Nouvelles méthodes pour l'apprentissage non-supervisé en grandes dimensions.

Hafiz Tiomoko Ali

université PARIS-SACLAY

CentraleSupélec

NNT : 2018SACLC074

Thèse de doctorat

Nouvelles méthodes pour l'apprentissage non-supervisé en grandes dimensions

Thèse de doctorat de l'Université Paris-Saclay
préparée à CentraleSupélec

École doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Traitement du Signal et des Images

Thèse présentée et soutenue à Gif-sur-Yvette, le 24/09/2018, par

M. Hafiz Tiomoko Ali

Composition du Jury :

Mme Michèle Sebag Professeur, Université Paris-Sud	Présidente
M. Pierre Borgnat Directeur de recherche (CNRS), ENS de Lyon	Rapporteur
M. Konstantin Avrachenkov Directeur de recherche (INRIA), Sophia Antipolis	Rapporteur
Mme Lenka Zdeborová Chargée de recherche (CNRS), CEA-Saclay	Examinatrice
M. Marc Lelarge Directeur de recherche (INRIA), ENS Paris	Examinateur
M. Jamal Najim Directeur de recherche (CNRS), Université Paris-Est	Examinateur
M. Frédéric Pascal Professeur, CentraleSupélec	Examinateur
M. Romain Couillet Professeur, CentraleSupélec	Directeur de thèse

Titre: Nouvelles méthodes pour l'apprentissage non supervisé en grandes dimensions.

Mots-clés: Apprentissage non supervisé, Classification de données en grandes dimensions, Détection de communautés, Théorie des matrices aléatoires, Inférence bayésienne.

Résumé: De nos jours, une pression industrielle très forte avec l'essor de l'intelligence artificielle pousse les chercheurs à comprendre le fonctionnement des algorithmes d'apprentissage automatisé (AA), surtout pour des données à grandes dimensions, dans le but de les améliorer. La réussite de ces algorithmes d'AA vient principalement du fait qu'ils s'adaptent à des données réelles et qu'ils utilisent des transformations non linéaires sur ces dernières. Cela rend cependant l'étude théorique de ces méthodes plus difficile. De plus, le phénomène de la "malédiction de la dimensionalité" ou "curse of dimensionality" en anglais fait que certaines intuitions au coeur d'algorithmes d'AA en petites dimensions ne sont plus vraies lorsque les données sont de très grandes tailles. Par contre, comme nous le verrons au cours de ce manuscript de thèse, ce phénomène de la "malédiction de la dimensionalité" est plutôt un avantage pour l'analyse théorique des algorithmes en grandes dimensions du fait notamment d'un phénomène important de "concentration asymptotique" de certaines quantités-clés. Cette thèse est focalisée sur les méthodes d'apprentissage non supervisées notamment de "clustering". La première partie de la thèse est dédiée au *clustering de graphes* tandis que la seconde partie est focalisée sur le *clustering de données*.

Dans le Chaptitre 3, nous proposons pour des modèles de graphes dits "denses" présentant une structure de communautés et une structure de degrés hétérogènes, une famille de méthodes dites *spectrales* utilisant des matrices d'affinités qui dépendent d'un paramètre de régularisation pour pallier à l'effet néfaste des degrés sur la lecture des communautés dans les vecteurs propres de ces matrices. Nous montrons que ces matrices de similarité sont asymptotiquement équivalentes à des matrices suivant un modèle de matrices aléatoires connu dit *spike* présentant une structure forte de communautés dans les vecteurs propres associés aux plus grandes valeurs propres de ces matrices. Une analyse asymptotique poussée des valeurs propres et vecteurs propres de ces matrices nous permet de **i)** déterminer *a priori* le paramètre de régularisation le plus approprié et donc la méthode spectrale optimale pour un type de graphe donné; **ii)** proposer une amélioration de l'algorithme Expectation Maximization (EM) utilisé dans la dernière étape du clustering spectral, en initialisant EM avec nos découvertes théoriques sur les limites asymptotiques des entrées des vecteurs propres.

Dans le Chapitre 4, nous proposons une nouvelle méthode pour la détection de communautés multi-graphes (c'est-à-dire communes et non communes) entre les différentes couches du graphe avec plusieurs types d'interaction. Nous proposons ici un modèle probabiliste de graphes à plusieurs couches qui tient compte de la disparité des communautés entre les couches. Sur la base de ce modèle, une approche bayésienne variationnelle est utilisée pour approximer la distribution a posteriori des variables latentes corrélées représentant les communautés des différentes couches dont les liens sont observés. L'algorithme proposé peut être appliqué à tout réseau multi-couches avec des liens pondérés ou non.

Enfin, le Chapitre 5 traite de la question du clustering spectral de données en grandes dimensions utilisant des similarités à base de noyaux (kernels). Nous proposons une analyse spectrale des matrices de similarités (en termes de produit scalaire) à noyau, de données provenant d'un mélange d'un petit nombre de vecteurs gaussiens. Nous montrons que ces matrices sont asymptotiquement équivalentes à des matrices aléatoires à modèles spikes comme dans le Chapitre 3. L'analyse précise de ces modèles spikes nous permet d'identifier de nouvelles fonctions à noyau paramétrisées de telle sorte à induire de meilleures performances comparées aux noyaux standards, et capables de discriminer les données à la fois sur la base de la différence entre leurs moyennes statistiques et/ou de la différence entre leurs covariances statistiques.

Title: New methods for large scale unsupervised learning.

Keywords: Unsupervised learning, High dimensional data clustering, Community detection, Random matrix theory, Bayesian inference.

Abstract: The industrial pressure of finding efficient Machine Learning (ML) algorithms especially for very large datasets pushes researchers to understand the behavior of those algorithms in order to improve them. The power of ML algorithms mostly comes from the fact that they are data-driven and rely on nonlinear transformations. This however makes the theoretical analysis of their performances more challenging. Additionally, due to the *curse of dimensionality* in ML, some original intuitions at the basis of algorithms adapted to small dimensional data are not valid anymore in high dimensions. However, as we shall see in the course of this thesis, this curse of dimensionality problem is turned into a *blessing* notably due to the asymptotic *concentration* of certain key quantities allowing to theoretically capture the behavior of ML algorithms in high dimensions. This thesis focuses on unsupervised learning methods and particularly on *clustering*. The first part is dedicated to the question of *graph clustering* while the second part focuses on *data clustering*.

