mixtures had the same dispersion, in the sense that $\mathrm{tr}[\mathbf{\Sigma}_1] = \ldots = \mathrm{tr}[\mathbf{\Sigma}_k]$. If we do not assume this anymore, the K-means criterion (4.1) is biased. This is a well known phenomenon[3]: K-means may tend to cut wide groups into severals subgroup and merge smaller groups together. Recall that the exact K-means (which is NP-hard) problem can be expressed as the linear matrix maximization problem $\widehat{\mathbf{B}}_K \in \arg\max_{\mathbf{B}\in\mathcal{P}} \langle \mathbf{X}\mathbf{X}^T, \mathbf{B}\rangle$. In order to have some intuition on its behavior, one may consider the population counterpart of K-means where one replaces $\mathbf{X}\mathbf{X}^T$ by its expectation, that is $\mathbf{B}_K \in \arg\max_{\mathbf{B}\in\mathcal{P}} \langle \mathbb{E}[\mathbf{X}\mathbf{X}^T], \mathbf{B}\rangle$. Unfortunately, this population K-means solution $\mathbf{B}_K$ may differ from the oracle solution $\mathbf{B}^*$ which corresponds to the hidden partition G^*. Indeed, it is not much difficult to show that the $\mathbf{B}^* \in \arg\max_{\mathbf{B}\in\mathcal{P}} \langle \mathbb{E}[\mathbf{X}]\,\mathbb{E}[\mathbf{X}^T], \mathbf{B}\rangle$. Looking closely at the population and the oracle programs, one sees that those two differ because $\mathbb{E}[\mathbf{X}\mathbf{X}^T] = \mathbb{E}[\mathbf{X}]\,\mathbb{E}[\mathbf{X}^T] + \mathbf{\Gamma}$, where $\mathbf{\Gamma}$ is diagonal matrix such that $\mathbf{\Gamma}_{aa} = \mathrm{tr}[\mathbf{\Sigma}_{k^*(a)}]$. If all the traces are equal, then $\mathbf{\Gamma}$ is proportional to the identity and it is therefore benign in the maximization of $\langle \mathbb{E}[\mathbf{X}]\,\mathbb{E}[\mathbf{X}^T] + \mathbf{\Gamma}, \mathbf{B}\rangle$. Unfortunately, when $\mathbf{\Gamma}$ has very high variations (different traces), this can highly bias the behavior of the K-means. To handle this issue, we propose in [A7, A13] to correct the relaxed K-means criterion by considering $\widehat{\mathbf{B}}_c \in \arg\max_{\mathbf{B}\in\mathcal{C}} \langle \mathbf{X}\mathbf{X}^T - \widehat{\mathbf{\Gamma}}, \mathbf{B}\rangle$, where $\widehat{\mathbf{\Gamma}}$ is a suitable estimator of $\mathbf{\Gamma}$. If we knew the true partition G^* in advance, we could easily estimate the $\mathrm{tr}(\mathbf{\Sigma}_k)$'s by plug-in. Unfortunately, this is not possible as G^* is precisely our goal. Still, a rough estimator $\mathbf{\Gamma}$ is sufficient to counter the bias, so that simple polynomial-time estimators of $\mathbf{\Gamma}$ are sufficient for our purpose –see [A7, A13] for the details.

Theorem 4.4 (Unequal trace case)**.** *Assume that, for all $k = 1, \ldots, K$, $\|\mathbf{\Sigma}_k\|_{op}\,\mathrm{tr}[\mathbf{\Sigma}_k] \lesssim \frac{n}{\log(n)}\|\mathbf{\Sigma}_k\|_F^2$. If the signal-to-noise ratio s^2 satisfies $s^2 \gtrsim K$, then, with high-probability, the proportion of misclassified points by corrected relaxed K-means satisfies $err(\widehat{G}, G^*) \leq e^{-c's^2}$.*

This condition on the covariances is mild: it allows covariance matrices whose singular values decay fast towards zero or conversely covariance matrices whose condition number is bounded by $n/\log(n)$. Provided that this condition is satisfied, then the corrected relaxed K-means achieves the same error bounds as in the equal trace case.

4.2.3 Recovery bounds for Stochastic Block models

We temporarily leave aside Gaussian mixture models and now turn to the Stochastic Block Models (SBM) described in Definition 4.2. We apply the relaxed K-means procedure to the adjacency matrix $\mathbf{X}$ to recover the unknown partition.

Let us first provide some intuition on why the K-means criterion seems suited to recovering the partition G^* in a SBM. For two nodes a and b in the same group G_k, the expectation $\mathbb{E}[\mathbf{X}_{a:}]$ of the a-th row matches the one of the b-th row to the exception of the a and b coordinates. As a consequence, up to the symmetries and up to mild changes, the distribution of the n rows of $\mathbf{X}$ is qualitatively not that different from a mixture of subGaussian distributions, for which K-means is expected to work well. From a graph perspective, (relaxed) K-means applied to the adjacency matrix $\mathbf{X}$ will tend to group together nodes that share similar pattern of connectivity.

As in the previous subsection, we still assume for the sake of presentation that the true partition is approximately balanced, that is $\min_k |G_k^*| \asymp n/K$. Again, to ease the exposition we assume in Theorem 4.5 below that the maximum connection probability $L = \max_{k,l} \mathbf{Q}_{k,l}$ satisfies $L \geq \log(n)/n$. The interesting case of sparser graphs $L \in (1/n, \log(n)/n)$ is also handled in [A13], but

[3] see e.g. https://en.wikipedia.org/wiki/K-means_clustering

the definition of the relaxed K-means criterion needs to be slightly modified to prevent the criterion to select too unequal groupe sizes.

Similarly to the sub-Gaussian setting, we define the minimum distance between two groups as

$$\Delta^2 := \min_{j\neq k} \Delta_{jk}^2 \; ; \quad \Delta_{jk}^2 := \sum_{\ell} \frac{n}{K}(\mathbf{Q}_{k\ell} - \mathbf{Q}_{j\ell})^2 = \frac{n}{K}\|\mathbf{Q}_{k:} - \mathbf{Q}_{j:}\|_2^2 \; , \tag{4.6}$$

which represents the signal strength in our analysis. Note that Δ^2 corresponds to the minimum square Euclidean distance between the expected rows $\mathbb{E}[\mathbf{X}_a]$ and $\mathbb{E}[\mathbf{X}_b]$ where a and b belong to distinct groups of the SBM. Since the sparsity parameter L plays the role of a proxy for the variance (in analogy to σ^2 for Gaussian mixtures), we consider the SNR $s^2 = \Delta^2/L$.

The next theorem provides a recovery bound for relaxed K-means in stochastic block models. In contrast to the Gaussian mixture setting, no correction is needed to debias the criterion.

Theorem 4.5. *Assume that $L \gtrsim \log(n)/n$. Provided that $s^2 \gtrsim K$, then, with high probability, the proportion of misclassified nodes satisfies $err(\widehat{G}, G^*) \leq e^{-c's^2}$.*

The statement Theorem 4.5 is quite similar to 4.3. Both theorems states that at least the SNR s^2 is higher than K, then the missclassification error is smaller than $e^{-c's^2}$.

An SDP not tied to the assortative case. We recall that an assortative SBM is a model where the connection probabilities (Q_{kk}) inside a group is higher than the connection probabilities (Q_{kl}) between distinct groups. All previous SDP methods used for SBM clustering take their origin, in some way or others, in Goemans-Williamson SDP relaxation of max-cut problem [101], which is tailored to the assortative case. In contrast, (4.2) is derived as a convex relaxation of K-means and can handle a wide range of settings going beyond the assortative case usually handled by SDP algorithms.

Assortative case. Still, to start the discussion, we explicit our rate and our SNR in the toy assortative model, where $\mathbf{Q} = q\mathbf{J}_K + (p-q)\mathbf{I}_K$ with $p < q$. In this case, $s^2 = 2n(p-q)^2/(pK)$, and we obtain the same rate of exponential decay as in [82, 92, 3, 57, 204], but without the tight constants of [92, 204] in the exponential rate. Yet, we stress that we are also able to deal with groups with unknown size. Besides, Theorem 4.5 ensures perfect recovery for

$$\frac{(p-q)^2}{p} \gtrsim \frac{K(K \vee \log(n))}{n} \; . \tag{4.7}$$

matching the best known results (up to constants) for polynomial-time algorithms [54]. When $K \leq \log(n)$, this condition matches the information theoretical condition $\frac{(p-q)^2}{p} \gtrsim \frac{K\log(n))}{n}$ [54] but, for larger larger K, there is a multiplicative gap of the order of $\log(n)/K$ which is conjectured to be unavoidable for polynomial-time procedures [33, 3].

Partial recovery for general models. To the best of our knowledge, outside the assortative case, the only other exponentially decaying misclassification error is stated in [3] for a quite different procedure. Abbe and Sandon only focus on the sparse regime $\mathbf{Q} = \mathbf{Q}_0/n$ where $\mathbf{Q}_0$ is a fixed matrix and $n \to \infty$. The results are not completely comparable, because we obtain faster convergence rates (at least by a factor K inside the exponential) but those hold under stronger signal condition than those of [3]. We again emphasize that Theorem 4.5 is also valid in denser regimes than that of [3].

To conclude this section, we recall the bottom line of our work [A13]. The generic relaxed K-means procedure is able to achieve near state-of-the art recovery bounds, this for all forms of matrices $\mathbf{Q}$, beyond the classical assortative case and even in the challenging sparse case where the connection probability are smaller than $\log(n)/n$.

4.3 Large K asymptotic

We now come back to the Gaussian mixture model (Definition 4.1) and discuss the difficulty of recovering the hidden partition G^* when the number of K groups is large. As this asymptotic with respect to K is still poorly understood, we focus on the toy model of iso-volumetric spherical covariance matrices $\mathbf{\Sigma} = \mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_k = \sigma^2 \mathbf{I}_p$ with a nearly balanced partition, that is $\min_k |G_k^*| \asymp n/K$. In this case, our signal-to-noise ratio simplifies as $s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{n\Delta^4}{pK\sigma^4}$.

In Theorem 4.3, we stated that, as soon $s^2 \gtrsim K$, the relaxed K-means estimator $\widehat{G}$ achieves a misclassification error of the order of $e^{-c's^2}$. We argued that the error $e^{-c's^2}$ cannot be improved as it matches the optimal error for arguably simpler problems such as supervised classification. However, the signal condition $s^2 \gtrsim K$ needs to be discussed. Two questions are in order: (i) What is the information-theoretical minimum signal condition for approximate recovery and what is the corresponding optimal classification error? (ii) Is there a gap between the information-theoretical minimum signal and performances of polynomial-time procedures? Although these twin questions have been much investigated for SBM problems, this is less the case for Gaussian mixtures. Besides, the situation turns out to be much more complex. Contrary to the remainder of the manuscript, the purpose of this section is mainly to discuss open problems and to introduce a few conjectures.

4.3.1 Information-theoretical threshold

Before addressing the questions (i) and (ii), let us make a detour with the related problem of parameter estimation in Gaussian mixtures models. Contrary to the previous section, we consider here the usual (non-conditional) definition of the Gaussian mixture models where one observes a sample from a random vector with density $f = \sum_{k=1}^K \pi_k \phi_{\mu_k}$. Here, ϕ_{μ_k} stands for the density of the Gaussian distribution with mean μ_k and covariance $\sigma^2 \mathbf{I}_p$. Given an n-sample of f, the goal is to estimate the parameters (π_k, μ_k) of the model. Contrary to the clustering problem, parameter estimation can be dealt with without any separation condition between the μ_k's. The minimax rate for estimating these parameters has recently been tightly characterized by Doss et al. [73] and turns out to be really slow. In fact, even in dimension 1 ($p = 1$), the sample size n has to be exponentially larger with respect to K to estimate well the parameters. This is not unexpected as, for very close means μ_k, it is really challenging to disentangle the different components of the mixtures. This has spurred a recent line of research in theoretical computer science, where the aim is estimate[4] the parameters with a polynomial sample size ($n = \text{poly}(p, K)$)[5] under suitable conditions on the parameters μ_k's. In particular, the objective is to characterize the minimal separation between the μ_k's which is needed to be able to estimate the parameters with polynomial sample size. Note that, if we are able to approximately estimate the parameters, then a simple plug-in classifier will be able to achieve approximate recovery provided that the separation distance Δ is large enough. In fact, Regev and Vijayaraghavan [173] have shown the following phase transition phenomenon: if $s^2 \lesssim \log(K)$, then an exponentially large (in K) sample size is necessary for accurate

[4] In this literature, approximate estimation should be understood as estimating the parameters up to an error which is small compared to the separation distance.

[5] $\text{poly}(a, b)$ is any finite-degree polynomial with respect to a and b.

mean estimation. Conversely, if $s^2 \gtrsim \log(K)$ is large and n is large ($n \geq \text{poly}(K,p)$), then an exponential-time procedure is able to estimate approximately the parameters. More recently, Kwon and Caramanis [138] have improved this by showing that $s^2 \gtrsim \log(K)$ is sufficient for approximate parameter estimation as long as $n \gtrsim K^3 p$ –see also [177] for related results. Although none of these results are able to handle the high-dimensional case, this and other informal arguments lead us to make the following conjecture for Question (i) above.

Conjecture 4.1 (Information-Theoretical threshold for Clustering)**.**

(a) If $s^2 \gtrsim \log(K)$, then, exact *K-means satisfies $err(\widehat{G}, G^*) \leq e^{-cs^2}$ with high probability.*

(b) Conversely, if $s^2 \lesssim \log(K)$, then no clustering procedure achieves partial recovery.

If true, this conjecture would extend the result of [138, 173] in two ways: first it would entail that the $\log(K)$ threshold is valid in arbitrary dimension p (which can be even larger than n). Second, it would prove that the optimal error rate remains driven by $e^{-c's^2}$ above the threshold.

4.3.2 Polynomial-time threshold

Taking for granted that Conjecture 4.1 is true, one may then wonder whether the gap between the condition $s^2 \gtrsim K$ for relaxed K-means and the information-theoretical threshold $s^2 \gtrsim \log(K)$ is intrinsic or not. Alternatively, is there a computational-statistical tradeoff that prevents any polynomial-time clustering procedure to achieve approximate recovery below the threshold $s^2 \gtrsim K$? With this level of generality, this conjecture turns out to be false. Consider for instance the case where $n \gg K = p \gg \log(n)$. In that situation, it is not hard to show that a simple distance clustering method that groups together points at a distance smaller than $\sigma^2[p + c\sqrt{K\log(n)} + c\log(n)]$ for a large constant c achieves perfect recovery provided that[6] $s^2 = \Delta^2/\sigma^2 \gtrsim \sqrt{K\log(n)}$ which is much smaller than K. Building upon this intuition, Vempala and Wang [189] have shown that in a large sample size asymptotic $n \gtrsim p^3K^2\log(pK)$, one may project the data onto a low-dimensional subspace and apply a simple distance clustering method to achieve perfect recovery under the condition $\Delta^2/\sigma^2 \gtrsim \sqrt{K\log(n) + \log(n)}$. In fact, it turns out that even the $\sqrt{K}$ threshold can be beaten by polynomial-time procedures provided that the sample size is large enough. Indeed, in the related problem of parameter estimation, some recent works [67, 111, 136] have shown that, for any fixed $\epsilon > 0$, a separation condition $\Delta/\sigma \geq K^\epsilon$ is sufficient for approximately estimating the means in polynomial-time, provided that the sample size is larger than $n \geq \text{poly}(p, K^{1/\epsilon})$. This implies that, if $s^2 = \Delta^2/\sigma^2 \geq K^\epsilon \vee \sqrt{\log(n)}$, then perfect clustering is possible as long as $n \geq \text{poly}(p, K^{1/\epsilon})$. Their procedures share connections makes heavy use of high-moment estimation. In summary, when the sample size is really large, there does not seem to exist a significant gap between polynomial-time and non-polynomial time methods, although procedures that seem to fill the gap do not proceed from distance-clustering ideas but rely on Gaussian mixture density estimation techniques.

In contrast, none of the aforementioned result is valid in a high-dimensional setting where $p \gtrsim n$. In fact, the problem seems to be qualitatively different in this regime. As an example, we have considered in [A11] an asymptotic toy problem[7] where both p and n go to infinity while n/p converges to a constant $\alpha \in (0, \infty)$. K is fixed but should be considered as large. Then, we consider a K-group mixture model with common covariance matrix $\mathbf{\Sigma} = \mathbf{I}_p$ and where the means $\mu_1, \dots \mu_K \in \mathbb{R}^p$ are sampled in the following way: for each $i = 1, \dots, p$, the vectors $(\mu_{1,i}, \mu_{2,i}, \dots, \mu_{K,i})$ are

[6] recall that for $n \leq p$, we have $s^2 = \Delta^2/\sigma^2$

[7] Actually, this work also deals with other statistical problems such as sparse PCA or submatrix localization.

independent and follow the Gaussian distribution $\mathcal{N}(0, \rho/p(\mathbf{I}_K - K^{-1}\mathbf{J}_K))$ so that the means μ_k's are constrained to satisfy $\sum_{k=1}^K \mu_k = 0$. Equipped with this notation, we have $\mathbb{E}[\|\mu_k - \mu_l\|_2^2] = 2\rho$. Then, the counterpart $\underline{s}^2$ of signal-to-noise ratio s^2 of the previous section where we replace Δ by the expected distance and n/p by its limit simplifies as $\underline{s}^2 = 2\rho \wedge \frac{4\rho^2\alpha}{K}$. In [A11], we are interested in the slightly different problem of testing whether the matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has been sampled according to this model against the alternative that $\mathbf{X}$ is made of i.i.d. standard normal entries.

Theorem 4.6 (Theorem 3 in [A11]). *Define* $\rho^{upper} = 2\sqrt{K\log(K)/\alpha} + 2\log(K)$ *and* $\rho^{lower} = \sqrt{2(K-1)\log(K-1)/\alpha}$ *for* $K > 3$. *Then, detection is possible when* $\rho > \rho^{upper}$ *and detection is impossible when* $\rho < \rho^{lower}$.

With our parametrization $\underline{s}^2$, the condition $\rho > \rho^{upper}$ corresponds to $\underline{s}^2 \gtrsim \log(K)$ and therefore matches the positive part of Conjecture 4.1. The minimax lower bound only matches the conjecture when $\alpha \leq K/\log(K)$ which corresponds to $n \leq pK/\log(K)$.

What is particularly appealing with this toy model is that the columns ot the matrix $\mathbf{X}$ are independent normal and that their covariance matrix is a rank K-perturbation of the identity matrix. In this asymptotic setting where p/n converges to a constant, the behavior of the spectrum of $\mathbf{X}$ is predicted by the so-called BBP phase transition [13] in random matrix theory. In particular, tests based on the largest singular values of $\mathbf{X}$ cannot achieve detection if $\rho\sqrt{\alpha} < K - 1$, whereas the largest singular value of $\mathbf{X}$ achieves detection if $\rho\sqrt{\alpha} > K - 1$. For $\alpha \leq 1$ (which corresponds to $p > n$), then one can reparametrize this condition to $\underline{s}^2 \geq K - 1$. Obviously, this does not perclude the existence of a test which is not based on the largest singular values, but in this high-dimensional setting, this seems difficult to improve over the spectral methods. See also [144] for other non-rigorous physical arguments that support this conjecture. This leads us to the following conjecture for partial recovery.

Conjecture 4.2 (Statistical-computational gap). *In a high dimensional setting where* $p \geq nK$, *no polynomial-time procedure is able to achieve partial recovery when* $s^2 \lesssim K$.

Obviously, we cannot reasonably hope to solve this conjecture unconditionally. Still, one could hope to build upon the recent series of work of Brennan and Bresler [33, 34, 36, 35] to establish a computational lower bound conditionally to the hardness of planted clique.

To conclude this section, we come back to the low-dimensional case. We explained that a condition of the form $s^2 \gtrsim K^\epsilon$ with $\epsilon \in (0, 1/2)$ is sufficient for partial recovery in polynomial time provided that the sample size n is large enough. The exact dependency of this minimal sample size seems really challenging to pinpoint. Still, the available polynomial procedures [67, 111, 136] that beat the $s^2 \gtrsim K$ boundary are really tailored to the Gaussian mixture density. In particular, resort to high-empirical moments of the data. In some ways, these procedures rely on the density of the distribution to estimate the parameters. This contrasts with typical clustering procedure such as K-means which only depends on the distance between the points. Hence, one may wonder whether the condition $s^2 \gtrsim K$ becomes a barrier for polynomial-time algorithms if we are working in another model where only the distances matter. A good candidate for this approach is the so-called semi-random model [27] perspective, where an adversary is allowed to modify the data but in a specific way which agrees with the hidden partition. This framework has been fruitfully applied to stochastic block models [159].

Awasthi and Vijayaraghavan have recently introduced in [11] such a suitable semi-random model for Gaussian mixtures. Given the partition G^* and the unperturbed observations $\mathbf{X}$, the adversary is allowed to modify each row X_a by possibly moving it closer to the mean $\mu_{k^*(a)}$. More specifically, the adversary picks some $\alpha_a \in [0, 1]$ and defines $Y_a = \mu_{k^*(a)} + \alpha_a(X_a - \mu_{k^*(a)})$. The statistician

is then given the matrix $\mathbf{Y} \in \mathbb{R}^{n\times p}$ of perturbed observations of $\mathbf{X}$. Obviously, this semi-random model completely breaks down the distribution of the Gaussian mixture model, while not making the clustering problem much more difficult. In some way, this could allow us to decipher the clustering viewpoint from the density estimation perspective.

In their paper, Awasthi and and Vijayaraghavan [11] establish that a suitable variant of Lloyd's algorithm achieves partial recovery in the semi-random model when $s^2 \gtrsim K\log(n)$ and the sample size n is large. It seems possible to extend some of the arguments in a high-dimensional setting. This leads us to the following stimulating conjecture.

Conjecture 4.3 (Statistical-computational gap for the semi-random model)**.** *No polynomial-time procedure is able to achieve partial recovery in the semi-random model of [11] when* $s^2 \lesssim K$.

4.4 Variable clustering

In this section, we move from point clustering and graph clustering to a slightly different problem which is referred to as *variable clustering*. The material described in this section mainly comes from [A7].

The problem of variable clustering is that of identifying similar variables in a n-dimensional random vector $X = (X_1, \dots, X_n)$ based on a p-sample[8] of the vector X. Here we do not require that two similar variables are highly-correlated but rather that they tend to share the same covariance with all the remaining variables. In [A7], we introduce a planted partition model, called G^*-block covariance model to formalize this problem.

Given a partition G^*, we associate a membership matrix $\mathbf{A} \in \mathbb{R}^{p\times K}$ defined by $\mathbf{A}_{ak} = 1$ if $a \in G^*_k$, and $\mathbf{A}_{ak} = 0$ otherwise.

Definition 4.7 (G^*-block covariance Model)**.** *Consider an n-dimensional mean zero random vector X with covariance matrix $\mathbf{\Omega}$. For a partition G^* of $[n]$, we say that X has a G^*-block covariance structure if*

$$\mathbf{\Omega} = \mathbf{A}\mathbf{C}^*\mathbf{A}^T + \mathbf{\Gamma} \ , \tag{4.8}$$

where $\mathbf{A}$ is relative to G^*, $\mathbf{C}^*$ is a symmetric $K \times K$ matrix, and $\mathbf{\Gamma}$ is a diagonal matrix.

With this property (4.8), for any two distinct coordinate $(a, b) \in [n]$, we have $\mathbf{\Omega}_{a,b} = \mathbf{C}^*_{k^*(a),k^*(b)}$. In other words, the covariance between any two variables only depends on the groups of these two variables. This structure of block-constant covariance matrix has been observed to hold, empirically, in a number of recent studies on the parcelation of the human brain, for instance [134, 100, 63, 202]. See also some real-world applications in [A7].

Connection with SBM. In some way, the definition of G^*-block covariance models is a covariance analogue of the stochastic block model. Indeed, with the notation of Definition 4.2, the expected adjacency matrix turns out to decompose as $\mathbb{E}[\mathbf{X}] = \mathbf{A}\mathbf{Q}\mathbf{A}^T - \mathrm{Diag}(\mathbf{A}\mathbf{Q}\mathbf{A}^T)$ because the diagonal of $\mathbf{X}$ is zero. Hence, the expected matrix of observation is a block-constant matrix (up to the diagonal). By contrast, in the G^*-block models, the covariance matrix of the random vector X is a block constant matrix (up to the diagonal).

Let us briefly provide two examples of G^*-block covariance models to illustrate the versatility of the model.

[8] It is more standard to write n for the sample size and p for the number of variables. Here, we exchange both notation for reasons of coherence to be explained later.

4.4.1 G^*-Latent Model

In a latent factor model, the components of X that belong to the same group can be decomposed as a sum of common latent factor (at the group level) and an uncorrelated fluctuation (at the individual level). More precisely, we have the following decomposition

$$X_a = Z_{k^*(a)} + E_a \ , \tag{4.9}$$

where $\mathrm{Cov}(Z_{k(a)}, E_a) = 0$ and the individual fluctuations E_a are uncorrelated. Writing $\mathbf{\Gamma}$ for the diagonal covariance matrix of E and $\mathbf{C}^*$ for the $k \times k$ covariance matrix of Z, one easily checks that that a G^*-latent model is a G^*-block covariance model.

Conversely, if X has a G^* block covariance structure and is normally distributed, then X can also be written as a G^*-latent model provided that $\mathbf{C}^* \succeq 0$. Note that $\mathbf{C}^*$ is not necessarily semi-definite positive in (4.8).

G^*-Latent Models and Gaussian mixture models with random means. Consider a Gaussian G^*-latent model whose covariance matrix $\mathbf{\Gamma}$ of E in (4.9) is block constant, ie there exists $\gamma \in \mathbb{R}_+^K$ such that $\mathbf{\Gamma}_{aa} = \gamma_{k^*(a)}$. Suppose that we observe a p-sample of X which is gathered in a $p \times n$ matrix $\mathbf{X}$. We also write $\mathbf{Z} \in \mathbb{R}^{K \times p}$ for the corresponding latent factor models. It turns out that, conditionally to $\mathbf{Z}$, $\mathbf{X}$ is distributed as a p-dimensional Gaussian mixture model with partition G^*, means $\mu_k = (\mathbf{Z}_{k,1}, \mathbf{Z}_{k,2}, \dots, \mathbf{Z}_{k,p})^T$ and covariances $\mathbf{\Sigma}_k = \gamma_k \mathbf{I}_p$. As a consequence, recovering the hidden partition in a Gaussian G^*-latent model is equivalent to recovering the partition in a spherical Gaussian mixture models with random means[9]. Note that, depending on $\mathbf{C}^*$, the means μ_k's may exhibit particular geometries.

4.4.2 Ising Block Model

Beyond normal distributions, G^*-block covariance models encompass other models that do not exhibit a latent structure. The Ising Block Model has been proposed in [21] for modelling social interactions. Here, the joint distribution of $X \in \{-1, 1\}^n$ is given by

$$f(x) = \frac{1}{\kappa_{\alpha,\beta}} \exp\left[\frac{\beta}{2p} \sum_{a \sim b} x_a x_b + \frac{\alpha}{2p} \sum_{a \nsim b} x_a x_b\right] \ , \tag{4.10}$$

where $\kappa_{\alpha,\beta}$ is a normalizing constant, and the notation $a \sim b$ means that both a and b belong to the same group of a partition G^*. The variables X_a may for instance represent the votes of U.S. senators on a bill [15]. For $\alpha > \beta$, the density (4.10) models the fact that senators belonging to the same political group tend to share the same vote. By symmetry of the density f, the covariance matrix $\mathbf{\Omega}$ of X decomposes as a G^*-block covariance model $\mathbf{\Omega} = \mathbf{A}\mathbf{C}^*\mathbf{A}^T + \mathbf{\Gamma}$ where $\mathbf{\Gamma}$ is diagonal.

4.4.3 Recovery bounds for G^*-block covariance models

In [A7], we introduce several metrics to quantify the distance between components that do not belong to the same group G^*. For the sake of conciseness, we discuss here one such distance $\Delta(\mathbf{C}^*)$, which is mostly relevant when $\mathbf{C}^*$ is semi-definite positive. We define

$$\Delta^2(\mathbf{C}^*) = \min_{j \neq k} \mathbf{C}^*_{jj} + \mathbf{C}^*_{kk} - 2\mathbf{C}^*_{jk} \tag{4.11}$$

[9] In fact, this analogy was already made (but without formalization) when discussing the results of [A11] in the previous section.

If $\mathbf{X}$ is also a G^*-latent model as in (4.9), then $\Delta(\mathbf{C}^*) = \min_{j\neq k} \mathbb{E}[(Z_j - Z_k)^2]$ is the minimum expected distance between the latent factors.

Given a p-sample $\mathbf{X} \in \mathbb{R}^{n\times p}$ of X, our goal is to recover the hidden partition G^* of variables from the data. In light of the analogy between Gaussian G^*-latent models and Gaussian mixture models, it is tempting to apply a corrected version of relaxed K-means as in Section 4.2.2.2. The correction of the K-means criterion is not required if the matrix $\mathbf{\Gamma}$ in (4.8) is proportional to the identity, but is important when the noise levels $\mathbf{\Gamma}_{jj}$ vary too much. To ease the presentation, we assume in the next theorem that the true partition G^* is approximately balanced.

Theorem 4.8 (Theorem 5.3 in [A7])**.** *Assume that the random vector X is sub-Gaussian and has a G^*-block covariance structure (4.8). Provided that*

$$\Delta^2(\mathbf{C}^*) \gtrsim \|\mathbf{\Gamma}\|_{op}\left[\sqrt{\frac{K(\log(n)\vee K)}{np}} + \frac{\log(n)\vee K}{p}\right] , \tag{4.12}$$

then corrected relaxed K-means perfectly recovers the partition G^* with high probability.

Since the above theorem applies to Gaussian G^*-latent model and since those can be interpreted as Gaussian mixture models (section 4.4.1), we can rewrite Condition (4.12) with the notation of Theorem 4.3. Indeed, the expected distance between the means is $p\Delta^2(\mathbf{C}^*)$. With this notation, the counterpart of s^2 is $s^{*2} = \frac{p\Delta^2(\mathbf{C}^*)}{\|\mathbf{\Gamma}\|_{op}} \wedge \frac{np\Delta(\mathbf{C}^*)}{K\|\mathbf{\Gamma}\|_{op}}$. Then, (4.12) is equivalent to Condition $s^{*2} \gtrsim [K \vee \log(n)]$, which also corresponds to the regime of perfect reconstruction in Theorem 4.3. In summary, both Theorems 4.3 and 4.8 conclusions are consistent.

Still, Theorem 4.8 applies beyond G^*-latent model and, in particular is also valid for Ising block Models (4.10).

From a minimax viewpoint, Condition (4.12) turns out to be minimax optimal, at least when $K \leq \log(n)$. For larger K, relaxed K-means seem to be sub-optimal by an additional K term. Again, this is consistent with the state-of-knowledge in Gaussian mixture modelling and the discussion in Section 4.3.

As a final point, we mention that the work [A7] on variable clustering precedes [A13]. Retrospectively, one may wonder whether it is not possible to establish partial recovery bounds in the spirit of Theorem 4.3 for general G^*-block covariance models.

Conjecture 4.4. *In the setting of Theorem 4.8, the penalized relaxed K-means estimator achieves $err(\hat{G}, G^*) \leq e^{-c's^{*2}}$ with high probability, provided that $s^{*2} \gtrsim K$.*

4.5 Detection thresholds for sparse GMM

In Section 4.2, we discussed the versatility and near-optimality of the K-means algorithm and its relaxations. Consider a Gaussian mixture model (as in Definition 4.1) with a common non spherical matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_K$. It was already pointed out that K-means is able to achieve good performances with respect to the distance $\Delta^2/\sigma^2 = \min_{j,k} \|\mu_j - \mu_k\|_2^2/\|\mathbf{\Sigma}\|_{op}$, but it does not adapt to the larger Mahalanobis distance $\min_{j,k}[\mu_j - \mu_k]^T\mathbf{\Sigma}^{-1}[\mu_j - \mu_k]$, on which the Bayes classifier is based. Another possible weakness of K-means, or more generally of classical distance clustering algorithms, is that they are not adaptive to specific structures of the data. For instance, in some high-dimensional clustering problems, it is hypothesized that the differences $\mu_k - \mu_j$ are sparse, which means that the two groups k and j only differ through a few features. The latter problem

is referred in the literature as sparse clustering. Hopefully, one would hope to build upon this structural assumption to improve the clustering error. Arguably, K-means (or classical spectral clustering algorithms) will not be able to deal efficiently with sparsity as these algorithms are invariant with respect to any rotation of the data.

This section is mainly devoted to a joint work with E. Arias-Castro [A17]. There, we provide some intuition on the sparse clustering problem through the prism of signal detection and feature selection in a sparse Gaussian mixture model. We consider a testing problem where, given a sample $X_1, \dots, X_n$ of p-dimensional random vector X, we want to test whether X is sampled according to a two-class Gaussian mixture distributions or from a single Gaussian distribution. More formally, we consider the general testing problem

$$H_0 : X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \mathbf{\Sigma}) \ , \quad \text{for some } \mu \in \mathbb{R}^p, \text{ and } \mathbf{\Sigma} \succeq 0 \ ; \tag{4.13}$$

versus

$$H_1 : X_1 \dots, X_n \overset{\text{iid}}{\sim} \pi\mathcal{N}(\mu_0, \mathbf{\Sigma}) + (1-\pi)\mathcal{N}(\mu_1, \mathbf{\Sigma}) \ , \text{for some } \mathbf{\Sigma} \succeq 0, \ \mu_0 \neq \mu_1 \in \mathbb{R}^p, \text{ and } \pi \in (0,1). \tag{4.14}$$

We are specifically interested in settings where the difference in means $\Delta\mu := \mu_1 - \mu_0$ is k-sparse. Arguably, the detection problem is simpler than the clustering problem, as in low-dimension (e.g. $p = 1$), one is able to decipher a mixture of two gaussian when $\Delta\mu$ arbitrarily small (provided that the sample size is large enough), whereas partial recovery of the unknown corresponding partition is possible only if $\Delta\mu = \Omega(1)$. Still, the qualitative difference between detection and clustering become much thinner in high-dimensional sparse regimes.

It turns out that the detection problem of (4.13) against (4.14) depends a lot on the knowledge of the covariance matrix $\mathbf{\Sigma}$. We first discuss the case of known covariance and then turn to the unknown covariance.

4.5.1 Known Covariance

In this subsection, we assume that $\mathbf{\Sigma}$ is known and that the problem is high-dimensional. We aim at characterizing the minimum signal-to-noise ratio $R_0 = \Delta\mu^\top \mathbf{\Sigma}^{-1} \Delta\mu$ in Mahalanobis distance which allows to decipher (4.13) from (4.14). This quantity is called the minimax detection distance, in accordance with the formalism of Chapter 2. The results are summarized in Table 4.1 below.