In Chapter 3, for realistic *dense* random graph models with communities and heterogeneous degree distributions, we derive a family of modified spectral clustering algorithms using similarity matrices depending on a regularization parameter to handle the deleterious effects of degree heterogeneity. Those similarity matrices are shown to be asymptotically equivalent to a family of spiked random matrices which exhibit strong structures such that (a small number of) communities can be extracted from the eigenvectors associated to the dominant eigenvalues of those matrices. For graphs with a large number of nodes, a thorough study of the eigenvalues and eigenvectors of those random matrices allows to **i)** derive an online selection of the most appropriate regularization parameter and thus an improved clustering method adapted to these graphs **ii)** propose an improvement of the Expectation Maximization (EM) algorithm in the last step of the spectral method, which uses the asymptotic limit of the eigenvectors entries' statistics to initialize EM instead of a random initialization.

In Chapter 4, we propose a new method for the detection of overlaping and non-overlapping communities in multi-layer graphs. Here, a probabilistic multi-graphical model accounting for disparing communities between the different layers is proposed. Based on this model, a variational Bayes approach is used to approximate the intractable posterior distribution of the latent communities of the different layers given the observed graphs. The proposed algorithm which can be applied to any number of network layers is applied to a real genome-wide fibroblast proliferation dataset, revealing important biological insights on the interplay between functional and spatial relationships between the genes.

Finally, Chapter 5 treats the question of high dimensional data spectral clustering using kernels. We proceed to a spectral analysis of large dimensional *inner product* kernel matrices, based on Gaussian mixture inputs, which are shown to be equivalent to tractable random matrices in the family of spiked models. A precise analysis of the spiked random matrix led us to identify new parametrized kernel functions outperforming standard kernels in discriminating data classes through their differences in statistical means and/or differences in statistical covariances.

... I dedicate this thesis to my father Djafarou and my mother Adiza,
for the perfect examples they have always been since my childhood...

Acknowledgments

My engineering background allowed me to have a general knowledge on how to apply many different technologies. However, something was missing: How do they work? This curiosity led me to go for a PhD in order to learn and apply the theoretical tools which could allow me to better understand different technologies in order to improve on them. By the end of this aventure, I can assert that I have reached this objective and I have gained a lot in scientific maturity. This would have really been difficult without the support of many people surrounding me.

Many thanks to:

- My PhD advisor, Professor Romain Couillet. I could write a whole story about him as he is incredibly amazing in many aspects. From the beginning to the end of my PhD, he has always guided me towards the right path. In my opinion, he pocesses all the qualities of a PhD advisor and I always say that it is a great chance to have him as an advisor. Thank you Romain!

- Professor Mérouane Debbah, who has identified me when I was studying in an engineering school in Algeria, to work in his lab at Supélec. Working with such an amazing researcher and leader with a great personality has helped me a lot during this path. Thank you Mérouane!

- Abla Kammoun (an amazing researcher!) who was my main collaborator during this thesis. To all my colleagues and friends that I have met during this whole path. It was a great honor for me to have spent amazing moments with them!

- Michigan friends. During my PhD, I had the opportunity to go for a 5-month internship at the University of Michigan where I have met amazing people (that I thank all!) and learnt many things. I would like to thank Professor Alfred Hero for hosting me into his lab and for his contributions in my research.

- My sister Chérifath and my brothers Malik and Kémal for their overall support. My siblings are big fighters; they always work hard (as our parents always taught us) to get to their objectives. My uncle Idrissou (and his family) who is like a father for me. My big family in Benin (uncles, aunts, cousins,..) for their support since my childhood.

- My parents have always been perfect examples for me since my childhood. I always owe them all my successes for the perfect education they offered me. My father decided to go for a PhD after he retired and he got it at the age of 55: this was a strong motivation for me to move forward! Thank you “papa” et “maman”!

Contents

Notations

$\mathbf{X}$	Matrix.
$\mathbf{x}$	Column vector by default.
X_{ij}	Entry (i, j) of matrix $\mathbf{X}$.
x_i	Entry i of vector $\mathbf{x}$.
$\{f(i,j)\}_{i,j}^{n}$	Matrix of size $n \times n$ with (i, j) entry $f(i, j)$.
$\mathbf{X}^\mathsf{T}$	Transpose of $\mathbf{X}$.
$\operatorname{tr}\mathbf{X}$	Trace of matrix $\mathbf{X}$.
$\det\mathbf{X}$	Determinant of matrix $\mathbf{X}$.
$\|\mathbf{X}\|$	Spectral norm of symmetric matrix $\mathbf{X}$.
$\operatorname{diag}(x_1, \cdots, x_n)$	Diagonal matrix with (i, i) entry x_i.
$\mathbb{R}$	Space of real numbers.
$\mathbb{C}$	Space of complex numbers.
x^+	Right-limit of the real x.
x^-	Left-limit of the real x.
$(x)^+$	For $x \in \mathbb{R}$, $\max(x, 0)$.
$\mathcal{R}(z)$	Real part of the complex z.
$\mathcal{I}(z)$	Imaginary part of the complex z.
$f'(x)$	First derivative of the function f.
$f''(x)$	Second derivative of the function f.
$f'''(x)$	Third derivative of the function f.
$f^{(p)}(x)$	Derivative of order p of the function f.
$1_A(x)$	Indicator function of the set A, $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ if $x \notin A$.
$\delta(x)$	Dirac delta function, $\delta(x) = 1_{\{0\}}(x)$.
$\mathcal{S}$	Support of distribution function.

$\limsup_n x_n$	limit superior of the series $x_1, x_2, \ldots$.
$\liminf_n x_n$	limit inferior of the series $x_1, x_2, \ldots$, $\liminf_n x_n = -\limsup_n x_n$.
$x_n \to l$	Simple convergence of the series $x_1, x_2, \ldots$ to l.
$x_n = o(y_n)$	Upon existence, $x_n/y_n \to 0$ as $n \to \infty$.
$x_n = \mathcal{O}(y_n)$	There exists K such that $x_n \leq K y_n$ for all n.
$n/p \to c$	As $n \to \infty, p \to \infty$ $n/p \to c$.
$\mu_{\mathbf{X}}$	Probability distribution of the eigenvalues of $\mathbf{X}$.
$\mathbb{E}[X]$	Expectation of the random variable X.
$Var(X)$	Variance of the random variable X.
$X \sim \mu$	X is a random variable with density μ.
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Real Gaussian distribution of mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
$x_n \xrightarrow{\text{a.s.}} l$	Almost sure convergence of the series $x_1, x_2, \ldots$ to l.
$\mu_n \Rightarrow \mu$	Weak convergence of the distribution function serie $\mu_1, \mu_2, \cdots$ to μ.
S^c	Complementary of the set S.
$x \triangleq y$	x is defined as y.

List of Figures

List of Tables

Chapter 1

Introduction (Français)

L'apprentissage automatique (AA) est probablement le domaine le plus populaire de nos jours en sciences techniques aussi bien dans le milieu industriel qu'académique. La puissance de l'AA réside dans sa capacité à traiter de manière automatique et rapide tous types de données qui peuvent être du texte, de l'image, du son, de la vidéo, etc., traitements à partir desquels des informations importantes peuvent être tirées. L'idée derrière l'AA est simple : elle consiste à "apprendre" une fonction *boîte noire* $\mathbf{y} = f(\mathbf{x})$ associant des entrées $\mathbf{x}$ à des sorties $\mathbf{y}$, en utilisant des exemples dits d'apprentissage pour lesquels les sorties sont soit connues d'avance (*apprentissage supervisé*) ou inconnues (*apprentissage non supervisé*). La tâche d'apprentissage est dite de *classification* lorsque les sorties $\mathbf{y}$ sont à valeurs discrètes et dite de *régression* lorsque les sorties sont à valeurs continues. En apprentissage supervisé, les paramètres de la fonction sont appris par le système et ensuite utilisés pour prédire la sortie de nouvelles entrées (test) tandis qu'en apprentissage non supervisé, les exemples d'apprentissage sont directement utilisés pour catégoriser les entrées. Nous nous intéressons dans cette thèse au *clustering*, méthode de classification non-supervisée la plus populaire en AA.

De manière générale, le clustering d'objets consiste à les regrouper de telle sorte que les objets soient très similaires à l'intérieur de chaque groupe. Les objets susmentionnés peuvent soit être des données (qui peuvent être représentées sous forme de vecteurs) ou des graphes (représentant des interactions entre un ensemble de noeuds). Lorsque l'on regroupe des données de manière non-supervisée, on parle de *clustering de données* [Shalev-Shwartz and Ben-David, 2014] où les données sont regroupées en fonction de leurs similarités où la similarité (entre une paire de données) est calculée en utilisant une certaine "mesure de proximité". Le problème du clustering est dit de *détection de communautés* ou *clustering de graphes* [Newman, 2010, Goldenberg et al., 2010, Fortunato, 2010] lorsque l'on a une base de données indiquant la présence ou non d'interactions entre différents individus et la tâche consiste à regrouper ces derniers de telle sorte que le nombre d'interactions entre individus appartenant au même groupe soit important. Dans la majorité des bases de données d'AA, les vecteurs sont généralement non linéairement séparables. Pour ce faire, la classification de ces vecteurs consiste généralement à tout

d'abord les projeter dans un espace à plus grande dimension où ils ont tendance à être linéairement séparables avant de leur appliquer la mesure de proximité qui permettra de faire la classification. On peut imaginer qu'il sera très coûteux de calculer la mesure de proximité entre une paire de vecteurs (généralement un produit scalaire) dans ce nouvel espace à grande dimension. Heureusement, il existe une *astuce* connue sous le nom du *kernel-trick* [Schölkopf, 2001] qui montre que l'application d'une fonction (appelée *noyau*) au produit scalaire entre les vecteurs dans leur espace de départ est équivalente au produit scalaire entre les projetés de ces vecteurs dans le plus grand espace. Le clustering est connu sous le nom de *clustering à noyau* lorsqu'une fonction à noyau est utilisée pour classifier des données non linéairement séparables. Il est facile de voir que le *clustering de données* ou *le clustering de données à base de noyau* sont équivalentes à faire de la classification de données sur un graphe dont les poids sont donnés par les valeurs des similarités.

1.1 Détection de communautés dans les graphes

La détection de communautés dans les graphes (réseaux) est l'un des problèmes les plus fondamentaux de l'analyse de données car elle permet d'explorer et d'analyser les graphes représentant des interactions du monde réel dans de nombreux domaines incluant la sociologie [Goldenberg et al., 2010], la biologie et la médecine [Chen and Yuan, 2006, Marcotte et al., 1999, Cline et al., 2007], le transport [Guimera et al., 2005], l'Internet des Objets [Linden et al., 2003, Clauset et al., 2004]. Un graphe $\mathcal{G}$ est défini comme une paire d'ensembles $(\mathcal{V}, \mathcal{E})$ avec $\mathcal{V}$ l'ensemble des nœuds et $\mathcal{E}$ est un sous-ensemble de toutes les paires d'interactions $\mathcal{V} \times \mathcal{V}$. Le graphe est dit *dirigé* lorsque les éléments de l'ensemble $\mathcal{E}$ sont ordonnés et *non dirigé* sinon. Dans de nombreuses applications, des poids réels peuvent être affectés à chaque élément de $\mathcal{E}$ indiquant la force de la connexion et résultant en un *graphe dit pondéré* tandis que des poids binaires correspondent à un *graphe non pondéré*. Les noeuds d'un graphe peuvent être impliqués dans différents types d'interactions (par exemple des acteurs impliqués différemment dans divers réseaux sociaux), ou les interactions peuvent évoluer au fil du temps (graphes dynamiques) ; dans ces cas, on parle de *graphes multi-couches*. La terminologie *graphe à simple couche* sera utilisée quand un seul type d'interaction caractérise le graphe.