Sparsity regimes	Minimax detection distances	Near-optimal test
$k \leq \dfrac{n}{\log(ep/n)}$	$\left[\dfrac{k\log(ep/k)}{n}\right]^{1/2}$	Top sparse eigenvalue
$\dfrac{n}{\log(ep/n)} \leq k \leq (np)^{\zeta/2}$	$\dfrac{k\log(ep/k)}{n}$	Top sparse eigenvalue
$k \geq \sqrt{np}$	$\sqrt{p/n}$	Top eigenvalue

Table 4.1: Minimax detection distances and near-optimal tests as a function of the sparsity k when $\mathbf{\Sigma}$ is known and $p \geq n$. The minimax detection distances are expressed in terms of the signal-to-noise ratio $R_0 = \Delta\mu^\top \mathbf{\Sigma}^{-1} \Delta\mu$. Here, ζ denotes any arbitrary constant in $(0,1)$. The top and top sparse eigenvalues are referring to the eigenvalues of a (modified version) of the empirical covariance matrix

Some comments are in order. First, for $k \geq \sqrt{np}$, the detection rate $\sqrt{p/n}$ corresponds, up to reparametrization, to $s^2 \asymp 1$ with the signal-to-noise ratio formalism of Section 4.2. This entails, that in the dense high-dimensional regime, detection arises roughly at the same time as partial

recovery of the partition. Besides, sparsity does not play a role in this regime as the detection distance is the same as for $k = p$.

For smaller k, the optimal detection rate is achieved by the top sparse eigenvalue of a modified empirical covariance matrix. Unfortunately, no algorithm is known for computing it in polynomial time. In [A17], we provide some polynomial time testing procedures (in the spirit of sparse PCA), but those suffer from suboptimal performances by a factor which can be as large as $\sqrt{k}$. This is not completely unexpected as this problem shares the same taste as sparse PCA.

In the last years, there has been a renewed interest in this two-group model of sparse clustering (with $\mathbf{\Sigma} = \mathbf{I}_p$). In particular, it was recently shown that this gap between polynomial-time methods and minimax ones turns out to be unavoidable [148].

4.5.2 Unknown Covariance

When the covariance matrix $\mathbf{\Sigma}$ is unknown, we first prove that, when $p \gg n$, the minimax detection rate for the Mahalanobis distance is exponentially large, even in the simplest situation where $k = 1$. This entails that, even for one-sparse differences, detection is not possible even for very large Mahalanobis distance. This is due to the fact that $p \gg n$, the covariance matrix cannot be efficiently estimated so that the direction of higher variation $\mathbf{\Sigma}^{-1/2}\Delta\mu$ is difficult to localize. For this reason, we consider the weaker loss function $R_1 = \|\Delta\mu\|_2^4/[\Delta\mu^T\mathbf{\Sigma}\Delta\mu] \leq R_0$, and we established the minimax detection distance for (4.13) against (4.14). Although I do not reproduce here the precise results, three interesting phenomenons can be mentioned. First, this minimax detection distance highly depends on the symmetry of mixture that is whether $\pi = 1/2$ or $\pi \neq 1/2$ in (4.14), the asymmetric case $\pi \neq 1/2$ being much easier. Second, a computational-statistical trade-off seems to arise, even when $k = p$ (no sparsity). Third, the minimax optimal tests are based on a new approach which aims at finding the direction on which the first absolute moment is the highest possible. Recently, this approach has been used by Davis et al. [64] (see their Appendix E) to achieve optimal clustering rate in non-sparse clustering with unknown covariance matrix $\mathbf{\Sigma}$.

In the last year, recent works have unveiled interesting phenomenons in (non-sparse) clustering with unknown common covariance $\mathbf{\Sigma}$. In particular, a computational gap seems to arise when $p \in [n, n^2]$ and when the objective is to recover the cluster at the optimal rate in Mahalanobis distance –see [64]. These results to not overlap much with our own work as we were rather interested in the weaker metric $R_1 \leq R_0$[10] but focus on a higher-dimensional setting $p \gg n$. In any case, there remain many open problems to understand the optimal clustering rates and their counterpart with polynomial-time procedures when one considers a possibly-sparse mixture with unknown covariance matrix.

A closing comment from the application side

This manuscript is mainly dedicated to my work on mathematical statistics. Still, as a research scientist within INRAE, I am also involved and exposed to practical clustering and bi-clustering problems (e.g. in [A19]) mostly with assessments of cultivated biodiversity[11] and with seed exchange networks.

Although my own theoretical works are mainly concerned with K-means, SDP relaxation of K-means or, to a less extent, spectral clustering algorithms, I do not tend to apply these methods, but rather focus on likelihood-based methods optimized by variational-EM algorithms.

[10] Recall that R_0 is the Mahalanobis distance.

[11] https://forgemia.inra.fr/nicolas.verzelen/blockmodels4inventories

The main reason for this is that clustering or bi-clustering problems do not arise alone. In practice, the dataset to cluster is coming along with many control and explanatory covariates. Although there has been a recent interest in SDP formulation and spectral methods that use side information, the available methods are much less versatile than the machinery on likelihood maximization techniques [155]. See for instance the econetwork package[12].

[12] https://plmlab.math.cnrs.fr/econetproject/econetwork

Chapter 5

Other unsupervised-learning Problems

This last chapter is dedicated to other unsupervised learning problems that I have been interested in in the last few years. In some ways, this chapter is much more heterogeneous than the previous ones: the models, the objectives, and the techniques highly differ from one problem to another. Still, all these problems have in common the fact that we aim to recover some latent labels in some objects. In clustering, these hidden labels correspond to a partition of the objects. Here, we consider different objectives: recovering a partition with side-information on its form (change-point detection/segmentation problems), recovering a permutation of the objects (seriation/ranking problems).

The first section is mainly dedicated to change-point detection and is based on two joint works with M. Fromont, M. Lerasle, P. Reynaud-Bouret [P4] and with E. Pilliat and A. Carpentier [P3]. In the second section, I shortly discuss a recent work on seriation [P1] (with C. Giraud and Y. Issartel) together with some on-going work on ranking problems.

5.1 Change-point detection

Change-point detection has a long history that comes back to the seminal work of Wald [191] and lead to flourishing lines –see [187, 165] for recent surveys. Earlier work were mainly devoted to the problems of detecting and localizing change in the mean of univariate time series. Important applications e.g. in genomics [166] or finance have spurred a recent trend towards the detection of variations in more complex times series that live in a high-dimensional space [124] or even belong to a non-Euclidean space [59]. In this section, I describe two recent joint works. In the first one [P4], we focus on the toy model of Gaussian univariate mean change-point detection, and derive the tight optimal rates for detection and localization of the change-points. In the second one [P3], we introduce a strategy for general change-point problems and we apply it to several settings including changes in a high-dimensional mean vector, changes in the covariance matrix, or non-parametric changes in the distribution... In each case, this allows us to derive the tight optimal conditions for change-point detection.

5.1.1 Optimal univariate mean-Change-point detection and Localization

Let us consider the prototypical problem of univariate change-point analysis. Let $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ denote a random vector with

$$Y_i = \theta_i^* + \epsilon_i \ , \quad i = 1, \ldots, n \ , \tag{5.1}$$

where $\theta^* = (\theta_1^*, \ldots, \theta_n^*)$ in $\mathbb{R}^n$ and the *noise* random vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is made of independent standard Gaussian random variables. Given Y, the objective is to find the coordinates at which the mean vector θ^* is varying –those are called change-points. Since we consider θ^* as a piece-wise constant vector, we may define it through its change-points. There exists an integer $0 \leq K \leq n-1$, a vector of integers $\tau^* = (\tau_1^*, \ldots, \tau_K^*)$ satisfying $1 = \tau_0^* < \tau_1^* < \ldots < \tau_K^* < \tau_{K+1}^* = n+1$, a vector $\mu = (\mu_1, \ldots, \mu_{K+1})$ in $\mathbb{R}^{K+1}$ satisfying $\mu_k \neq \mu_{k+1}$ for all k in $\{1, \ldots, K\}$ such that $\theta_i^* = \sum_{k=1}^{K+1} \mu_k \mathbf{1}_{\tau_{k-1}^* \leq i < \tau_k^*}$. See Figure 5.1 below. Then, τ_k^* is called the *position* of the k-th change-point and $\Delta_k = \mu_{k+1} - \mu_k$ is called the *height* of the k-th change-point. We focus here on the situation where the number of change-points K is unknown but is larger than one –see [P4] for the specific case $K \leq 1$.

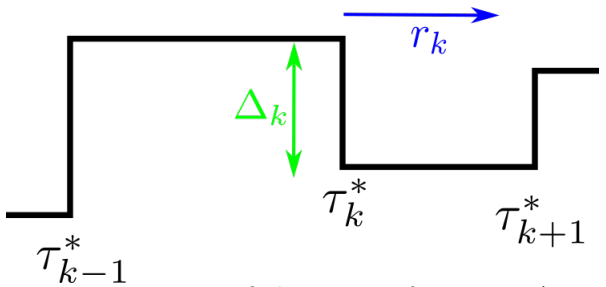

Figure 5.1: Change-points τ_{k-1}^*, τ_k^*, τ_{k+1}^* of the vector θ^*. Here, Δ_k stands for the height of the k-th change-point and $r_k = \sqrt{\frac{(\tau_{k+1}^* - \tau_k^*)(\tau_k^* - \tau_{k-1}^*)}{\tau_{k+1}^* - \tau_{k-1}^*}}$ for the corresponding length.

With this notation, our objective is to build an estimator $\hat{\tau}$ of τ^* of the change-points. Such an estimator is to be analyzed from two related but distinct perspectives: *detection* and *localization*.

(a) We say that the vector $\widehat{\tau}$ *detects* a true change-point τ_k^* if there exists an index l such that $\widehat{\tau}_l \in [(\tau_{k-1}^* + \tau_k^*)/2; (\tau_k^* + \tau_{k+1}^*)/2]$, i.e. $\widehat{\tau}_l$ is closer to τ_k^* than any other true change-point. Conversely, we say that $\widehat{\tau}$ detects a *spurious* change-point if there exist τ_k^*, $\widehat{\tau}_l$ and $\widehat{\tau}_{l'}$ such that both $\widehat{\tau}_l$ and $\widehat{\tau}_{l'}$ are close to τ_k^*, in the sense that they belong to $[(\tau_{k-1}^* + \tau_k^*)/2; (\tau_k^* + \tau_{k+1}^*)/2]$. The challenge is then to build an estimator $\widehat{\tau}$ which, with high probability, detects all *significant* true change-points and does not detect any spurious change-point. For this purpose, we need to introduce the energy $\mathbf{E}_k$ of a true change-point τ_k^*.

$$\mathbf{E}_k = |\Delta_k| \sqrt{\frac{(\tau_{k+1}^* - \tau_k*)(\tau_k^* - \tau_{k-1}^*)}{\tau_{k+1}^* - \tau_{k-1}^*}} \ . \tag{5.2}$$

Up to multiplicative factor, $\mathbf{E}_k^2 \asymp \Delta_k^2[(\tau_{k+1}^* - \tau_k*) \wedge (\tau_k^* - \tau_{k-1}^*)]$ is the square change-point height times the distance of τ_k^* to the closest change-point. Intuitively, $\mathbf{E}_k^2$ is the bias that we suffer if we estimate θ^* by a piece-wise constant vector with change-points in $(\tau_1^*, \tau_2^*, \ldots, \tau_{k-1}^*, \tau_{k+1}^*, \ldots, \tau_K^*)$. Similar notions of energies of a change-point often appear in the literature [194, 87, 85].

(b) If a true change-point τ_k^* has been detected, then one would aim at localizing it as best as possible, that is at having the distance $d_{H,1}(\widehat{\tau}, \tau_k^*) = \min_{i=1,\dots,|\widehat{\tau}|} |\widehat{\tau}_i - \tau_k^*|$ between τ_k^* and the closest estimated change-point as small as possible. This quantity $d_{H,1}(\widehat{\tau}, \tau_k^*)$ is referred as the *localization error* of $\widehat{\tau}$ for τ_k^*. Aside from the localization error of a specific change-point, one also may be interested in more global localization error such as the Hausdorff error or the Wasserstein error which respectively correspond to the supremum and the sum of the localization errors.

State of the art. It was recently proved in Wang et al. [194] that as long as the minimum energy $\min_{k=1,\dots,K} \mathbf{E}_k^2$ is large compared[1] to $\log(n)$, then all true change-points are detected, and those are uniformly localized at the rate $\log(n)/[\min_k \Delta_k^2]$ –see also [87, 195, 85] for related results. If all change-points are equi-spaced and the jump size Δ_k^2 are of the same order, these energy condition and localization turn out to be minimax optimal up to $\log(n)$ terms [194]. Interestingly, such near optimal properties are achieved by both the penalized least-square estimators and greedy methods based on the CUSUM statistic [194].

Our contribution. In [P4], we establish the tight minimal condition for change-point detection as well as the tight localization rate for a significant change-point. In particular, we close the logarithmic gaps between the known minimax lower and upper bounds. Besides, both detection and localization properties are still proved to hold in settings where we allow for an arbitrarily large number of nuisance change-points that have a low energy. More precisely, we first establish that a change-point τ_k^* is significant and therefore can be detected as long as

$$\mathbf{E}_k^2 \gtrsim \log\left(\frac{n}{(\tau_{k+1}^* - \tau_k^*) \wedge (\tau_k^* - \tau_{k-1}^*)}\right) . \tag{5.3}$$

This condition turns out to minimax. In comparison to the $\log(n)$ condition of [194], the logarithm term in (5.3) can be much smaller. For instance, if all the change-points are nearly equi-spaced, it simplifies as $\mathbf{E}_k^2 \gtrsim \log(K)$. Besides, our procedures detects change-points satisfying (5.3) even in the presence of other change-points with a very small energy.

Regarding the localization error, we establish a transition phenomenon from a regional to a local problem. As soon as the energy of a true change-point τ_k^* satisfies (5.3), then it can be localized at the rate $1/(\Delta_k^2)$. Besides, the localization errors of all significant change-points can behave like independent sub-exponential random variable. This allows us to recover the tight localization errors (with the right logarithm) for a specific change-point ($d_{H,1}(\widehat{\tau}, \tau_k^*)$), for Hausdorff distance, and for Wasserstein distance.

We introduce two procedures achieving all these optimality properties. The first one is a penalized least-squares type estimator with a multiscale penalty that promote equi-spaced change-points positions. As the corresponding penalty is additive, this estimator is easily computed by (pruned) dynamic programming [128]. In contrast to the BIC-type penalty studied recently in [194], this allows us to recover the optimal logarithmic terms. As an alternative to the penalized least-squares estimator, we promote a two-step bottom-up aggregation method based on the aggregation of many CUSUM tests. It is shown to satisfy the same optimality property as the previous penalized procedure while enjoying a quasi-linear computational complexity.

[1] In fact, the results in [194, 87, 195, 85] are slightly weaker than that, because they consider the smaller $\mathbf{E}_{min} = \min_k |\Delta_k| \min_k |\tau_{k+1}^* - \tau_k^*|^{1/2}$ which is smaller than $\min_k \mathbf{E}_k$

5.1.2 Optimal change-point detection for general problems

We now turn to more general change-point problems beyond the previous toy model. In the most general form, we consider a sequence $Y = (Y_1, Y_2, \dots, Y_n)$ in some measured space $\mathcal{Y}^n$. For $t = 1, \dots, n$, we write $\mathbb{P}_t$ for the marginal distribution of Y_t. We are also given a functional Γ mapping the probability distribution $\mathbb{P}_t$ to some space $\mathcal{V}$. With this formalism, the purpose of change-point analysis is to detect changes in the sequence $(\Gamma(\mathbb{P}_1), \Gamma(\mathbb{P}_2), \dots, \Gamma(\mathbb{P}_n))$ in $\mathcal{V}^n$ and to estimate the positions of these changes.

Up to our knowledge, this general framework encompasses most offline change-point detection problems. For Gaussian mean univariate change-point setting, we have $\mathcal{Y} = \mathbb{R}$, the distribution $\mathbb{P}_t$ corresponds to the normal distribution with mean $\theta_t \in \mathbb{R}$ and variance σ^2 and $\Gamma(\mathbb{P}_t) = \theta_t$. In the (heteroscedastic) mean univariate change-point problem, the distribution $\mathbb{P}_t$ is not necessarily Gaussian and, in particular, the variance of Y_t is allowed to vary with t. Still, one is only interested in detecting variations of $\Gamma(\mathbb{P}_t) = \int x d\mathbb{P}_t = \mathbb{E}[Y_t]$. By contrast, in the *variance* univariate change-point problems, one focuses on changes in the variance of Y_t. This can be done by considering $\Gamma(\mathbb{P}_t) = \int x^2 d\mathbb{P}_t - [\int x d\mathbb{P}_t]^2 = \mathrm{Var}(Y_t)$. If one is interested in possibly nonparametric changes in the distributions, then the functional Γ is simply taken to be the identity map. In semi-parametric quantile change-point detection [127], the univariate distributions $\mathbb{P}_t$ can be arbitrary whereas $\Gamma(\mathbb{P}_t)$ is a quantile of $\mathbb{P}_t$.

Extending the notation of the previous subsection, we define an integer $0 \leq K \leq n-1$ and a vector of integers $\tau^* = (\tau_1^*, \dots, \tau_K^*)$ satisfying $1 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = n+1$ such that $\Gamma(\mathbb{P}_t)$ is constant over each interval $[\tau_k, \tau_{k+1} - 1]$ and $\Gamma(\mathbb{P}_{\tau_k - 1}) \neq \Gamma(\mathbb{P}_{\tau_k})$. In [P3], we introduce a generic approach for estimating τ^*.

In a nutshell, our generic procedure is based on a bottom-up aggregation of local homogeneity tests. This idea is to compute homogeneity tests $T_{l,r}$ of the hypothesis {$\Gamma(\mathbb{P}_t)$ is constant on [l-r,l+r)}, this is for a suitable collection of scales r and locations l. Then, we aggregate all these tests $T_{l,r}$ to build $\widehat{\tau}$ by adding locations l such that $T_{l,r} = 1$ for some r and such that, at all smaller scales r', we did not find any change point at positions l' such that $[l' - r', l' + r') \cap [l - r, l + r) \neq \emptyset$. Interestingly, we are able to translate properties of the collection of tests $(T_{l,r})$ into properties of the corresponding estimator $\widehat{\tau}$ of τ^*. In particular, a control of the family-wise-error rate (FWER) of $(T_{l,r})$ implies that $\widehat{\tau}$ does not estimate any *spurious* change-points. Conversely, control of type II error probabilities of some specific tests $(T_{l,r})$ imply that $\widehat{\tau}$ *detects* specific change-points. The idea of bottom-up aggregation of local homogeneity tests is not new and related (but still different) procedures have been introduced e.g. in [137, 48]. Nevertheless, none of these procedures come up with statistical guarantees, except in the toy model of Gaussian mean univariate change-point discussed in the previous subsection.

In summary, we are able to reduce the problem of change-point detection to multiple homogeneity testing. In fact, rather than general homogeneity tests, we are more interested in two-sample testing problems of the form {$\Gamma(\mathbb{P}_t)$ is constant on [l-r,l+r)} against {$\Gamma(\mathbb{P}_{l+r}) = \ldots = \Gamma(\mathbb{P}_{l-1}) \neq \Gamma(\mathbb{P}_l) = \ldots = \Gamma(\mathbb{P}_{l-r})$}. Hence, change-point detection mostly boils down to optimal multiple two-sample testing.

As an application of the generic procedure, we apply in [P3] our methodology to *sparse high-dimensional Gaussian mean* change-point model, *high-dimensional covariance* change-point models and *non-parametric* change-point model. In each case, we derive the tight minimal energy conditions for the detection of change-points and we show that our procedure achieves detection under these minimal conditions. Thereby, we improve previous results [167, 193, 48] in the literature. Interestingly, our methodology allows for the presence of nuisance low energy change-points (as in

the previous subsection) that are not to be detected, but do not either perturb the detection of significant change-points.

5.1.3 Research directions

Localization in general change-point problems. The generic procedure of the previous subsection leads to tight minimal detection properties in various settings. Unfortunately, it does not seem to exhibit optimal localization errors in general –recall the difference between detection and localization explained in Section 5.1.1. A second refinement step needs to be added in the procedure to yield minimax localization errors. In the univariate mean change-point problem of Section 5.1.1, there is a simple transition from detection to localization errors in the parametric rate $1/\Delta_k^2$. For more complex change-points problems, the situation is certainly trickier – see e.g. [192] for localization error in change-point detection in network time series. As a simple example, consider a mean multivariate problem where the θ_i^*'s are now p-dimensional multivariate vectors. In that case, a change-point can be detected as long as the difference Δ_k between the means at a true change-point τ_k^* is large enough in Euclidean norm. For very large values of the corresponding energy $\min(\tau_k^* - \tau_{k-1}^*, \tau_{k+1}^* - \tau_k^*)\|\Delta_k\|_2^2$, not only can the change-point be detected, but the direction of Δ_k can also be estimated, so that projecting the data on the corresponding estimated direction $\widehat{\Delta}_k$, one can hope to localize the change-point at the parametric rate $1/\|\Delta_k\|_2^2$. In contrast, if the signal strength is much lower, the change-point is able to be detected, but the direction Δ_k cannot be localized. For this reason, the localization error of $d_{H,1}(\widehat{\tau}, \tau_k^*)$ is expected to be much higher than $1/\|\Delta_k\|_2^2$ in this case and should in particular depend on p. Hence, we expect the optimal localization error to exhibit several regimes. In light of this, we doubt that it is possible to craft a generic procedure that achieves the optimal localization errors in various problems. Still, pinpointing the tight localization error in emblematic problems is an exciting open problem.

Open Problem 5.1 (Localization of change-points)**.** *Derive the tight localization error for sparse high-dimensional mean change-point problems and covariance change-point problems.*

Segmentation on general graphs. In some way, one can interpret the problem of change-point detection as a very specific instance of clustering problem where the clusters are constrained to be intervals of $[n]$. This explains why change-points can be detected and the unknown partition can be partially reconstructed under much weaker separation conditions than for Gaussian mixture models (see Chapter 4) where there is no constraint on the partition. Between these two extreme settings where the partition is completely arbitrary (Gaussian mixture models) or is extremely constrained (change-point detection), it would be really exciting to investigate the general problem of signal segmentation on a graph.

Open Problem 5.2 (Segmentation on general graphs)**.** *Consider an undirected graph $\mathcal{G} = ([n], E)$ with n vertices. For each vertex $i = 1, \dots, n$, one observes $Y_i \sim \mathcal{N}(\theta_i^*, \mathbf{I}_p)$, with unknown mean $\theta_i^* \in \mathbb{R}^p$. Let $G^* = (G_1^*, \dots, G_K^*)$ denote the partition of $[n]$ that groups together identical values of θ_i^*. Provided that the partition G^* has a small boundary on the graph $\mathcal{G}$, what is the minimal difference between the means so that one is able to partially recover G^*? What is the minimum reconstruction error of G^*?*

In the univariate case ($p = 1$), the denoising version of this problem where one aims at estimating θ^* in l_2 distance has attracted a lot of attention and is now well understood –see e.g. [81]. In contrast, there are much fewer results on the segmentation error – but see [203]. Even in the univariate case, the minimal separation condition for approximate recovery remains unknown when the number of groups K is larger than 2.

5.2 Seriation and localization in 1-dimensional space

Suppose that we are given n objects and that we observe a noisy affinity matrix $\mathbf{A} = (\mathbf{A}_{ij})_{1\leq i,j\leq n}$, which provides similarity measurements between pairs of objects. These may correspond to real valued scores or to binary information, as when the matrix $\mathbf{A}$ encodes a similarity graph.

In [P1], we consider the 1D latent localization problem, where we seek to recover the 1D latent positions of n objects from $\mathbf{A}$. Such problems arise in archeology for relative dating of objects or graves [174], in 2D-tomography for angular synchronization [179, 60], in bioinformatics for reads alignment in *de novo* sequencing [171], in computer science for time synchronization in distributed networks [99, 80], or in matchmaking problems [32].

In such 1D latent models [108], the symmetric affinity matrix $\mathbf{A}$ is assumed to be sampled as follows. $\mathbf{A}_{ii} = 0$ (by convention) and

$$\mathbf{A}_{ij} = f(x_i^*, x_j^*) + \mathbf{E}_{ij}, \quad \text{for } 1 \leq i < j \leq n \ , \tag{5.4}$$

where
(i) $x_1^*, \dots, x_n^*$ are n unobserved latent positions spread on the unit sphere $\mathcal{C}$ in $\mathbb{R}^2$,
(ii) $f : \mathcal{C} \times \mathcal{C} \to [0,1]$ is unobserved, symmetric, decreasing with the geodesic distance $d(x,y)$, and
(iii) $[\mathbf{E}_{ij}]_{1\leq i<j\leq n}$ are some independent sub-Gaussian random variables.

This non-parametric framework is very flexible for fitting pairwise affinity data. It encompasses the circular random geometric graph models and the toroidal statistical seriation models defined below. The main difference with the graphon model, discussed in Chapter 3, is that the function f is constrained to be decreasing with the distance.

Definition 5.1 (Random Geometric Graph [65, 68, 169, 97])**.** *Let $\mathcal{C}$ denote the unit sphere in $\mathbb{R}^2$ endowed with the geodesic distance d. In the circular random geometric graph model, the edges are sampled independently with probability $\mathbb{P}[\mathbf{A}_{ij} = 1] = g(d(x_i^*, x_j^*))$, where $g : [0,\pi] \mapsto [0,1]$ is a non-increasing function and $x_1^*, \dots, x_n^* \in \mathcal{C}$ are the latent positions of the nodes on the sphere.*

Definition 5.2 (Toroidal R-Matrices and Statistical toroidal Seriation)**.** *A Robinson matrix (R-matrix) is any symmetric matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ whose entries decrease when moving away from the diagonal, i.e. such that $\mathbf{B}_{i,j} \geq \mathbf{B}_{i+1,j}$ and $\mathbf{B}_{i,j} \geq \mathbf{B}_{i,j-1}$, for all $1 \leq j \leq i \leq n$. A matrix $\mathbf{F}$ is called a pre-R matrix, when there exists a permutation $\sigma \in \Sigma_n$ of $[n]$, such that $\mathbf{F}_\sigma = [\mathbf{F}_{\sigma(i),\sigma(j)}]_{i,j}$ is an R-matrix. The noisy seriation problem [84] amounts to find, from a noisy observation of a pre-R matrix $\mathbf{F}$, a permutation σ^* such that $\mathbf{F}_{\sigma^*}$ is a R-matrix. This problem appears in genomic sequencing [94], in interval graph identification [88], and in envelope reduction for sparse matrices [17]. Here, we are interested in a variation of problem where we consider the consider the set $[n]$ as a torus with the corresponding distance $d(i,j) = \min(|j-i|, |n+i-j|)$ for any $1 \leq i,j \leq n$. A toroidal R-matrix is any symmetric matrix $\mathbf{B}$ whose entries decrease when moving away from the diagonal with respect to the toroidal distance: $\mathbf{B}_{i,j} \geq \mathbf{B}_{i+1,j}$ when $d(i,j) < d(i+1,j)$ and $\mathbf{B}_{i,j} \geq \mathbf{B}_{i,j+1}$ when $d(i,j) < d(i,j+1)$. As above, a pre-toroidal R-matrix is defined as a permutation of a toroidal R-matrix and the statistical seriation model is defined analogously [172]. We can recast this model as a latent space model (5.4) on the regular grid $\mathcal{C}_n$ of the unit sphere $\mathcal{C}$ corresponding to the n-th unit roots, endowed with the geodesic distance on $\mathcal{C}$.*

5.2.1 Our contribution

In [P1], our overall goal is to recover from $\mathbf{A}$ the latent positions $x^* = (x_1^*, \dots, x_n^*) \in \mathcal{C}^n$, with some high-confidence, simultaneously for all individual positions x_i^*. Since the global error of an

estimator $\hat{x}$, say $d_2(\hat{x}, x^*) = \sqrt{\sum_{i=1}^n d(\hat{x}_i, x_i^*)^2}$, provides limited information on each individual error $d(\hat{x}_i, x_i^*)$, we focus instead on the maximum error

$$d_\infty(\hat{x}, x^*) = \max_{i=1,\dots n} d(\hat{x}_i, x_i^*) \ . \tag{5.5}$$

Unfortunately, controlling $d_\infty(\hat{x}, x^*)$ is impossible as the latent positions are not identifiable from $\mathbf{A}$. Indeed, for any bijective map $\varphi : \mathcal{C} \to \mathcal{C}$, we have $f(x, y) = f \circ \varphi^{-1}(\varphi(x), \varphi(y))$ for all $x, y \in \mathcal{C}$, with the notation $f \circ \varphi^{-1}(x, y) := f(\varphi^{-1}(x), \varphi^{-1}(y))$. Even if we would enforce some strong shape constraints, such as $f(x, y) = 1 - \alpha d(x, y)$ with $\alpha > 0$, the distribution of the data would still be invariant by orthogonal transformation of the latent positions[2] since $f(x, y) = f(Qx, Qy)$ for any orthogonal transformation Q of $\mathcal{C}$. Informally, our remedy is to control $d_\infty(\hat{x}, x^*)$ for *some* representative x^* of the latent positions.

In [P1], we propose some estimators $\hat{x}$ achieving, with high-probability, a maximum error $d_\infty(\hat{x}, x^*)$ of the order of $\sqrt{\log(n)/n}$, under the assumptions that the latent positions $x_1^*, \dots, x_n^*$ are sufficiently spread on $\mathcal{C}$ and that $f(x, y)$ is a bi-Lipschitz function of $d(x, y)$. The $\sqrt{\log(n)/n}$ estimation rate is shown to be optimal. To the best of our knowledge, these are the first optimal results on maximum error $d_\infty(\hat{x}, x^*)$ in latent space models with unknown and non-parametric affinity function f.

Our estimation procedures proceeds in two main stages: (1) we start with an initial estimator $\hat{x}^{(1)}$ with a global control in $d_1(x, y) := \sum_{i=1}^n d(x_i, y_i)$ distance; (2) then, for each point, we refine this first estimator to get a control in d_∞ distance. This second step has a polynomial computational complexity and, under appropriate assumptions, it allows to recover the desire rate $d_\infty(\hat{x}^{(2)}, x^*) = O(\sqrt{\log(n)/n})$ provided that the have an initial control $d_1(\hat{x}^{(1)}, x^*) = O(\sqrt{n \log(n)})$. We propose two estimators fulfilling this requirement:
(a) a first one, which requires no additional assumptions, but which has a super-polynomial computational complexity;
(b) a spectral seriation algorithm, adapted from [172], which has a polynomial computational complexity, but for which we prove a $O\big(\sqrt{n \log(n)}\big)$ control only for a class of random geometric graphs.

5.2.2 Related work and Perspectives

In the last decade, the analysis of interaction data has spurred a lot of work in machine learning and statistics. Most of them rely on the case where the affinity function is known or belong to a known parametric model. Our modeling assumptions in [P1], with only shape constraints on f, offer a more flexible setting to fit data. In this subsection, we discuss related work and perspectives on such non-parametric models.

Seriation from pairwise affinity. In [P1], we deal with the toroidal seriation problem (see Definition 5.2). In the vanilla seriation problem, we recall that we are given a noisy observation $\mathbf{A}$ of a pre-R matrix $\mathbf{F}$ and we seek to find a latent order σ^* such that $\mathbf{F}_{\sigma^*}$ is a R-matrix. In the noiseless case, this can be done efficiently by spectral methods [9] or by convex optimization methods [84]. In the noisy case, Jannssen and Smith [119] have recently introduced as estimator achieving $\max_{i\in[n]} |\widehat{\sigma}_i - \sigma_i^*| \lesssim \sqrt{n} \log^5(n)$, under some complex assumptions on the matrix $\mathbf{F}$. Although their assumptions and the setting are slightly differs from ours, the localization rates are (up to logarithmic factors and to the scaling) comparable to ours for toroidal seriation. Beside,

[2] See [P1] for a proposer discussion of the identifiability issues

Cai and Ma [42] have recently established the optimal convergence rate of $\widehat{\sigma}$ under the additional assumption that the matrix $\mathbf{F}$ is Toeplitz. They also conjecture the existence a computational gap for this Toeplitz seriation problem. In any case, the optimal convergence rate for general noisy seriation remains unknown. Besides, the best possible performances of polynomial-time procedures remain largely unknown.

Open Problem 5.3. *For general unknown pre R-matrices $\mathbf{F}$, characterizing the best estimation error for σ^* which is achievable by a polynomial-time procedure.*

Ranking and Skills estimation. Moving away from affinity observations, we discuss a related class of problem where we want to rank players in a game. The observations $\mathbf{A}_{ij}$ now correspond to noisy comparisons between i and j – think e.g. as the results of a match between i and j. As in seriation problems, the goal is recover a latent order σ^* from the noisy matrix $\mathbf{A}$. Still, in contrast to seriation problems, the expected permuted matrix $\mathbf{F}_{\sigma^*} = \mathbb{E}[\mathbf{A}_{\sigma^*}]$ is assumed to be a bi-isotonic matrix, in the sense that $(\mathbf{F}_{\sigma^*})_{ij} \leq (\mathbf{F}_{\sigma^*})_{i+1j}$ and $(\mathbf{F}_{\sigma^*})_{ij} \leq (\mathbf{F}_{\sigma^*})_{ij+1}$. Besides, the symmetry assumption of $\mathbf{F}$ in seriation is replaced by $(\mathbf{F}_{\sigma^*})_{ij} = 1 - (\mathbf{F}_{\sigma^*})_{ji}$. This model introduced by Shah et al. [178] is referred as the SST model in the literature. Estimation of the latent order σ^* in the SST and related models has stimulated a lot of recent works [146, 154, 50], but the best possible performances achievable by polynomial-time methods remain unknown and it is not clear whether there exists or not a computational gap for this problem [146]. I am currently working on related questions with A. Carpentier and E. Pilliat.

Open Problem 5.4. *In the SST model, characterizing the best estimation error for σ^* which is achievable by polynomial-time procedures.*

As a side remark, let us mention that there exist several parametric counterparts to the ranking problems in the SST model, the most popular being the Bradley-Luce-Terry (BLT) model [32]. According to that model the observations $\mathbf{A}_{ij}$ are independent Bernoulli outcomes with mean $f(x_i^*, x_j^*) = \phi(x_i^* - x_j^*)$, where $x_i^* \in \mathbb{R}$ represents the skill of individual i and $\phi(x) = e^x/(1+e^x)$ is the sigmoid function. The estimation of the permutation σ^* or of the skills x_i^*'s can be efficiently performed using a spectral algorithm or two-steps variants of it [55, 162, 56]. In particular, these polynomial time procedures are shown to achieve the exact minimax rates and the problem does not exhibit any computational gap. However, the function f is known in BLT, which makes the problem significantly easier than SST model.

Bibliography

[2] Emmanuel Abbe. "Community detection and stochastic block models: recent developments". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6446–6531.

[3] Emmanuel Abbe and Colin Sandon. "Community Detection in General Stochastic Block Models: Fundamental Limits and Efficient Algorithms for Recovery". In: *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*. FOCS '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 670–688. ISBN: 978-1-4673-8191-8.

[4] David Amar, Ron Shamir, and Daniel Yekutieli. "Extracting replicable associations across multiple studies: Empirical Bayes algorithms for controlling the false discovery rate". In: *PLoS computational biology* 13.8 (2017), e1005700.

[5] A. A. Amini and E. Levina. "On semidefinite relaxations for the block model". In: *ArXiv e-prints* (June 2014).

[6] E. Arias-Castro, E. Candes, and Y. Plan. "Global Testing under Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism". In: *Annals of Statistics* 39 (2011), pp. 2533–2556.