Pour un graphe donné, la détection de communautés consiste à extraire des groupes de nœuds *cachés* ou *latents* de telle sorte qu'il y ait beaucoup d'arêtes à l'intérieur des communautés et peu d'arêtes entre elles. La recherche en détection de communautés sur des graphes à simple couche a été et continue d'être un domaine très actif tant sur le plan théorique que sur le plan algorithmique. Dans les réseaux actuels du monde réel, les individus peuvent être impliqués dans différents types d'interactions, conduisant ainsi à des réseaux/graphes multi-couches/multi-relationnels/multi-dimensionnels. Par exemple, les employés de grandes entreprises peuvent être connectés les uns aux autres en fonction d'activités similaires d'une part et en fonction de leurs activités sociales d'autre part [Oselio et al., 2015]. En génomique, les gènes peuvent être liés soit par leurs in-

teractions transcriptionnelles (relations fonctionnelles), par exemple par la similarité du profil RNA-seq, soit par les interactions de la chromatine (relations spatiales), mesurées par capture de conformation de la chromatine (Hi-C) [Dixon et al., 2015, Boulos et al., 2017, Dekker et al., 2017]. La récente croissance de ces réseaux dynamiques avec interactions hétérogènes a ainsi fait appel à de nouvelles méthodes pour la détection de communautés sur des graphes multi-couches. Deux grandes classes de méthodes sont généralement utilisées pour résoudre le problème de détection de communautés (pour les graphes à simple couche ou multi-couches) : les méthodes *d'inférence statistique* et des méthodes basées sur une *métrique à optimiser*.

Le problème de détection de communautés décrit dans la section 1.1.1 est spécifique aux graphes à simple couche où les noeuds du graphe sont seulement impliqués dans un seul type d'interactions.

1.1.1 Détection de communautés sur graphes à simple couche

Telles qu'indiquées ci-dessus, les méthodes d'inférence statistique sont l'une des grandes catégories d'approches pour la détection de communautés. Elles consistent à associer le graphe observé à un modèle statistique tenant compte de la structure de communautés latente et à estimer les paramètres du modèle (parmi lesquels l'attribution des nœuds aux communautés). Le modèle statistique (incluant une structure de communautés) le plus fondamental qui est souvent utilisé pour l'inférence de communautés sur graphes est le modèle à bloc stochastique (MBS) ou Stochastic Block Model (SBM) en anglais. Soit $\mathcal{G}$, un graphe comportant n sommets (ou noeuds) à k communautés $\mathcal{C}_1, \ldots, \mathcal{C}_k$ avec g_i le groupe assigné au noeud i , le MBS définit une matrice d'adjacence $\mathbf{A} \in \{0,1\}^{n \times n}$ avec A_{ij} variables aléatoires de Bernoulli indépendantes ayant pour paramètre $P_{g_i g_j}$ où P_{ab} représente la probabilité qu'un noeud de la classe $\mathcal{C}_a$ soit connecté à un noeud de la classe $\mathcal{C}_b$. La principale limite de ce modèle est qu'il est seulement adapté aux graphes homogènes où tous les nœuds ont le même degré moyen dans chaque communauté (par ailleurs, les tailles de classes sont souvent prises égales). Un modèle plus réaliste, le MBS à degrés corrigés (MBSDC), a été proposé dans [Coja-Oghlan and Lanka, 2009, Karrer and Newman, 2011] pour tenir compte de l'hétérogénéité des degrés à l'intérieur des communautés. Pour le même graphe $\mathcal{G}$ défini ci-dessus, en définissant q_i, $1 \leq i \leq n$, comme étant des poids intrinsèques qui affectent la probabilité pour le noeud i de se connecter à tout autre nœud de réseau , la matrice d'adjacence $\mathbf{A} \in \{0,1\}^{n \times n}$ du graphe généré par le MBSDC est telle que A_{ij} sont des variables aléatoires indépendantes de Bernoulli avec paramètre $q_i q_j C_{g_i g_j}$, où $C_{g_i g_j}$ est un facteur de correction tenant compte de la communauté d'appartenance de chaque noeud.

La seconde classe de méthodes pour la détection de communautés, basée sur une *métrique* à optimiser, consiste à tout d'abord définir une métrique cohérente avec la définition formelle des communautés (par exemple, la modularité [Newman, 2006b], le ratio de coupure (ratio-cut) [Wei and Cheng, 1989]) puis à maximiser la métrique choisie sur toutes les partitions possibles du graphe. Formulé en tant que tel, le problème de dé-

tection de communautés s'avère être NP-difficile [Brandes et al., 2007]. La littérature dans la détection de communautés sur graphe à simple couche s'est concentrée sur la recherche de méthodes d'approximation polynomiale du problème original en proposant des approches heuristiques [Newman, 2004], d'optimisation [Duch and Arenas, 2005] et des méthodes spectrales [Ng et al., 2002, Newman, 2006b, Newman, 2016]. La relaxation de l'optimisation de la modularité ou de l'optimisation du ratio de coupure, des valeurs de labels discrètes à des valeurs continues, conduit à des méthodes dites spectrales (résumées dans l'Algorithme 1 ci-dessous) qui consistent à extraire les communautés des nœuds en utilisant les vecteurs propres associés aux valeurs propres dominantes de matrices de similarité (matrice d'adjacence, matrice de modularité, matrice laplacienne) représentant le graphe. La matrice de similarité utilisée pour les méthodes spectrales dépend de la métrique considérée; par exemple, l'optimisation de la métrique de modularité conduit à une méthode spectrale utilisant la matrice de modularité, la métrique du ratio de la coupure correspond à la matrice d'adjacence tandis que l'optimisation du ratio de coupure normalisé induit l'utilisation de la matrice laplacienne normalisée. Il est montré dans [Nadakuditi and Newman, 2012] que le seuil de détectabilité des communautés par les méthodes spectrales utilisant la matrice d'adjacence (et ses variantes) correspond au seuil optimal donné par les approches Bayes optimales lorsque les graphes aléatoires considérés sont *denses* (c'est-à-dire que le degré moyen de ces graphes augmente avec la taille du graphe). Cependant, il existe un écart important entre le seuil de détectabilité de ces méthodes spectrales [Kawamoto and Kabashima, 2015] et le seuil Bayes optimal [Decelle et al., 2011b, Decelle et al., 2011a] dans le régime de graphes parcimonieux (c'est-à-dire, quand le degré maximum des nœuds ne croît pas avec le nombre de nœuds); ce qui montre la sous-optimalité de ces méthodes spectrales dans le régime parcimonieux [Krzakala et al., 2013]. Dans le simple cas du MBS avec 2 communautés de même taille, la partie positive du résultat sur la transition de phase [Decelle et al., 2011a] est prouvée dans [Mossel et al., 2013, Massoulié, 2014] tandis que la partie négative est prouvée dans [Mossel et al., 2015]. Il a été montré plus tard qu'une méthode spectrale utilisant les matrices Non-Backtracking [Krzakala et al., 2013] et Bethe Hessian [Saade et al., 2014] comble l'écart avec le seuil du Bayes optimal dans le régime parcimonieux. Une étude poussée des valeurs propres et vecteurs propres de la matrice Non-Backtracking dans [Bordenave et al., 2015] a conduit à prouver la conjecture sur l'optimalité des méthodes spectrales basées sur la Non-Backtracking. Cependant, la convergence de la distribution des valeurs propres de la matrice Non-Backtracking (une matrice non-symétrique) reste un problème ouvert. D'un point de vue de la théorie des matrices aléatoires, la raison de l'échec des méthodes spectrales utilisant des variantes de la matrice d'adjacence dans le régime parcimonieux réside dans le fait que le spectre (distribution des valeurs propres) de ces matrices ne se concentre pas ; il peut y avoir des valeurs propres éloignées, dont les vecteurs propres sont localisés autour de quelques noeuds (appelés *hubs*) et ainsi la structure de communauté globale est perdue dans les vecteurs propres qui sont normalement utilisés dans la classification spectrale. Revenons au régime "dense" où le spectre de *la famille des matrices d'adjacence* se comporte bien. [Nadakuditi and Newman, 2012] a étudié le spectre de la matrice de modularité pour un MBS symétrique, ce qui a conduit à une caractérisation explicite du seuil de détectabilité dans ce régime. A travers une caractérisation du