[7] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[8] David Arthur and Sergei Vassilvitskii. "K-means++: The Advantages of Careful Seeding". In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '07. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035. ISBN: 978-0-898716-24-5.

[9] Jonathan E Atkins, Erik G Boman, and Bruce Hendrickson. "A spectral algorithm for seriation and the consecutive ones problem". In: *SIAM Journal on Computing* 28.1 (1998), pp. 297–310.

[10] Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. "The Hardness of Approximation of Euclidean k-Means". In: *31st International Symposium on Computational Geometry (SoCG 2015)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2015.

[11] Pranjal Awasthi and Aravindan Vijayaraghavan. "Clustering semi-random mixtures of gaussians". In: *International Conference on Machine Learning*. 2018, pp. 294–303.

[12] David Azriel and Armin Schwartzman. "The Empirical Distribution of a Large Number of Correlated Normal Variables". In: *Journal of the American Statistical Association* 110.511 (2015), pp. 1217–1228.

[13] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *The Annals of Probability* 33.5 (2005), pp. 1643–1697.

[14] Debapratim Banerjee and Zongming Ma. "Optimal hypothesis testing for stochastic block models with growing degrees". In: *arXiv preprint arXiv:1705.05305* (2017).

[15] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data". In: *Journal of Machine learning research* 9.Mar (2008), pp. 485–516.

[16] Yannick Baraud. "Non-asymptotic minimax rates of testing in signal detection". In: *Bernoulli* 8.5 (2002), pp. 577–606. ISSN: 1350-7265.

[17] Stephen T Barnard, Alex Pothen, and Horst Simon. "A spectral algorithm for envelope reduction of sparse matrices". In: *Numerical linear algebra with applications* 2.4 (1995), pp. 317–334.

[18] A. Belloni, V. Chernozhukov, and L. Wang. "Square-root Lasso: Pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4 (2011), pp. 791–806.

[19] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J. Roy. Statist. Soc. Ser. B* 57.1 (1995), pp. 289–300. ISSN: 0035-9246.

[20] Quentin Berthet and Philippe Rigollet. "Complexity Theoretic Lower Bounds for Sparse Principal Component Detection". In: *Proceedings of the 26th Annual Conference on Learning Theory*. Ed. by Shai Shalev-Shwartz and Ingo Steinwart. Vol. 30. Proceedings of Machine Learning Research. Princeton, NJ, USA: PMLR, Dec. 2013, pp. 1046–1066.

[21] Quentin Berthet, Philippe Rigollet, and Piyush Srivastava. "Exact recovery in the Ising blockmodel". In: *The Annals of Statistics* 47.4 (2019), pp. 1805–1834.

[22] P. Bickel, Y. Ritov, and A. Tsybakov. "Simultaneous analysis of lasso and Dantzig selector". In: *Ann. Statist.* 37.4 (2009), pp. 1705–1732. ISSN: 0090-5364.

[23] Peter J Bickel and Aiyou Chen. "A nonparametric view of network models and Newman–Girvan and other modularities". In: *Proceedings of the National Academy of Sciences* 106.50 (2009), pp. 21068–21073.

[24] Peter J Bickel, Aiyou Chen, and Elizaveta Levina. "The method of moments and degree distributions for network models". In: *The Annals of Statistics* 39.5 (2011), pp. 2280–2301.

[25] Peter J Bickel and Purnamrita Sarkar. "Hypothesis testing for automated community detection in networks". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.1 (2016), pp. 253–273.

[26] Lucien Birgé and Pascal Massart. "Estimation of integral functionals of a density". In: *The Annals of Statistics* 23.1 (1995), pp. 11–29.

[27] Avrim Blum and Joel Spencer. "Coloring random and semi-random k-colorable graphs". In: *Journal of Algorithms* 19.2 (1995), pp. 204–234.

[28] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. "Nonbacktracking spectrum of random graphs: Community detection and nonregular Ramanujan graphs". In: *Annals of Probability* 46.1 (2018), pp. 1–71.

[29] C. Borgs, J.T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. "Convergent sequences of dense graphs II. Multiway cuts and statistical physics". In: *Ann. of Math. (2)* 176.1 (2012), pp. 151–219. ISSN: 0003-486X.

[30] C. Borgs, J.T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. "Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing". In: *Adv. Math.* 219.6 (2008), pp. 1801–1851. ISSN: 0001-8708.

[31] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. "Independent filtering increases detection power for high-throughput experiments". In: *Proceedings of the National Academy of Sciences* 107.21 (2010), pp. 9546–9551.

[32] Ralph Allan Bradley and Milton E Terry. "Rank analysis of incomplete block designs: I. The method of paired comparisons". In: *Biometrika* 39.3/4 (1952), pp. 324–345. ISSN: 00063444.

[33] Matthew Brennan and Guy Bresler. "Average-case lower bounds for learning sparse mixtures, robust estimation and semirandom adversaries". In: *arXiv preprint arXiv:1908.06130* (2019).

[34] Matthew Brennan and Guy Bresler. "Optimal average-case reductions to sparse pca: From weak assumptions to strong hardness". In: *Conference on Learning Theory*. PMLR. 2019, pp. 469–470.

[35] Matthew Brennan, Guy Bresler, and Wasim Huleihel. "Reducibility and computational lower bounds for problems with planted sparse structure". In: *Conference On Learning Theory*. PMLR. 2018, pp. 48–166.

[36] Matthew Brennan, Guy Bresler, and Wasim Huleihel. "Universality of computational lower bounds for submatrix detection". In: *Conference on Learning Theory*. PMLR. 2019, pp. 417–468.

[37] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. "Testing for high-dimensional geometry in random graphs". In: *Random Structures & Algorithms* 49.3 (2016), pp. 503–532.

[38] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[39] Cristina Butucea and Yuri I Ingster. "Detection of a sparse submatrix of a high-dimensional noisy matrix". In: *Bernoulli* 19.5B (2013), pp. 2652–2688.

[40] Cristina Butucea, Enno Mammen, Mohamed Ndaoud, and Alexandre B Tsybakov. "Variable selection, monotone likelihood ratio and group sparsity". In: *arXiv preprint arXiv:2112.15042* (2021).

[41] Cristina Butucea, Mohamed Ndaoud, Natalia A Stepanova, and Alexandre B Tsybakov. "Variable selection with Hamming loss". In: *The Annals of Statistics* 46.5 (2018), pp. 1837–1875.

[42] T Tony Cai and Rong Ma. "Matrix Reordering for Noisy Disordered Matrices: Optimality and Computationally Efficient Algorithms". In: *arXiv preprint arXiv:2201.06438* (2022).

[43] T. Tony Cai and Zijian Guo. "Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity". In: *Ann. Statist.* 45.2 (2017), pp. 615–646. ISSN: 0090-5364.

[44] T. Tony Cai, X. Jessie Jeng, and Jiashun Jin. "Optimal detection of heterogeneous and heteroscedastic mixtures". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73.5 (2011), pp. 629–662. ISSN: 1369-7412.

[45] T. Tony Cai and Mark G. Low. "Nonquadratic estimators of a quadratic functional". In: *Ann. Statist.* 33.6 (2005), pp. 2930–2956. ISSN: 0090-5364.

[46] T. Tony Cai and Mark G. Low. "Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional". In: *Ann. Statist.* 39.2 (2011), pp. 1012–1041. ISSN: 0090-5364.

[47] Emmanuel Candes and Terence Tao. "The Dantzig selector: statistical estimation when p is much larger than n". In: *Ann. Statist.* 35.6 (2007), pp. 2313–2351. ISSN: 0090-5364.

[48] Hock-Peng Chan and Hao Chen. "Multi-sequence segmentation via score and higher-criticism tests". In: *arXiv preprint arXiv:1706.07586* (2017).

[49] Moses Charikar, Sudipto Guha, éva Tardos, and David B. Shmoys. "A Constant-Factor Approximation Algorithm for the k-Median Problem". In: *Journal of Computer and System Sciences* 65.1 (2002), pp. 129–149. ISSN: 0022-0000.

[50] Sabyasachi Chatterjee and Sumit Mukherjee. "Estimation in tournaments and graphs under monotonicity constraints". In: *IEEE Transactions on Information Theory* 65.6 (2019), pp. 3525–3539.

[51] Sourav Chatterjee. "Matrix estimation by universal singular value thresholding". In: *The Annals of Statistics* 43.1 (2014), pp. 177–214.

[52] Kehui Chen and Jing Lei. "Network cross-validation for determining the number of communities in network data". In: *Journal of the American Statistical Association* 113.521 (2018), pp. 241–251.

[53] Xin Chen and Anderson Y Zhang. "Optimal clustering in anisotropic gaussian mixture models". In: *arXiv preprint arXiv:2101.05402* (2021).

[54] Yudong Chen and Jiaming Xu. "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 882–938.

[55] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. "Spectral method and regularized MLE are both optimal for top-K ranking". In: *The Annals of Statistics* 47.4 (2019), pp. 2204–2235.

[56] Yuxin Chen and Changho Suh. "Spectral MLE: Top-K Rank Aggregation from Pairwise Comparisons". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 371–380.

[57] Peter Chin, Anup Rao, and Van Vu. "Stochastic Block Model and Community Detection in Sparse Graphs: A spectral algorithm with optimal rate of recovery". In: *Proceedings of The 28th Conference on Learning Theory*. Ed. by Peter Grünwald, Elad Hazan, and Satyen Kale. Vol. 40. Proceedings of Machine Learning Research. Paris, France: PMLR, Mar. 2015, pp. 391–423.

[58] Yunjin Choi, Jonathan Taylor, and Robert Tibshirani. "Selecting the number of principal components: Estimation of the true rank of a noisy matrix". In: *The Annals of Statistics* (2017), pp. 2590–2617.

[59] Lynna Chu and Hao Chen. "Asymptotic distribution-free change-point detection for multivariate and non-euclidean data". In: *The Annals of Statistics* 47.1 (2019), pp. 382–414.

[60] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer. "Graph Laplacian Tomography From Unknown Random Projections". In: *IEEE Transactions on Image Processing* 17.10 (2008), pp. 1891–1899.

[61] Olivier Collier, Laëtitia Comminges, and Alexandre B Tsybakov. "Minimax estimation of linear and quadratic functionals on sparsity classes". In: *arXiv preprint arXiv:1502.00665* (2015).

[62] ENCODE Project Consortium et al. "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project". In: *Nature* 447.7146 (2007), p. 799.

[63] R Cameron Craddock, G Andrew James, Paul E Holtzheimer, Xiaoping P Hu, and Helen S Mayberg. "A whole brain fMRI atlas generated via spatially constrained spectral clustering". In: *Human brain mapping* 33.8 (2012), pp. 1914–1928.

[64] Damek Davis, Mateo Diaz, and Kaizheng Wang. "Clustering a mixture of gaussians with unknown covariance". In: *arXiv preprint arXiv:2110.01602* (2021).

[65] Yohann De Castro, Claire Lacour, and Thanh Mai Pham Ngoc. "Adaptive estimation of nonparametric geometric graphs". In: *Mathematical Statistics and Learning* 2.3 (2020), pp. 217–274.

[66] Persi Diaconis and Svante Janson. "Graph limits and exchangeable random graphs". In: *Rend. Mat. Appl. (7)* 28.1 (2008), pp. 33–61. ISSN: 1120-7183.

[67] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. "List-decodable robust mean estimation and learning mixtures of spherical gaussians". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 1047–1060.

[68] Josep Diaz, Colin McDiarmid, and Dieter Mitsche. "Learning random points from geometric graphs or orderings". In: *Random Structures & Algorithms* 57.2 (2020), pp. 339–370.

[69] Lee H. Dicker. "Variance estimation in high-dimensional linear models". In: *Biometrika* 101.2 (2014), pp. 269–284. ISSN: 0006-3444.

[70] Thorsten Dickhaus. *Simultaneous statistical inference*. With applications in the life sciences. Springer, Heidelberg, 2014, pp. xiv+180. ISBN: 978-3-642-45181-2; 978-3-642-45182-9.

[71] David L. Donoho and Michael Nussbaum. "Minimax quadratic estimation of a quadratic functional". In: *J. Complexity* 6.3 (1990), pp. 290–323. ISSN: 0885-064X.

[72] David Donoho and Jiashun Jin. "Higher criticism for detecting sparse heterogeneous mixtures". In: *Ann. Statist.* 32.3 (2004), pp. 962–994. ISSN: 0090-5364.

[73] Natalie Doss, Yihong Wu, Pengkun Yang, and Harrison H Zhou. "Optimal estimation of high-dimensional location Gaussian mixtures". In: *arXiv preprint arXiv:2002.05818* (2020).

[74] Bradley Efron. "Correlation and large-scale simultaneous significance testing". In: *J. Amer. Statist. Assoc.* 102.477 (2007), pp. 93–103. ISSN: 0162-1459.

[75] Bradley Efron. "Doing thousands of hypothesis tests at the same time". In: *Metron - International Journal of Statistics* LXV.1 (2007), pp. 3–21.

[76] Bradley Efron. "Empirical Bayes estimates for large-scale prediction problems." English. In: *J. Am. Stat. Assoc.* 104.487 (2009), pp. 1015–1028. ISSN: 0162-1459; 1537-274X/e.

[77] Bradley Efron. "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis." English. In: *J. Am. Stat. Assoc.* 99.465 (2004), pp. 96–104. ISSN: 0162-1459; 1537-274X/e.

[78] Bradley Efron. "Microarrays, empirical Bayes and the two-groups model". In: *Statist. Sci.* 23.1 (2008), pp. 1–22. ISSN: 0883-4237.

[79] Bradley Efron, Robert Tibshirani, John D. Storey, and Virginia Tusher. "Empirical Bayes analysis of a microarray experiment". In: *J. Amer. Statist. Assoc.* 96.456 (2001), pp. 1151–1160. ISSN: 0162-1459.

[80] Jeremy Elson, Richard M. Karp, Christos H. Papadimitriou, and Scott Shenker. "Global Synchronization in Sensornets". In: *LATIN 2004: Theoretical Informatics*. Ed. by Martín Farach-Colton. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 609–624. ISBN: 978-3-540-24698-5.

[81] Zhou Fan and Leying Guan. "Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs". In: *The Annals of Statistics* 46.6B (2018), pp. 3217–3245.

[82] Yingjie Fei and Yudong Chen. "Exponential error rates of SDP for block models: Beyond Grothendieck's inequality". In: *IEEE Transactions on Information Theory* 65.1 (2018), pp. 551–571.

[83] Yingjie Fei and Yudong Chen. "Hidden integrality of SDP relaxations for sub-Gaussian mixture models". In: *Conference On Learning Theory*. PMLR. 2018, pp. 1931–1965.

[84] Fajwel Fogel, Rodolphe Jenatton, Francis Bach, and Alexandre d'Aspremont. "Convex relaxations for permutation problems". In: *Advances in Neural Information Processing Systems*. 2013, pp. 1016–1024.

[85] Klaus Frick, Axel Munk, and Hannes Sieling. "Multiscale change point inference". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.3 (2014), pp. 495–580.

[86] Alan Frieze and Ravi Kannan. "Quick approximation to matrices and applications". In: *Combinatorica* 19.2 (1999), pp. 175–220. ISSN: 0209-9683.

[87] Piotr Fryzlewicz. "Tail-greedy bottom-up data decompositions and fast multiple change-point detection". In: *The Annals of Statistics* 46.6B (2018), pp. 3390–3421.

[88] Delbert Fulkerson and Oliver Gross. "Incidence matrices and interval graphs". In: *Pacific journal of mathematics* 15.3 (1965), pp. 835–855.

[89] Chao Gao, Yu Lu, Zongming Ma, and Harrison H Zhou. "Optimal estimation and completion of matrices with biclustering structures". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 5602–5630.

[90] Chao Gao, Yu Lu, and Harrison H Zhou. "Rate-optimal graphon estimation". In: *The Annals of Statistics* 43.6 (2015), pp. 2624–2652.

[91] Chao Gao and Zongming Ma. "Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing". In: *Statistical Science* 36.1 (2021), pp. 16–33.

[92] Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. "Achieving Optimal Misclassification Proportion in Stochastic Block Models". In: *J. Mach. Learn. Res.* 18.1 (Jan. 2017), pp. 1980–2024. ISSN: 1532-4435.

[93] Chao Gao and Anderson Y Zhang. "Iterative algorithm for discrete structure recovery". In: *arXiv preprint arXiv:1911.01018* (2019).

[94] Gemma C Garriga, Esa Junttila, and Heikki Mannila. "Banded structure in binary matrices". In: *Knowledge and information systems* 28.1 (2011), pp. 197–226.

[95] Matan Gavish and David L Donoho. "The optimal hard threshold for singular values is $4/\sqrt{3}$". In: *IEEE Transactions on Information Theory* 60.8 (2014), pp. 5040–5053.

[96] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *Ann. Statist.* 42.3 (2014), pp. 1166–1202. ISSN: 0090-5364.

[97] E. N. Gilbert. "Random Plane Networks". In: *Journal of the Society for Industrial and Applied Mathematics* 9.4 (1961), pp. 533–543.

[98] Christophe Giraud. *Introduction to high-dimensional statistics*. Vol. 139. Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL, 2015, pp. xvi+255. ISBN: 978-1-4822-3794-8.

[99] A. Giridhar and P. R. Kumar. "Distributed Clock Synchronization over Wireless Networks: Algorithms and Analysis". In: *Proceedings of the 45th IEEE Conference on Decision and Control*. 2006, pp. 4915–4920.

[100] M.F. Glasser et al. "A Multi-modal parcelation of human cerebral cortex". In: *Nature* 536 (2016), pp. 171–178.

[101] Michel X Goemans and David P Williamson. "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming". In: *Journal of the ACM (JACM)* 42.6 (1995), pp. 1115–1145.

[102] D. Goldstein. "Common genetic variation and human traits". In: *New England Journal of Medicine* 360 (2009), pp. 1696–1698.

[103] T. R. Golub et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring". In: *Science* 286.5439 (1999), pp. 531–537. ISSN: 0036-8075.

[104] Bruce Hajek, Yihong Wu, and Jiaming Xu. "Computational lower bounds for community detection on random graphs". In: *Conference on Learning Theory*. PMLR. 2015, pp. 899–928.

[105] Yanjun Han, Jiantao Jiao, and Rajarshi Mukherjee. "On estimation of L_r-norms in Gaussian white noise models". In: *Probability Theory and Related Fields* 177.3 (2020), pp. 1243–1294.

[106] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. "Minimax estimation of divergences between discrete distributions". In: *IEEE Journal on Selected Areas in Information Theory* 1.3 (2020), pp. 814–823.

[107] Ingrid Hedenfalk et al. "Gene-Expression Profiles in Hereditary Breast Cancer". In: *New England Journal of Medicine* 344.8 (2001), pp. 539–548.

[108] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. "Latent space approaches to social network analysis". In: *Journal of the american Statistical association* 97.460 (2002), pp. 1090–1098.

[109] Marc Hoffmann and Richard Nickl. "On adaptive inference and confidence bands". In: *Ann. Statist.* 39.5 (2011), pp. 2383–2409. ISSN: 0090-5364.

[110] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps". In: *Social networks* 5.2 (1983), pp. 109–137.

[111] Samuel B Hopkins and Jerry Li. "Mixture models, robustness, and sum of squares proofs". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 1021–1034.

[112] Peter J Huber. *Robust statistics*. Vol. 523. John Wiley & Sons, 2004.

[113] Il'dar Abdulovich Ibragimov and Rafail Zalmanovich Has' Minskii. *Statistical estimation: asymptotic theory*. Vol. 16. Springer Science & Business Media, 2013.

[114] Yuri I. Ingster. "Asymptotically minimax hypothesis testing for nonparametric alternatives. I". In: *Math. Methods Statist.* 2.2 (1993), pp. 85–114. ISSN: 1066-5307.

[115] Yuri I. Ingster. "Asymptotically minimax hypothesis testing for nonparametric alternatives. II". In: *Math. Methods Statist.* 2.3 (1993), pp. 171–189. ISSN: 1066-5307.

[116] Yuri I. Ingster. "Asymptotically minimax hypothesis testing for nonparametric alternatives. III". In: *Math. Methods Statist.* 2.4 (1993), pp. 249–268. ISSN: 1066-5307.

[117] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. Springer Science & Business Media, 2012.

[118] Lucas Janson, Rina Foygel Barber, and Emmanuel Candes. "EigenPrism: inference for high dimensional signal-to-noise ratios". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4 (2017), pp. 1037–1065.

[119] Jeannette Janssen and Aaron Smith. "Reconstruction of line-embeddings of graphons". In: *Electronic Journal of Statistics* 16.1 (2022), pp. 331–407.

[120] Adel Javanmard and Andrea Montanari. "Confidence intervals and hypothesis testing for high-dimensional regression". In: *J. Mach. Learn. Res.* 15 (2014), pp. 2869–2909. ISSN: 1532-4435.

[121] Wei Jiang and Weichuan Yu. "Controlling the joint local false discovery rate is more powerful than meta-analysis methods in joint analysis of summary statistics from multiple genome-wide association studies". In: *Bioinformatics* 33.4 (Dec. 2016), pp. 500–507. ISSN: 1367-4803.

[122] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. "Minimax estimation of functionals of discrete distributions". In: *IEEE Transactions on Information Theory* 61.5 (2015), pp. 2835–2885.

[123] Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. "Optimal adaptivity of signed-polygon statistics for network testing". In: *The Annals of Statistics* 49.6 (2021), pp. 3408–3433.

[124] Moritz Jirak. "Uniform change point tests in high dimension". In: *The Annals of Statistics* 43.6 (2015), pp. 2451–2483.

[125] Julie Josse and François Husson. "Selecting the number of components in principal component analysis using cross-validation approximations". In: *Computational Statistics & Data Analysis* 56.6 (2012), pp. 1869–1879.

[126] Lou Jost. "Entropy and diversity". In: *Oikos* 113.2 (2006), pp. 363–375.

[127] Laura Jula Vanegas, Merle Behr, and Axel Munk. "Multiscale quantile segmentation". In: *Journal of the American Statistical Association* (2021), pp. 1–14.

[128] Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. "Optimal detection of changepoints with a linear computational cost". In: *J. Amer. Statist. Assoc.* 107.500 (2012), pp. 1590–1598. ISSN: 0162-1459.

[129] Vladimir Koltchinskii. "Asymptotically efficient estimation of smooth functionals of covariance operators". In: *Journal of the European Mathematical Society* 23.3 (2020), pp. 765–843.

[130] Vladimir Koltchinskii. "Estimation of smooth functionals in high-dimensional models: bootstrap chains and Gaussian approximation". In: *arXiv preprint arXiv:2011.03789* (2020).

[131] Vladimir Koltchinskii and Karim Lounici. "Concentration inequalities and moment bounds for sample covariance operators". In: *Bernoulli* 23.1 (2017), pp. 110–133.

[132] Vladimir Koltchinskii and Mayya Zhilova. "Efficient estimation of smooth functionals in Gaussian shift models". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 57.1 (2021), pp. 351–386.

[133] Vladimir Koltchinskii and Mayya Zhilova. "Estimation of smooth functionals in normal models: bias reduction and asymptotic efficiency". In: *The Annals of Statistics* 49.5 (2021), pp. 2577–2610.

[134] Ru Kong, Jingwei Li, Csaba Orban, Mert R Sabuncu, Hesheng Liu, Alexander Schaefer, Nanbo Sun, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, et al. "Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion". In: *Cerebral cortex* 29.6 (2019), pp. 2533–2551.

[135] Weihao Kong and Gregory Valiant. "Estimating learnability in the sublinear data regime". In: *Advances in Neural Information Processing Systems* 31 (2018).

[136] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. "Robust moment estimation and improved clustering via sum of squares". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 1035–1046.

[137] Solt Kovàcs, Housen Li, Peter Bühlmann, and Axel Munk. "Seeded Binary Segmentation: A general methodology for fast and optimal change point detection". In: *arXiv preprint arXiv:2002.06633* (2020).

[138] Jeongyeol Kwon and Constantine Caramanis. "The EM algorithm gives sample-optimality for learning mixtures of well-separated gaussians". In: *Conference on Learning Theory*. PMLR. 2020, pp. 2425–2487.

[139] B. Laurent and P. Massart. "Adaptive estimation of a quadratic functional by model selection". In: *Annals of Statistics* 28.5 (2000), pp. 1302–1338. ISSN: 0090-5364.

[140] Namgil Lee, Ah-Young Kim, Chang-Hyun Park, and Sung-Ho Kim. "An improvement on local FDR analysis applied to functional MRI data". In: *Journal of neuroscience methods* 267 (2016), pp. 115–125.

[141] Jing Lei. "A goodness-of-fit test for stochastic block models". In: *The Annals of Statistics* 44.1 (2016), pp. 401–424.

[142] Jing Lei and Alessandro Rinaldo. "Consistency of spectral clustering in stochastic block models". In: *Ann. Statist.* 43.1 (2015), pp. 215–237. ISSN: 0090-5364.

[143] Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. "On estimation of the L r norm of a regression function". In: *Probability theory and related fields* 113.2 (1999), pp. 221–253.

[144] Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborová. "Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering". In: *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2016, pp. 601–608.

[145] Boris Ya Levit. "Asymptotically efficient estimation of nonlinear functionals". In: *Problemy Peredachi Informatsii* 14.3 (1978), pp. 65–72.

[146] Allen Liu and Ankur Moitra. "Better algorithms for estimating non-parametric models in crowd-sourcing and rank aggregation". In: *Conference on Learning Theory*. PMLR. 2020, pp. 2780–2829.

[147] S. Lloyd. "Least Squares Quantization in PCM". In: *IEEE Trans. Inf. Theor.* 28.2 (Sept. 1982), pp. 129–137. ISSN: 0018-9448.

[148] Matthias Löffler, Alexander S Wein, and Afonso S Bandeira. "Computationally efficient sparse clustering". In: *arXiv preprint arXiv:2005.10817* (2020).

[149] Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou. "Optimality of spectral clustering in the gaussian mixture model". In: *The Annals of Statistics* 49.5 (2021), pp. 2506–2530.

[150] László Lovász. *Large networks and graph limits*. Vol. 60. American Mathematical Society Colloquium Publications. American Mathematical Soc., 2012, pp. xiv+475. ISBN: 978-0-8218-9085-1.

[151] László Lovász and Balázs Szegedy. "Limits of dense graph sequences". In: *Journal of Combinatorial Theory, Series B* 96.6 (2006), pp. 933–957.

[152] Y. Lu and H. H. Zhou. "Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants". In: *ArXiv e-prints* (Dec. 2016).

[153] Brendan Maher. "Personal genomes: The case of the missing heritability". In: *Nature* 456.7218 (Nov. 2008), pp. 18–21. ISSN: 0028-0836.

[154] Cheng Mao, Ashwin Pananjady, and Martin J Wainwright. "Towards optimal estimation of bivariate isotonic matrices with unknown permutations". In: *The Annals of Statistics* 48.6 (2020), pp. 3183–3205.

[155] Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher. "Uncovering latent structure in valued graphs: a variational approach". In: *The Annals of Applied Statistics* 4.2 (2010), pp. 715–742.

[156] N. Meinshausen and P. Bühlmann. "High-dimensional graphs and variable selection with the Lasso". In: *Annals of Statistics* 34.3 (2006), pp. 1436–1462. ISSN: 0090-5364.

[157] Christopher J. Miller, Christopher Genovese, Robert C. Nichol, Larry Wasserman, Andrew Connolly, Daniel Reichart, Andrew Hopkins, Jeff Schneider, and Andrew Moore. "Controlling the False-Discovery Rate in Astrophysical Data Analysis". In: *The Astronomical Journal* 122.6 (2001), pp. 3492–3505.

[158] Dustin G Mixon, Soledad Villar, and Rachel Ward. "Clustering subgaussian mixtures by semidefinite programming". In: *Information and Inference: A Journal of the IMA* 6.4 (2017), pp. 389–415.

[159] Ankur Moitra, William Perry, and Alexander S Wein. "How robust are reconstruction thresholds for community detection?" In: *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM. 2016, pp. 828–841.

[160] Rajarshi Mukherjee and Subhabrata Sen. "On minimax exponents of sparse testing". In: *arXiv preprint arXiv:2003.00570* (2020).

[161] Mohamed Ndaoud. "Sharp optimal recovery in the Two Component Gaussian Mixture Model". In: *arXiv e-prints*, arXiv:1812.08078 (Dec. 2018), arXiv:1812.08078.

[162] Sahand Negahban, Sewoong Oh, and Devavrat Shah. "Rank Centrality: Ranking from Pairwise Comparisons". In: *Operations Research* 65.1 (2017), pp. 266–287.

[163] Mark Newman. *Networks*. Oxford university press, 2018.

[164] Richard Nickl and Sara van de Geer. "Confidence sets in sparse regression". In: *Ann. Statist.* 41.6 (2013), pp. 2852–2876. ISSN: 0090-5364.

[165] Yue S Niu, Ning Hao, and Heping Zhang. "Multiple change-point detection: A selective overview". In: *Statistical Science* 31.4 (2016), pp. 611–623.

[166] Adam B. Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. "Circular binary segmentation for the analysis of array-based DNA copy number data". In: *Biostatistics* 5.4 (2004), pp. 557–572.

[167] Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. "Optimal nonparametric change point analysis". In: *Electronic Journal of Statistics* 15.1 (2021), pp. 1154–1201.

[168] Jiming Peng and Yu Wei. "Approximating K-means-type Clustering via Semidefinite Programming". In: *SIAM J. on Optimization* 18.1 (Feb. 2007), pp. 186–205. ISSN: 1052-6234.

[169] Mathew Penrose. *Random geometric graphs*. Vol. 5. Oxford Studies in Probability. Oxford university press, 2003, pp. xiv+330. ISBN: 0-19-850626-0.

[170] Maxim Rabinovich, Aaditya Ramdas, Michael I Jordan, and Martin J Wainwright. "Optimal rates and trade-offs in multiple testing". In: *Statistica Sinica* 30.2 (2020), pp. 741–762.

[171] Antoine Recanati, Thomas Brüls, and Alexandre d'Aspremont. "A spectral algorithm for fast de novo layout of uncorrected long nanopore reads". In: *Bioinformatics* 33.20 (June 2017), pp. 3188–3194. ISSN: 1367-4803.

[172] Antoine Recanati, Thomas Kerdreux, and Alexandre d'Aspremont. "Reconstructing Latent Orderings by Spectral Clustering". In: *arXiv preprint arXiv:1807.07122* (2018).

[173] Oded Regev and Aravindan Vijayaraghavan. "On learning mixtures of well-separated gaussians". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. Oct. 2017, pp. 85–96.

[174] W. S. Robinson. "A Method for Chronologically Ordering Archaeological Deposits". In: *American Antiquity* 16.4 (1951), pp. 293–301. ISSN: 00027316.

[175] Armin Schwartzman. "Comment: "Correlated z-values and the accuracy of large-scale statistical estimates" [MR2752597]". In: *J. Amer. Statist. Assoc.* 105.491 (2010), pp. 1059–1063. ISSN: 0162-1459.

[176] Armin Schwartzman. "Empirical null and false discovery rate inference for exponential families". In: *The Annals of Applied Statistics* 2.4 (2008), pp. 1332–1359.

[177] Nimrod Segol and Boaz Nadler. "Improved convergence guarantees for learning Gaussian mixture models by EM and gradient EM". In: *Electronic Journal of Statistics* 15.2 (2021), pp. 4510–4544.

[178] Nihar B Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J Wainwright. "Stochastically Transitive Models for Pairwise Comparisons: Statistical and Computational Issues". In: *IEEE Transactions on Information Theory* 63.2 (2016), pp. 934–959.

[179] A. Singer. "Angular synchronization by eigenvectors and semidefinite programming". In: *Applied and Computational Harmonic Analysis* 30.1 (2011), pp. 20–36. ISSN: 1063-5203.

[180] J.L. Stein et al. "Identification of common variants associated with human hippocampal and intracranial volumes." In: *Nature Genetics* 44(5) (2012), pp. 552–61.

[181] Matthew Stephens. "False discovery rates: a new deal". In: *Biostatistics* 18.2 (2017), pp. 275–294.

[182] Sophia Sulis, David Mary, and Lionel Bigot. "A study of periodograms standardized using training datasets and application to exoplanet detection". In: *IEEE Transactions on Signal Processing* 65.8 (2017), pp. 2136–2150.

[183] Lei Sun and Matthew Stephens. "Solving the empirical Bayes normal means problem with correlated noise". In: *arXiv preprint arXiv:1812.07488* (2018).

[184] A. S. Szalay, A. J. Connolly, and G. P. Szokoly. "Simultaneous Multicolor Detection of Faint Galaxies in the Hubble Deep Field". In: *The Astronomical Journal* 117 (Jan. 1999), pp. 68–74.

[185] Endre Szemerédi. *Regular partitions of graphs*. Tech. rep. DTIC Document, 1975.

[186] Roberto Toro et al. "Genomic architecture of human neuroanatomical diversity". In: *Molecular Psychiatry* (2014).

[187] Charles Truong, Laurent Oudre, and Nicolas Vayatis. "Selective review of offline change point detection methods". In: *Signal Processing* 167 (2020), p. 107299.

[188] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. Springer, New York, 2009, pp. xii+214. ISBN: 978-0-387-79051-0.

[189] Santosh Vempala and Grant Wang. "A spectral algorithm for learning mixture models". In: *Journal of Computer and System Sciences* 68.4 (2004). Special Issue on FOCS 2002, pp. 841–860. ISSN: 0022-0000.

[190] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[191] Abraham Wald. "Sequential tests of statistical hypotheses". In: *The annals of mathematical statistics* 16.2 (1945), pp. 117–186.

[192] Daren Wang, Yi Yu, and Alessandro Rinaldo. "Optimal change point detection and localization in sparse dynamic networks". In: *The Annals of Statistics* 49.1 (2021), pp. 203–232.

[193] Daren Wang, Yi Yu, and Alessandro Rinaldo. "Optimal covariance change point localization in high dimensions". In: *Bernoulli* 27.1 (2021), pp. 554–575.

[194] Daren Wang, Yi Yu, and Alessandro Rinaldo. "Univariate mean change point detection: Penalization, CUSUM and optimality". In: *Electron. J. Stat.* 14.1 (2020), pp. 1917–1961.

[195] Tengyao Wang and Richard J. Samworth. "High dimensional change point estimation via sparse projection". In: *J R Stat Soc Ser B Stat Methodol* 80.1 (2018), pp. 57–83. ISSN: 1369–7412.

[196] Patrick J. Wolfe and Sofia C. Olhede. "Nonparametric graphon estimation". In: *arXiv preprint arXiv:1309.5936* (2013).

[197] Angélique B van't Wout, Ginger K Lehrman, Svetlana A Mikheeva, Gemma C O'Keeffe, Michael G Katze, Roger E Bumgarner, Gary K Geiss, and James I Mullins. "Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4+-T-cell lines". In: *Journal of virology* 77.2 (2003), pp. 1392–1402.

[198] Yihong Wu. "Lecture notes on information-theoretic methods for high-dimensional statistics". In: *Lecture Notes for ECE598YW (UIUC)* 16 (2017).

[199] Yihong Wu and Pengkun Yang. "Chebyshev polynomials, moment matching, and optimal estimation of the unseen". In: *The Annals of Statistics* 47.2 (2019), pp. 857–883.

[200] Yihong Wu and Pengkun Yang. "Minimax rates of entropy estimation on large alphabets via best polynomial approximation". In: *IEEE Transactions on Information Theory* 62.6 (2016), pp. 3702–3720.