vecteur propre dominant dans ce régime simple en utilisant notamment une conjecture sur le caractère gaussien des entrées du vecteur propre, [Nadakuditi and Newman, 2012] a déterminé le taux d'erreur de classification d'un algorithme spectral utilisant la matrice de modularité d'un graphe dense. Bien que les travaux de [Nadakuditi and Newman, 2012] soient seulement basés sur la matrice de modularité, les performances (et aussi le seuil de détectabilité) des méthodes spectrales utilisant toutes les matrices dans *la famille des matrices d'adjacence* sont asymptotiquement les mêmes pour des graphes générés suivant le MBS. Cependant, comme nous le montrons dans cette thèse, les performances des différentes méthodes spectrales pourraient être différentes dans le MBSDC dense.

La plupart des travaux théoriques pour la détection de communautés dans les MBSDC denses s'est focalisée sur la preuve de leur consistance asymptotique [1]. Les conditions suffisantes pour lesquelles les approches de détection de communautés basées sur le maximum de vraisemblance [Karrer and Newman, 2011] et les méthodes basées sur l'optimisation de la modularité [Newman, 2006b] sont faiblement et fortement consistantes, ont été établies dans [Zhao et al., 2012]. L'algorithme appelé CMM (Maximisation de la Modularité Convexifiée) a été proposé dans [Chen et al., 2015b] pour faire face à la complexité de calcul des méthodes de modularité/maximum de vraisemblance [Karrer and Newman, 2011, Newman, 2006b], en proposant une solution améliorée de la relaxation du problème d'optimisation de la modularité. En ce qui concerne les méthodes de classification spectrale, il a été montré [Coja-Oghlan and Lanka, 2009, Qin and Rohe, 2013, Jin et al., 2015, Gulikers et al., 2015] que lorsque les degrés sont très hétérogènes, les méthodes spectrales classiques ne parviennent pas à correctement détecter les communautés. Pour illustrer les limitations des méthodes spectrales sous le MBSDC, les deux graphes de la figure 1.1 représentent en 2-D le vecteur propre dominant 1 versus le vecteur propre 2 de la matrice de modularité standard et de la matrice de Bethe Hessian[2], lorsque trois quarts des nœuds se connectent avec un poids faible $q_{(1)}$ et un quart des nœuds avec un poids élevé $q_{(2)}$. Pour les deux méthodes, il est clair qu'un algorithme de classification de type k-means induirait de manière erronée une détection de communautés supplémentaires et même une confusion des communautés dans l'approche de Bethe Hessian. Ces communautés supplémentaires sont créées suite à des biais dûs aux poids intrinsèques hétérogènes q_i ; intuitivement, les nœuds partageant les mêmes poids de connexion intrinsèques tendent à créer leur propre sous-communauté à l'intérieur de chaque communauté, formant ainsi des sous-communautés supplémentaires à l'intérieur des communautés de base. Pour surmonter ce problème, un certain nombre de techniques de classification spectrales régularisées ont été proposées pour normaliser par les degrés, soit la matrice d'adjacence, soit les vecteurs propres dominants. Dans [Coja-Oghlan and Lanka, 2009,Gu-

[1] La consistance est définie principalement sous deux formes. De manière informelle, un algorithme de détection de communautés est **faiblement** consistant lorsque la fraction de nœuds mal classifiés s'annule asymptotiquement avec une probabilité élevée tandis qu'un algorithme de détection de communautés est **fortement** consistant lorsque que les labels assignés aux noeuds correspondent exactement au vrai labels avec une très grande probabilité.

[2] La méthode spectrale basée sur le Bethe Hessian (BH) [Saade et al., 2014] est basée sur l'union des vecteurs propres associés aux valeurs propres négatives de $H(r_c)$ et $H(-r_c)$ respectivement où $H(r) = (r^2-1)\mathbf{I}_n - r\mathbf{A} + \mathbf{D}$ pour $r_c = \frac{\sum_i d_i^2}{\sum_i d_i} - 1$ avec d_i le degré de nœud i ($\mathbf{D}$ et d_i sont définis par la suite).

likers et al., 2015], les auteurs ont proposé de grouper les noeuds en utilisant les vecteurs propres d'une matrice d'adjacence normalisée $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}$ avec $\mathbf{D}$ la matrice diagonale contenant les degrés observés sur la diagonale principale. Nous avons effectué précisément dans [Tiomoko Ali and Couillet, 2016b] une analyse spectrale de la matrice de modularité normalisée $\mathbf{D}^{-1}\left(\mathbf{A}-\frac{\mathbf{d}\mathbf{d}^{\mathsf{T}}}{\mathbf{d}^{\mathsf{T}}\mathbf{1}_n}\right)\mathbf{D}^{-1}$ (ce qui n'est pas différent de $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}$ d'un point de vue classification spectrale) dans un régime dense MBSDC où la classification n'est pas asymptotiquement triviale (c'est-à-dire ni trop facile, ni impossible). Nous y avons déterminé la transition de phase où la classification devient asymptotiquement possible et avons établi, pour de simples jeux de modèles, le taux d'erreur de classification asymptotique. Une autre approche pour palier aux biais induits par le MBSDC consiste à plutôt utiliser les vecteurs propres dominants (pré-normalisés par la matrice de degrés inverse $\mathbf{D}^{-1}$) de la matrice d'adjacence, qui est proposée sous le nom de l'algorithme SCORE dans [Jin et al., 2015]. [Qin and Rohe, 2013] a proposé d'utiliser les vecteurs propres de la matrice Laplacienne $\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$. Comme nous le montrerons dans ce manuscrit, certaines des méthodes précitées ont le désavantage d'avoir un spectre (distribution des valeurs propres) plus étendu et donc une difficulté à atteindre la transition de phase. Nous proposons dans cette thèse, une méthode générique qui englobe ces méthodes précédentes dans un régime MBSDC dense et qui identifie la meilleure matrice qui permet d'atteindre la transition de phase dans des scénarios de classification difficile. Nous indiquons également à travers nos résultats, la bonne normalisation à effectuer sur les vecteurs propres afin d'éviter les biais .

Comme indiqué précédemment, les travaux précités [Coja-Oghlan and Lanka, 2009, Gulikers et al., 2015, Jin et al., 2015] ont montré que sous certaines conditions de régularisation, une reconstruction presque parfaite ou totalement parfaite des communautés auxquelles appartient chaque noeud peut être obtenue asymptotiquement par les méthodes de normalisation précitées. Notre motivation dans cette première partie de la thèse est d'aller au-delà des résultats de consistance pour comprendre les performances des différents algorithmes de classification spectrale régularisée pour des graphes de très grandes tailles n fini. Pour cela, nous nous plaçons dans un régime où les communautés sont trop proches pour induire des reconstructions parfaites, de sorte que les différents algorithmes spectraux ne conduisent pas tous à une même classification triviale asymptotique. Comme indiqué ci-dessus, afin d'englober la plupart des méthodes susmentionnées, nous étudions ici une régularisation généralisée de la matrice d'adjacence [3] donnée, pour tout $\alpha \in \mathbb{R}$, par

$$\mathbf{L}_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} \mathbf{D}^{-\alpha}\left[\mathbf{A}-\frac{\mathbf{d}\mathbf{d}^{\mathsf{T}}}{2m}\right]\mathbf{D}^{-\alpha}$$

[3] Il est montré dans des simulations que le terme principal $\frac{\mathbf{d}\mathbf{d}^{\mathsf{T}}}{2m}$ (ne fournissant aucune information sur les communautés) n'a aucun impact asymptotique sur la performance de classification des graphes. Il est ajouté ici principalement pour faciliter l'exposition des résultats. On note en passant que $\mathbf{A}-\frac{\mathbf{d}\mathbf{d}^{\mathsf{T}}}{2m}$ correspond à la matrice de modularité [Newman, 2006a], donc $\mathbf{L}_\alpha$ peut être vu comme une matrice de modularité normalisée par " *alpha*".

où $\mathbf{d}$ est le vecteur des degrés ($d_i = \sum_{j=1}^{n} A_{ij}$), $\mathbf{D}$ est la matrice diagonale des degrés (contenant $\mathbf{d}$ sur la diagonale principale) et $m = \frac{1}{2}\mathbf{d}^\mathsf{T}\mathbf{1}_n$ est le nombre d'arêtes dans le graphe. En particulier, $\mathbf{L}_0$ est la matrice de modularité [Newman, 2006b, Jin et al., 2015], $\mathbf{L}_{\frac{1}{2}}$ est une modularité équivalente à la matrice laplacienne normalisée [Qin and Rohe, 2013, Chung, 1997] et $\mathbf{L}_1$ est la matrice utilisée dans [Coja-Oghlan and Lanka, 2009, Gulikers et al., 2015, Tiomoko Ali and Couillet, 2016c].

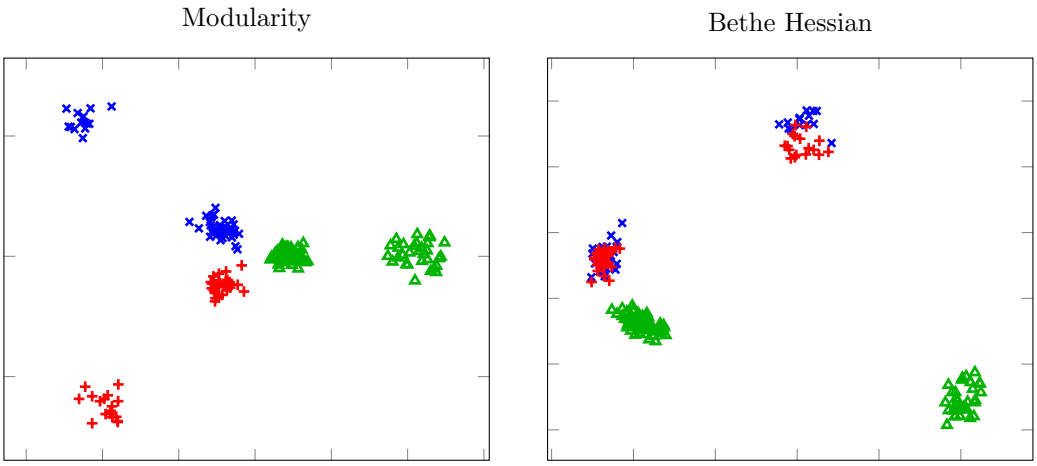


Figure 1.1: Two dominant eigenvectors (x-y axes) for $n = 2000$, $k = 3$ classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ of sizes $|\mathcal{C}_1| = |\mathcal{C}_2| = \frac{n}{4}$, $|\mathcal{C}_3| = \frac{n}{2}$, $\frac{3}{4}$ of the nodes having $q_i = 0.1$ and $\frac{1}{4}$ of the nodes having $q_i = 0.5$, matrix of weights $\mathbf{C} = \mathbf{1}_3\mathbf{1}_3^\mathsf{T} + \frac{100}{\sqrt{n}}\mathbf{I}_3$. Colors and shapes correspond to ground truth classes.