[201] Jiaming Xu, Laurent Massoulié, and Marc Lelarge. "Edge Label Inference in Generalized Stochastic Block Models: from Spectral Theory to Impossibility Results". In: *Conference on Learning Theory*. 2014, pp. 903–920.

[202] B.T. Yeo et al. "The organization of the human cerebral cortex estimated by intrinsic functional connectivity". In: *Journal of Neurophysiology* 106 (2011), pp. 1125–1165.

[203] Yi Yu, Oscar Hernan Madrid Padilla, and Alessandro Rinaldo. "Optimal partition recovery in general graphs". In: *arXiv preprint arXiv:2110.10989* (2021).

[204] Se-Young Yun and Alexandre Proutière. "Accurate Community Detection in the Stochastic Block Model via Spectral Algorithms". In: *CoRR* abs/1412.7335 (2014).

[205] Rong W Zablocki, Andrew J Schork, Richard A Levine, Ole A Andreassen, Anders M Dale, and Wesley K Thompson. "Covariate-modulated local false discovery rate for genome-wide association studies". In: *Bioinformatics* 30.15 (2014), pp. 2098–2104.

[206] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. "Consistency of community detection in networks under degree-corrected stochastic block models". In: *The Annals of Statistics* 40.4 (2012), pp. 2266–2292. ISSN: 0090-5364.

Chapter 6

Résumé en Français

Ce chapitre propose une version résumée en langue française du manuscrit d'habilitation.

6.1 Parcours scientifique

Ma thèse de doctorat [T1] portait sur l'inférence de graphes dans les modèles graphiques gaussiens. Puisque l'estimation de ces graphes peut se reformuler en termes de sélection de variables dans des modèles de régression linéaire [156], je me suis rapidement intéressé à l'analyse minimax des modèles de régression linéaire parcimonieux en grande dimension [A25].

Après l'achèvement de mon doctorat en 2008, j'ai commencé à travailler sur divers problèmes d'estimation minimax en statistique en grande dimension. J'ai été notamment intéressé par la compréhension des régimes de très haute dimension [A26] pour lesquels les vitesses optimales d'estimation et de test changent de forme.

À la même époque, j'ai été recruté comme chercheur en statistique à MISTEA au sein de l'INRAE. Dans ce contexte, j'ai commencé à discuter et à collaborer avec des biologiste, agronomes et collègues statisticiens au sein de mon institut. Cela m'a conduit à plusieurs travaux appliqués [A22] ainsi qu'à quelques collaborations en analyse de données fonctionnelles [A23, A27]. Celles-ci étaient motivés par l'analyse de phénotypes complexes (ex: courbe de croissance de plantes) qui est une thématique prioritaire pour l'INRAE. Depuis cette période, j'ai gardé un goût prononcé pour les modèles statistiques suffisamment simples pour être analysables mais assez complexes pour capturer les grandes lignes des problèmatiques concrêtes. A titre d'exemple, on peut citer l'estimation de l'héritabilité en génétique grâce à l'estimation du niveau de signal dans des modèles de régression linéaire en grande dimension [A14] – voir la section 2.3.

Mes intérêts de recherche récents ont été fortement remodelés par deux rencontres. Premièrement, Ery Arias-Castro m'a invité à collaborer avec lui sur des problèmes de détection de communautés [A21, A20]. Il m'a également introduit aux domaines de l'analyse de réseau et du clustering qui sont tous les deux centraux dans cette thèse. Deuxièmement, j'ai été invité à participer au consortium MIRES. Ce groupe interdisciplinaire d'anthropologues, de généticiens, d'écologues et de statisticiens cherche à fournir des méthodes pour l'analyse des réseaux d'échange de semences. Bien que j'aie peu de pub-

lications sur ce sujet (mais voir [A19]), j'y ai consacré une part importante de mon activité scientifique par le biais de discussions informelles avec des chercheur·e·s et étudiant·e·s, la réalisation de tutoriaux,... Certaines des questions statistiques relatives à cette thématique concernent l'analyse de réseaux et plus généralement l'apprentissage non supervisé et rejoignent ainsi mes travaux plus théoriques. Comme cette thèse décrit principalement mes thavaux mathématiques, je ne détaille pas plus cette activité.

Au cours de la période 2014–2018, j'ai eu la chance d'avoir des collaborations fructueuses et enrichissantes sur trois directions de recherche qui correspondent aux chapitres 2–4: Alexandra Carpentier m'a initié au problème de l'estimation de la complexité, ce qui a conduit à notre travail commun sur les tests de sparsité [A8, A5]. Avec Olga Klopp, nous avons par ailleur abordé le problème de l'estimation des graphons parcimonieux [A16, A9]. Avec Christophe Giraud, nous avons enfin fourni une analyse de la relaxation convexe du critére K-means [A13], ce tant pour du clustering de points que pour du clustering de graphe.

Ces dernières années, j'ai eu le plaisir de participer à l'encadrement de trois doctorants: Solène Thépaut, Yann Issartel (tous deux conjointement avec C. Giraud), et Emmanuel Pilliat (conjointement avec A. Carpentier et J. Salmon). Parallèlement, mes intérêts scientifiques ont évolué vers d'autres problèmes non supervisés tels que la détection de ruptures [P3, P4] ou les problèmes de sériation/classement qui sont décrits dans le chapitre 5.

Quelques mots sur ma démarche mathématique. La plupart de mes travaux mathématiques s'inscrivent fortement dans la théorie minimax. Depuis mon doctorat, ma démarche scientifique vise, pour un problèmes donné, à identifier la notion de signal et à établir des majorations et minorations précises du risque minimax associé en interrogeant notamment le rôle joué par la connaissance de certains paramêtres de nuisance (ex: le niveau de bruit, la distribution du bruit). Au-delà de mes penchants mathématiques, je suis convaincu que ce type d'analyse permet de se forger une intuition sur les quantités pertinentes et les hypothèses importantes pour la résolution de problèmes pratiques.

Organisation. Ce résumé est organisé en quatre sections qui peuvent être lues de manière presque indépendante. La première est dédiée aux problèmes de détection et d'estimation de fonctionnelles principalement dans les modèle de séquences gaussiennes et de régression linéaire. La section suivante est consacrée à l'analyse de réseaux, notamment la détection des communautés et l'estimation des graphons. Les deux dernières sections portent sur le clustering et la détection de ruptures. J'en profite pour glisser quelques éléments de mon projet de recherche.

6.2 Détection et estimation de fonctionnelles

L'estimation dans les modèles de régression linéaire en grande dimension a suscité beaucoup d'intérêt au cours des vingt dernières années [38, 98, 190]. Elle a donné lieu à des contributions fondamentales telles que la théorie du compressed sensing. Plus généralement, les idées développées se sont répandues bien au-delà de ce modèle spécifique et ont eu un impact profond en statistique et en apprentissage automatique.

Les aspects essentiels de l'estimation du paramètre de régression θ^* dans le modèle de régression linéaire parcimonieux sont compris depuis une quinzaine d'années [22, 47]. Cette section est principalement consacrée à des problèmes connexes pour lesquels on désire soit tester si $\theta^* = 0$ (détection de signal), tester si θ^* appartient à une classe spécifique, ou plus généralement estimer une fonction simple $f(\theta^*)$ de θ^* (estimation de fonctionnelle). Parmi les fonctionnelles classiques, on peut à penser à une coordonnée spécifique θ_i^* de θ^* [43, 120], ce afin d'estimer l'effet d'une covariable tout en prenant en compte les autres covariables, ou alors à la norme l_q de θ^* [105, 143], ou bien encore au rapport signal/bruit [69, 118, A14].

Dans cette section, j'axe principalement la discussion sur les deux modèles jouets suivants: le *modèle de séquence gaussienne* $Y = \theta^* + \epsilon$ et le *modèle de régression linéaire gaussien* $Y = \mathbf{X}\theta^* + \epsilon$. Dans la suite, n désigne la taille de l'échantillon et p le nombre de covariables (pour la régresson linéaire).

Les contributions décrites dans cette section suivent l'organisation générale suivante: (i) pour le problème de test et d'estimation considéré, déterminer les vitesses minimax notamment en fonction de la parcimonie de θ^*; (ii) Si possible, proposer une procédure en temps polynomial permettant d'atteindre la vitesse optimale; (iii) Si possible également, proposer une procédure qui s'adapte à la parcimonie inconnue, c'est-à-dire atteigne la vitesse optimale (qui dépend du nombre $\|\theta^*\|_0$ de composantes non nulles) sans connaissance préalable de $\|\theta^*\|_0$; (iv) étudier le rôle joué par la connaissance ou non de paramètres de nuisance tels que le niveau de bruit σ ou la covariance $\mathbf{\Sigma}$ entre les covariables.

6.2.1 Détection de signal

Le premier travail est assez ancien. En collaboration avec Y. Ingster et A. Tsybakov [A28], nous avons caractérisé la distance de séparation minimax pour le problème de détection de signal (c'est-à-dire le test de l'hypothèse nulle $\theta^* = 0$) en régression linéaire parcimonieuse à design gaussien. En d'autres termes, nous avons quantifié le niveau de signal $\mathbb{E}[\|\mathbf{X}\theta^*\|_2^2]$ minimum nécessaire pour réussir à détecter avec grande probabilité la présence de signal. Le procédure utilisée est une combinaison de statistiques du χ^2, du Higher-Criticism [72] et d'une U-statistique.

6.2.2 Estimation du SNR

En collaboration avec E. Gassiat, nous nous sommes intéressés [A14] à l'estimation du rapport signal/bruit (SNR) $\mathbf{E}[\|\mathbf{X}\theta^*\|_2^2]/\sigma^2$ dans le modèle de régression linéaire en grande dimension. Cette quantité est notamment centrale en génétique quantitative pour caractériser l'héritabilité d'un caractère phénotypique. Si le vecteur θ^* de paramètre est très parcimonieux (ex: $\|\theta^*\|_0 \log(p) \leq \sqrt{n}$), alors on peut simplement estimer θ^* avec une méthode de type square-root Lasso [18] et utiliser un estimateur plug-in pour le SNR. Cet estimateur simple atteint la vitesse paramétrique en $n^{-1/2}$. Néanmoins, pour certains phénotype complexes (ex [153]), le caractère parcimonieux du vecteur θ^* correspondant a été remis en cause. Ceci pose alors la question d'estimer le SNR lorsque $\|\theta^*\|_0$ est possiblement grand, voir $\|\theta^*\|_0 = p$, auquel cas il devient impossible d'obtenir une estimation précise de θ^*. Dans ce cadre, Dicker [69] a proposé une U-statistique, qui tout au moins lorsque les covariables sont indépendantes, estime à la vitesse paramétrique le SNR dans

le régime $p \asymp n$. Notre contribution dans [A14] est double: (i) nous avons d'abord caractérisé la vitesse minimax d'estimation du SNR en fonction de $\|\theta^*\|_0$, p, n et construit une procédure adaptative à la parcimonie inconnue. De façon intéressante, il est possible d'estimer le SNR de façon consistante tant que $p = o(n^2)$, ceux même en l'absence de parcimonie. (ii) Nous avons montré que, lorsque la covariance entre les covariables est inconnue, alors il devient impossible d'estimer de façon consistante le SNR lorsque $p \gg n$ en l'absence d'hypothèse de parcimonie.

Cela illustre le rôle central tenu par la connaissance des paramètres de nuisance (ici la loi du design) sur la difficulté de certains problèmes d'estimation de fonctionnelles.

6.2.3 Test et estimation de parcimonie

Avec A. Carpentier, nous nous sommes intéressés à la problématique générale d'estimation de la complexité d'un signal. De nombreuses procédures statistiques modernes cherchent à utiliser la structure sous-jacente du signal (ex: la parcimonie d'un vecteur, la régularité d'une fonction, le faible rang d'une matrice) pour améliorer les propriétés d'inférence ou de prédiction. Estimer/Tester la complexité revient alors à essayer d'inférer/tester ce niveau de parcimonie, cette régularité, ce rang... Dans le cadre des modèles de séquence gaussienne et de régression linéaires, nous avons ainsi considérer les problèmes de test et d'estimation de la parcimonie $\|\theta^*\|_0$ du vecteur de paramètres [A8, A5].

D'un point de vue mathématique, il s'agit d'un problème de test emblématique pour lequel les hypothèses nulles et alternatives sont toutes deux composées. Au contraire du problèmes de détection de signal pour lequel l'hypothèse nulle est simple, la difficulté du problème de test d'hypothèses $\{\|\theta^*\|_0 \leq k_0\}$ contre $\{\|\theta^*\|_0 = k_0 + \Delta\}$ dépend de la taille de ces deux espaces de paramètres. Dans [A8], nous avons caractérisé la distance minimax de séparation d'hypothèse pour tout $k_0 \geq 0$ et tout $\Delta > 0$ dans le modèle de séquence gaussienne. Notamment, lorsque la parcimonie de l'hypothèse nulle est faible ($k_0 \leq \sqrt{n}$), la distance de séparation minimax ne dépend pas de k_0. En revanche, pour $k_0 > \sqrt{n}$, les distance de séparation dépendent de façon subtiles de k_0 et de Δ. Nous obtenons également des résultats partiels dans le modèles de régression linéaire [A5].

La construction du résultat d'impossibilités repose sur des méthodes dites de 'moment matching' tel qu'introduites par Lepski et al. [143]: nous contruisons deux loi a prioris μ_0 et μ_1 sur les ensemble de parametres correspondant aux hypothèses nulles et alternatives respectivement. Si ces deux mesures ont leur $\log(n)$ premiers moments égaux, alors on peut montrer qu'il est impossible de tester si θ^* a été tiré selon μ_0 et μ_1. L'enjeu devient alors de construire de telles mesures μ_0 et μ_1 telles que μ_1 se concentre sur des valeurs de paramêtres θ^* le plus éloignés possible de l'hypothèse nulle.

6.2.4 Tests multiples avec distribution nulle inconnue

Le dernier travail vise à donner une justification théorique aux travaux d'Efron en tests multiples. En réexaminant de nombreux jeux de données classiques, Efron [74, 75, 77] a fait valoir que, dans de nombreux cas, la distribution nulle utilisée pour réaliser un test multiple est mal choisie et doit en fait être re-estimée à partir des données. En conséquence, le statisticien doit à partir des même données à la fois inférer des paramètres de l'hypothèse nulle et réaliser l'ensemble des test multiples. Efron considère un cadre

statisque similaire au modèle de séquence gaussienne: $Y = \theta^* + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ et σ inconnu. On cherche tester les hypothèse $H_{0,i} : \theta_i^* = \theta_0$ contre $H_{1,i} : \theta_i^* > \theta_0$ (tests unilatéraux) ou $H'_{1,i} : \theta_i^* \neq \theta_0$ (tests bilatéraux). La difficulté est que la valeur de référence θ_0 est inconnue et doit être estimée au même titre que σ.

Dans ses travaux, Efron a proposé une approche bayésienne mais il n'y avait, jusqu'à présent, aucune analyse fréquentiste de ce problème. A. Carpentier, S. Delattre, E. Roquain et moi-même considérons d'abord le problème de l'estimation de ces paramètres θ_0 et σ et nous montrons ensuite dans quels cas l'erreur d'estimation est assez faible pour qu'une méthode de Benjamini-Hochberg basée sur un plug-in des estimateurs a des performances contrôlées. Réciproquement, nous montrons dans quel cas, il est impossible de construire une procédure de tests multiples aussi performante que si la distribution nulle était connue à l'avance. Nous traitons à la fois le cas des alternative unilatérales [A4] et bilatérales [A1].

Nous nous éloignons également du modèle de séquence gaussienne en permettant aussi aux distributions alternatives d'être non gaussiennes. Dans ce cadre, l'estimation des paramètres θ_0 et σ se rapproche du cadre des modèles de contamination de Huber [112] en statistiques robustes. Les arguments utilisés combinent des idées d'estimation minimax robuste avec des des mécanismes de contrôle du taux de fausses découvertes (FDR).

6.3 Analyse de réseaux

Cette section est dédiée à l'analyse statistique de graphe. On observe la matrice d'adjacence $\mathbf{A} \in \{0,1\}^{n \times n}$ d'un graphe aléatoire simple non orienté à n noeuds. On s'intéresse à deux objectifs bien distincts: (i) détecter si le graphe est homogène ou s'il existe des groupes de noeuds inhabituellement connectés [A21, A3, A20] et (ii) estimer la distribution de $\mathbf{A}$ de façon non-paramétrique [A16, A9].

6.3.1 Détection de communautés

Le problème de détection de communauté parfois appelé problème du sous-graphe planté est le suivant. Sous l'hypothèse nulle, on observe un graphe d'Erdös-Renyii dont la probabilité de connection est p_0. Sous l'hypothèse alternative, il existe un sous-ensemble de m noeuds (ici $m \ll n$) dont la probabilité de connection entre eux vaut $p_1 > p_0$. Une instance particulière de ce problème avec $p_0 = 1/2$ et $p_1 = 1$ correspond à une version aléatoire du problème de la clique plantée, qui est central en théorie de la complexité [7].

En collaboration avec E. Arias-Castro, nous avons caractérisé l'ensemble des valeurs de (p_0, p_1, m) pour lequel il est possible de détecter avec certitude l'existence de cette communauté. Lorsque le graphe est relativement dense [A21] (p_0 est assez grand par rapport a $1/m$), alors le problème n'est pas structuellement très différent du problème de la détection de sous-matrices gaussiennes [39] et les tests optimaux correspondant calculent soit le nombre total d'arêtes, soit le nombre maximum d'arêtes dans les sous-graphes à m noeuds (test de scan). En revanche, lorsque le graphe est plus parcimonieux [A20], alors il existe des procédure plus subtiles de détection de communautés qui s'appuient sur la géométrie des graphes d'Erdös-Renyii et le comportement de processus de branchement de poissons multi-types.