Nous considérons ici un modèle MBSDC *dense* où $q_i = \mathcal{O}(1)$ (par rapport à n) [4]. Dans ce régime, lorsque les facteurs de correction $C_{g_i g_j}$ diffèrent de $\mathcal{O}(1)$, tous les algorithmes spectraux régularisés sont consistants (c'est-à-dire que le taux d'erreur de classification s'annule asymptotiquement). Afin de comparer les différents algorithmes, nous considérons un régime où les communautés sont à peine séparables en regardant la matrice d'affinité de classes $\mathbf{C}$, mais toujours identifiables. Ce régime est assuré lorsque les $C_{g_i g_j} = \mathcal{O}(1)$ individuellement mais diffèrent entre eux par un facteur de l'ordre $\mathcal{O}(n^{-\frac{1}{2}})$. Sous ce régime, nous avons étudié dans [Tiomoko Ali and Couillet, 2016a, Tiomoko Ali and Couillet, 2018] les valeurs propres dominantes et les vecteurs propres associés (utilisés pour la classification) de $\mathbf{L}_\alpha$ pour des graphes denses de grandes dimensions suivant le MBSDC dans le régime non-trivial susmentionné. L'étude des valeurs propres nous permet de démontrer qu'il existe un *optimal* $\mathbf{L}_\alpha$ non trivial que l'on peut estimer à partir du graphe observé. Dans [Tiomoko Ali and Couillet, 2018], nous caractérisons le contenu asymptotique des vecteurs propres dominants de $\mathbf{L}_\alpha$ lequel nous a conduit à déterminer les valeurs limites des moyennes et covariances des centroïdes (correspondant à chaque communauté) qu'un algorithme Expectation Maximization (EM) aurait estimé après conver-

[4] À la lumière des précédentes méthodes de classification spectrale sur des graphes denses, la même analyse est valable pour q_i d'ordre aussi petit que $\frac{\log n}{n}$.

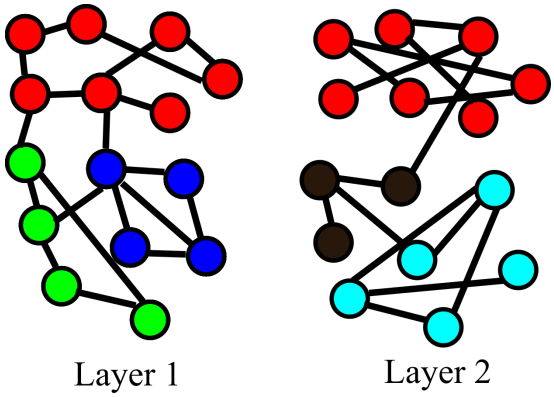


Figure 1.2: Heterogeneous multilayer network. Shared communities (in red) and unshared communities in different colors for each layer.

gence (en supposant que les entrées des vecteurs suivent une distribution d'un mélange fini de variables gaussiennes). Ces trouvailles sont alors utilisées en pratique pour initialiser l'algorithme EM en lieu et place de l'initialisation aléatoire qui est souvent utilisée.

1.1.2 Détection de communautés sur graphe multi-couches

L'extraction de structures de communautés partagées par les différentes couches ainsi que des structures distinctes entre les différentes couches pourraient être un bon moyen de comprendre les réseaux réels actuels qui sont souvent représentés sous forme de graphes multi-couches. Il existe donc un besoin pressant de nouvelles méthodes pour la détection de communautés dans les graphes multi-couches avec de telles structures de communautés hétérogènes. L'extraction des communautés indépendamment dans chaque couche (en utilisant des méthodes classiques de détection de communautés pour graphes à simple couche) est sous-optimale car cette approche n'exploite pas les informations communes à plusieurs couches. L'effort de recherche actuel vise à développer des méthodes d'inférence conjointe par agrégation des différentes couches [De Domenico et al., 2015a, Nicosia and Latora, 2015, De Bacco et al., 2017]. L'approche d'agrégation la plus simple consiste à réduire le réseau multi-couches en un réseau à simple couche sur lequel des méthodes classiques de détection de communautés peuvent être appliquées [Tang et al., 2009, Tang et al., 2012, Comar et al., 2012, Zhang et al., 2013, Amelio and Pizzuti, 2014, De Domenico et al., 2015b, Taylor et al., 2016, Kim et al., 2017]. Alternativement, certains chercheurs ont suggéré d'effectuer la détection de communautés séparément sur chaque couche suivie d'une agrégation par consensus des communautés trouvées à travers les différentes couches [Mucha et al., 2010, Xiang et al., 2012, Oselio et al., 2014, Amelio and Pizzuti, 2014, Paul and Chen, 2016]. Une autre approche consiste à étendre les MBS à une seule couche aux graphes multi-couches [Sweet et al., 2014, Han et al., 2015, Paul and Chen, 2015, Peixoto, 2015, Stanley et al., 2016, Valles-Catala et al., 2016, Reyes and Rodriguez, 2016, Barbillon et al., 2017], et utiliser des approches d'inférence statistique spécifiques à l'architecture multi-couches.

Comme indiqué ci-dessus, en pratique, certaines communautés peuvent être partagées entre les différentes couches, tandis que d'autres ne le sont peut-être pas (voir Figure 1.2). Cependant, peu de méthodes dans la littérature considèrent explicitement ce scénario général. L'algorithme d'extraction multi-couches proposé dans [Wilson et al., 2017] permet l'identification de communautés sur graphes multi-couches où les communautés pourraient être partagées entre un sous-ensemble de couches. [Wilson et al., 2017] minimise une fonction de coût et prend en compte les similitudes et les différences entre les communautés des différentes couches. Bien que le modèle utilisé dans [Wilson et al., 2017] soit réaliste lorsqu'on considère les connexions de graphes multi-couches, la méthode est seulement limitée aux graphes non-pondérés. L'approche proposée dans [Boden et al., 2012] étend [Zeng et al., 2006] à des graphes multi-couches pondérés et permet l'extraction de sous-graphes denses cohérents (appelés cliques) partagés par des sous-ensembles de couches. Cependant, ces dernières méthodes sont limitées à l'identification de communautés très denses et pourraient être inefficaces lorsque les graphes sont parcimonieux.