6.3.2 Estimation de graphons

A l'opposé de modèles d'Erdös-Renyii ou de problèmes à trois paramètres (p_0, p_1, m), nous considérons le problème d'estimation de la distribution de $\mathbf{A}$ sous l'unique hypothèse que cette distribution est échangeable, c'est à dire invariante par permutation des noeuds. Les travaux d'Aldous-Hoover et de Diaconis-Jansson [66] ont montré que la distribution de $\mathbf{A}$ peut alors être caractérisée par un graphon [150], qu'on considére comme une fonction mesurable $W : [0,1] \times [0,1] \to [0,1]$. Le modèle aléatoire correspondant, dit du W-graphe aléatoire, stipule que (a) on tire uniformément pour chaque noeud i une étiquette $\xi_i \in [0,1]$, puis (b) pour chaque paire de noeuds i et j, on tire une arête avec probabilité $W[\xi_i, \xi_j]$.

Avec O. Klopp et A. Tsybakov, nous considérons le problème d'estimation du graphon W lorsque lorsque la matrice $\mathbf{A}$ est tirée selon un tel modèle de W-graphe aléatoire possiblement parcimonieux, c'est à dire que la probabilité de connection vaut $\rho_n W[\xi_i, \xi_j]$ où ρ_n est possiblement petit. Nous étendons les résultats précédents de Gao et al. [90] dans deux directions différentes: premièrement, nous caractérisons la vitesse minimax d'estimation de la matrice $(W[\xi_i, \xi_j])_{i,j}$ en norme de Frobenius, ce pour toute valeur ρ_n. Deuxièmement, nous établissons la vitesse optimale d'estimation du graphon W en distance δ_2 qui s'interpréte comme une distance l_2 entre fonctions. La difficulté de ce deuxième résultat provient du fait que les graphons souffrent de graves soucis d'identifiabilité. Pour rendre le problème identifiable, la distance δ_2 est donc définie sur des classes d'équivalence [150] de graphons pour une certaine notion dite d'isomorphisme faible. Il est alors difficile de minorer la distance δ_2 entre des classes d'équivalence, ce qui rend la preuve des minorations minimax assez technique.

Dans [A9], O. Klopp et moi-même considérons l'estimation de graphons par rapport à la distance cut. En comparaison de δ_2, cette métrique a la vertu de mieux traduire des propriétés structurelle du graphe telles que des nombres d'homomorphismes. Le travail [A9] est plus intéressant d'un point de vue conceptuel que pratique. Nous y démontrons en effet qu'un estimateur trivial (i.e. prendre les donnée brutes $\mathbf{A}$ sans les lisser) s'avère être quasi-toujours optimal. Néanmoins, les estimateurs classiques basés sur le seuillage de valeurs singulières de $\mathbf{A}$ sont également optimaux. Estimer au mieux un graphon en distance cut est donc trivial en pratique. Notre résultat principal est une caractérisation fine de la vitesse optimale d'estimation d'un graphon pour cette métrique cut, lorsque le graphon correspond à un modèle à blocs stochastique (voir la section suivante pour la définition). La preuve utilise notamment des variations du lemme de régularité de Szemerédi [185].

6.4 Clustering

6.4.1 Modèle à Partition Plantée

Les problèmes de clustering consistent à regrouper des "objets" similaires. Ces objets peuvent être des points dans un espace métrique, les noeuds d'un graphe, des courbes,... Ces problèmes ont suscité des travaux très divers tant en statistique qu'en informatique théorique. Dans cette thèse, on s'intéresse unique à la perspective probabilitiste, dite de la partition cachée ou plantée. De manière informelle, cette perspective postule qu'il existe une vraie partition inconnue $G^* = (G_1, \dots, G_K)$ de ces n objets en K groupes.

L'ensemble de données $\mathbf{X}$ est alors supposé avoir été tiré selon une distribution $\mathbb{P}_{G^*}$ dont les caractéristiques dépendent de G^*. L'objectif est donc de retrouver cette partition cachée G^* à partir d'une observation $\mathbf{X}$ de $\mathbb{P}_{G^*}$. Dans ce formalisme, nous pouvons résumer le processus de génération des données et d'analyse statistique comme suit

$$G^* \underset{\mathbb{P}_{G^*}}{\to} \mathbf{X} \to \widehat{G} \ .$$

Etant donnée une partition G^* de $[n]$ et $a \in [n]$, on note dans la suite $k^*(a)$ pour le groupe du a-ième objet. Certains des modèles probabilitistes les plus classiques de clustering s'inscrivent dans ce cadre. Par exemple, on peut citer les modèles de mélange gaussien (GMM) ou les modèles à blocs stochastiques (SBM) définis ci-dessous.

Définition 6.1 (Modèle de mélange gaussien (GMM) conditionnel)**.** *Soient K vecteurs $\mu_1, \dots, \mu_K \in \mathbb{R}^p$, K matrices de covariance $\mathbf{\Sigma}_1, \dots, \mathbf{\Sigma}_K \in \mathbb{R}^{p\times p}$ et une partition G^* de $[n]$ en K groupes. Pour $a = 1, \dots, n$, les lignes X_a de $\mathbf{X} = (X_1, \dots, X_n)^T \in \mathbb{R}^{n\times p}$ sont indépendantes et satisfont $X_a \sim \mathcal{N}(\mu_{k^*(a)}, \mathbf{\Sigma}_{k^*(a)})$.*

Cette définition diffère légérement du cadre classique des mélanges gaussiens car la partition G^* est considérée comme fixe au lieu d'avoir été tirée selon une loi multinomiale.

Définition 6.2 (Modèles à blocs stochastiques (SBM) conditionnel)**.** *Soient $\mathbf{Q} \in [0,1]^{K\times K}_{\text{sym}}$ une matrice symétrique de probabilités et G^* une partition de $[n]$. La matrice $\mathbf{X} \in \mathbb{R}^{n\times n}$ correspond à la matrice d'adjacence d'un graphe simple non orienté, c'est-à-dire que $\mathbf{X}$ est symétrique et sa diagonale est nulle. Pour tout $1 \le a < b \le n$, les $\mathbf{X}_{a,b}$ sont indépendants et satisfont $\mathbf{X}_{a,b} \sim \mathcal{B}(\mathbf{Q}_{k^*(a),k^*(b)})$.*

Comme précédemment, ceci diffère légèrement de la définition habituelle des SBM [110] car la partition G^* est considérée fixe. Lorsque la matrice $\mathbf{Q}$ vaut $\mathbf{Q} = (\alpha - \beta)\mathbf{I}_K + \beta\mathbf{J}_K$ où $\mathbf{J}_K$ est la matrice constante 1 et $0 < \beta < \alpha < 1$, on obtient le modèle classique, dit d'affiliation, pour lequel la probabilité α de connection intra-groupe est plus grande que la probabilité β inter-groupes. Plus généralement, le problème de clustering est identifiables pour des matrices $\mathbf{Q}$ générales dès lors que toutes les lignes de $\mathbf{Q}$ sont distinctes.

Dans ces deux modèles, l'objectif général est d'estimer à partir des données brutes $\mathbf{X}$ une partition $\widehat{G}$ qui est aussi proche que possible de la vraie partition G^*. Si possible, la procédure correspondante doit s'exécuter en temps polynomial par rapport à la taille (K, n, p) du problème. Ce problème a suscité beaucoup d'intérêt tant pour les modèles de mélange gaussien que pour les modèles à blocs stochastiques. De nombreuses procédures différentes ont été étudiées, notamment des méthodes spectrales [142, 149], des programmes semi-définis [5, 158], les algorithmes de Loyd [152] ou plus généralement les algorithmes itératifs [93]. Pour les SBM, des procédures plus spécifiquement adaptées aux graphes parcimonieux telles que [3, 28] ont également été proposées –voir l'article de synthèse d'Abbe [2].

6.4.2 Analyse d'une version convexifiée de K-means

Lorsque les objets que l'on souhaite regrouper correspondent à des vecteurs dans un espace euclidien, l'une des approches de clustering les plus courantes est basée sur la minimisation

du critère K-means [147]. En écrivant $X_a \in \mathbb{R}^p$ pour l'objet $a \in [n]$, le critère K-means d'une partition $G = (G_1, \dots, G_k)$ de $[n]$ est défini comme suit

$$\mathrm{Crit}(G) = \sum_{k=1}^{K} \sum_{a \in G_k} \left\| X_a - \frac{1}{|G_k|} \sum_{b \in G_k} X_b \right\|_2^2, \tag{6.1}$$

où $\|.\|_2$ est la norme euclidienne. Comme la minimisation exacte du critère est NP-dure [10], on a en pratique souvent recours à procédures de minimisation itérative telles que l'algorithme de Lloyd [147] et ses variantes [8], mais ces procédures ne convergent que vers un minimum local du critère, à moins que l'initialisation ne soit suffisamment proche du minimum global. Comme alternative, Peng et Wei [168] ont proposé de relâcher le critère K-means en un programme semi-défini (SDP) suivi d'une étape d'arrondi. La procédure correspondante est calculable en temps polynomial.

Dans un travail joint [A13] avec C. Giraud, nous avons étudié cette version SDP de K-means aussi bien pour les GMM que pour les SBM. Dans ces deux modèles nous caractérisons la proportion d'individus mal classés et nous montrons que celle-ci décroit exponentiellement vite avec le rapport signal/bruit (voir la définition dans le chapitre 4).

Pour être plus concret, considérons un GMM avec une partition G^* équilibrée ($|G_k^*| \asymp n/K$) et une matrice de covariance commune $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_K = \sigma^2 \mathbf{I}_p$, le cas général étant traité dans [A13]. Notons $\Delta^2 = \min_{j \neq k} \|\mu_j - \mu_k\|_2^2$ la distance entre groupes. On définit alors le niveau de signal

$$s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{n\Delta^4}{Kp\sigma^4}.$$

Nous établissons que, si $s^2 \gtrsim K$, alors, avec grande probabilité, la proposition d'individus mal classés dans la partition estimée est plus petite que $e^{-c's^2}$ où $c' > 0$. Cette erreur de classification est optimale (à constante c' près) [161]. L'intérêt de ce résultat par rapport à la littérature existante est qu'il est valables dans un cadre général, incluant des matrices de covariance différentes et possiblement non sphériques et des cadres de grande dimension $p \gg n$.

6.4.3 Clustering avec K grand

Néanmoins, la condition ($s^2 \gtrsim K$) de signal nécessaire pour notre résultat de convergence exponentiel de l'erreur n'est pas minimale. Notamment, il est connu [173] que, tout au moins en petite dimension ($p \ll n$), il est possible de construire un estimateur $\widehat{G}$ dont l'erreur de classification est meilleure qu'une partition tirée au hasard, dès que le SNR satisfait $s^2 \gtrsim \log(K)$. Nous conjecturons que l'estimateur exact K-means qui minimise exactement (6.1) atteint une erreur en $e^{-c's^2}$ dès que $s^2 \gtrsim \log(K)$.

En revanche, la méthode correspondante souffre d'une complexité exponentielle. Tout au moins en grande dimension (lorsque p est au moins de l'ordre n), nous conjecturons également que la condition de signal $s^2 \gtrsim K$ ne peut pas être affaiblie pour des procédures de complexité polynomiale. Ce point est notamment discuté dans un travail joint [A11] avec J. Banks, C. Moore, J. Xu, et R. Vershynin.

6.5 Détection de rupture

J'ai récemment consacré deux travaux à la détection de ruptures avec M. Fromont, M. Lerasle et P. Reynaud-Bouret d'une part [P4] et avec E. Pilliat et A. Carpentier [P3] d'autre part. Dans ce résumé, je décris essentiellement le premier.

Considérons le modèle prototypique de détection de ruptures univariées $Y = \theta^* + \epsilon$ où $\theta^* = (\theta_1^*, \dots, \theta_n^*)$ est dans $\mathbb{R}^n$ et ϵ est un bruit gaussien standard. A partir de Y, l'objectif est de trouver les points de ruptures, c'est à dire les coordonnées où le vecteur moyenne θ^* varie. Par reparamétrisation, il existe un entier $0 \leq K \leq n-1$, un vecteur d'entiers $\tau^* = (\tau_1^*, \dots, \tau_K^*)$ satisfaisant $1 = \tau_0^* < \tau_1^* < \dots < \tau_K^* < \tau_{K+1}^* = n+1$, et un vecteur $\mu = (\mu_1, \dots, \mu_{K+1})$ dans $\mathbb{R}^{K+1}$ satisfaisant $\mu_k \neq \mu_{k+1}$ pour tout k dans $\{1, \dots, K\}$ tel que $\theta_i^* = \sum_{k=1}^{K+1} \mu_k \mathbf{1}_{\tau_{k-1}^* \leq i < \tau_k^*}$. Voir la figure 6.1. Alors, τ_k^* est appelé la *position* du k-ième saut et $\Delta_k = \mu_{k+1} - \mu_k$ est appelé la *hauteur* du k-ième saut.

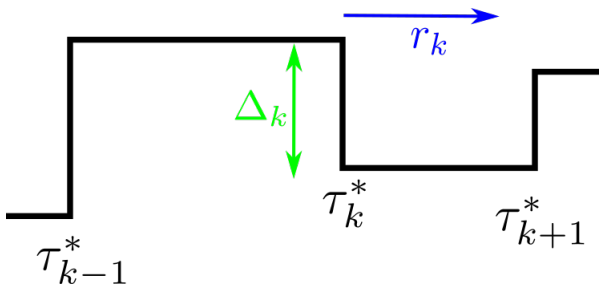

Figure 6.1: Sauts τ_{k-1}^*, τ_k^*, τ_{k+1}^* du vecteur θ^*. Ici, Δ_k corresponds à la hauteur du k-ième saut et $r_k = (\tau_{k+1}^* - \tau_k*) \wedge (\tau_k^* - \tau_{k-1}^*)$ à sa longueur correspondante.

Avec ces notations, notre objectif est de construire un estimateur $\hat{\tau}$ de τ^* des positions de saut qui soit optimal tant en terme de *détection* des sauts qu'en terme de *localisation* des sauts:

(a) On dit qu'un estimateur $\widehat{\tau}$ *détecte* un vrai saut τ_k^* s'il existe un saut estimé $\widehat{\tau}_l$ qui est plus proche de τ_k^* qu'il ne l'est de n'importe quel autre saut. Inversement, on dit que $\widehat{\tau}$ détecte à tort un saut s'il existe deux sauts estimés $\widehat{\tau}_l$ et $\widehat{\tau}_{l'}$ qui sont proches d'un même saut τ_k^*. Le défi consiste alors à construire un estimateur $\widehat{\tau}$ qui, avec grande probabilité, détecte tous les vrais sauts significatifs tout n'en détectant aucun à tort. À cette fin, nous introduisons l'énergie $\mathbf{E}_k^2 = \Delta_k^2[(\tau_{k+1}^* - \tau_k*) \wedge (\tau_k^* - \tau_{k-1}^*)]$ d'un vrai saut comme le carré de la hauteur saut multiplié par la distance de τ_k^* au saut le plus proche. Dans [P4], nous caractérisons les conditions minimales sur l'énergie du saut qui permettent de détecter le saut.

(b) Si un véritable saut τ_k^* a été détecté, on cherche à le localiser le mieux possible, c'est-à-dire à minimiser la distance $d_{H,1}(\widehat{\tau}, \tau_k^*) = \min_{i=1,\dots,|\widehat{\tau}|} |\widehat{\tau}_i - \tau_k^*|$ entre τ_k^* et le saut estimé le plus proche. Cette quantité $d_{H,1}(\widehat{\tau}, \tau_k^*)$ est appelée *erreur de localisation* de $\widehat{\tau}$ pour τ_k^*. Outre l'erreur de localisation d'un saut spécifique, nous caractérons dans [P4] la vitesse optimale pour les erreurs de Hausdorff et de Wasserstein, qui correspondent respectivement au supremum et à la somme des erreurs de localisation.

Nous construisons également dans [P4] deux procédures de détection de ruptures qui sont à la fois optimale en détection et en localisation. La première procédure est basée

sur une versions pénalisée de l'estimateur des moindres carrés tandis que la deuxième est une méthode d'agrégation de tests locaux dont le coût computationnel est quasi-linéaire.

Dans [P3], nous étendons cette analyse à des modèles plus généraux incluant la détection de ruptures pour des séries temporelle multivariée ou la détection de ruptures non-paramétrique.

Dans le futur, je compte m'intéresser avec mes collègues au problème de segmentation sur graphe qui peut s'interpréter comme un continuum entre le clustering classique et la détection de ruptures.

Problème ouvert 6.1 (Segmentation sur un graphe général)**.** *Considérons un graphe non orienté $\mathcal{G} = ([n], E)$ avec n sommets. Pour chaque sommet $a = 1, \dots, n$, on observe $Y_a \sim \mathcal{N}(\theta_a^*, \mathbf{I}_p)$ de moyenne inconnue $\theta_a^* \in \mathbb{R}^p$. Soit $G^* = (G_1^*, \dots, G_K^*)$ la partition de $[n]$ qui regroupe les valeurs identiques de θ_a^*. Sous réserve que la partition G^* ait une petite frontière sur le graphe $\mathcal{G}$, quelle est la différence minimale entre les moyennes pour que l'on soit capable de reconstruire partiellement G^* ? Quelle est l'erreur minimale de reconstruction de la partition G^* ?*

Remarquons que si $\mathcal{G}$ est un graphe ligne, alors il s'agit d'un problème de détection de ruptures multivariées. Si $\mathcal{G}$ est un graphe complet (ou alors un graphe sans arête), alors ce problème est équivalent à problème de clustering dans un modèle de mélange gaussien tel que décrit dans la section précédente.

Bien que le problème d'estimation du vecteur θ^* soit relativement bien compris, au moins dans le cas univarié [81], il n'existe que des résultats très partiels en ce qui concerne l'estimation de la partition G^* –voir par exemple [203].