En collaboration avec l'Université du Michigan, nous avons proposé dans [Tiomoko Ali et al., 2018c], une méthode qui permet de détecter simultanément des communautés partagées et non partagées entre les différentes couches d'un graphe pondéré. Nous avons adopté une approche d'inférence statistique basée sur un modèle à blocs stochastiques où les arêtes du graphe peuvent être pondérés et qui tient compte du fait qu'une partie des communautés est partagée entre les différentes couches. En raison de la forme trop complexe de la distribution a posteriori des labels d'appartenance de chaque noeud à une communauté, nous avons utilisé une approche Bayes variationnelle pour approximer cette distribution a posteriori, puis nous utilisons les paramètres de cette distribution variationnelle pour déduire les communautés partagées et privées (à chaque couche) du modèle multi-couches utilisé. Cela généralise les travaux [Aicher et al., 2014, Zhang and Zhou, 2017] conçus pour des graphes pondérés à une seule couche.

1.2 Méthodes spectrales à base de fonctions à noyau sur des données à grandes dimensions

La *malédiction de la dimensionnalité* en AA augmente de façon exponentielle le besoin d'avoir une quantité importante de données d'apprentissage lorsque la dimension des données augmente afin d'éviter les phénomènes de sur-apprentissage. En outre, le processus d'apprentissage est très lent lorsque les données sont de très grandes dimensions. La solution la plus naturelle au problème de la malédiction de la dimensionnalité est de réduire la dimension de manière à conserver les informations les plus importantes dans les données, opération appelée *réduction de la dimensionnalité*. Par exemple, les pixels dans les images ont tendance à être fortement corrélés et ainsi, ne conserver qu'un résumé des pixels représentatifs réduit énormément la dimensionnalité des données. L' ACP (Analyse en Composantes Principales), de loin la technique de réduction de dimensionnalité la plus répandue, consiste à trouver les hyperplans synthétisant les caractéristiques les plus importantes des données, puis à projeter les données sur ceux-ci. Les hyperplans sont choisis de

manière à minimiser la moyenne quadratique de la différence entre les données originales et leurs projections sur les axes des hyperplans à identifier. Les axes des hyperplans sont appelés les *composantes principales* et correspondent aux plus grands vecteurs propres dans la décomposition en valeurs singulières (SVD) de la matrice d'affinité des données. Une fois que la dimensionnalité est réduite avec les informations les plus importantes conservées, les algorithmes d'AA classiques peuvent être appliqués avec moins de complexité et plus d'efficacité. Les approches de *clustering spectral* peuvent être considérées comme des variantes de la ACP et consistent à regrouper les données en utilisant seulement quelques composantes principales. Nous nous intéressons ici à l'étude du *clustering spectral* sur des *données à très grandes dimensions*, dont le comportement sera complètement différent de l'intuition originale du clustering spectral en petites dimensions.

Algorithm 1: Algorithme spectral

1: Calculer les ℓ vecteurs propres $\mathbf{u}_1, \ldots, \mathbf{u}_\ell \in \mathbb{R}^n$ correspondant aux valeurs propres les plus domiantes (plus petites ou plus grandes) de l'une des matrice d'affinité du réseau (adjacence, modularité, Laplacienne) de taille $n \times n$.
2: Empiler les vecteurs $\mathbf{u}_i$ en colonne dans une matrice $\mathbf{W} = [\mathbf{u}_i, \ldots, \mathbf{u}_\ell] \in \mathbb{R}^{n \times \ell}$.
3: Soit $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^\ell$ les lignes de $\mathbf{W}$. Classifier $\mathbf{r}_i \in \mathbb{R}^\ell$, $1 \leq i \leq n$ dans l'un des k groupes en utilisant un algorithme de classification en faible dimension (e.g., k-means [Hartigan and Wong, 1979] ou Expectation Maximization (EM) [Ng et al., 2012]). Le label assigné à $\mathbf{r}_i$ correspond au label du noeud i.

Nous considérons les algorithmes de clustering spectral (Algorithme 1) appliqués à n vecteurs de données disons $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$ que nous souhaitons assigner à des classes distinctes. Une approche de base consisterait à trouver des hyperplans (dans l'espace $\mathbb{R}^p$) séparant les différents vecteurs en classes distinctes. Cette méthode peut être considérée comme équivalente à l'application d'une analyse en composantes principales (ACP) ou d'une méthode spectrale sur la matrice d'affinité contenant les produits scalaires entre les données. Cependant, cette séparation n'est valide que lorsque les données sont linéairement séparables, ce qui n'est pas le cas dans la plupart des bases de données utilisées en AA. Pour faire face à cette difficulté, des méthodes à base de noyau ont été introduites et consistent à projeter les objets non linéairement séparables dans un espace de dimension plus élevée où les hyperplans peuvent être trouvés pour séparer linéairement les données. L'idée est donc de déplacer $\mathbf{x}_i$ vers $\Phi(\mathbf{x}_i)$ avec $\Phi : \mathbb{R}^p \to \mathbb{R}^P$ (avec $P \gg p$) et utiliser une méthode spectrale sur la matrice de covariance empirique dans l'espace $\mathbb{R}^P$. L'opération précédente étant évidemment très coûteuse pour P assez grand, l'astuce du noyau (kernel-trick) [Schölkopf, 2001] a été introduite pour considérer à la place la similarité $f(\mathbf{x}_i, \mathbf{x}_j)$ entre les vecteurs $\mathbf{x}_i$ et $\mathbf{x}_j$ où f est une fonction telle que $\Phi(\mathbf{x}_i)^\mathsf{T}\Phi(\mathbf{x}_j) = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ ou $\Phi(\mathbf{x}_i)^\mathsf{T}\Phi(\mathbf{x}_j) = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$. De telles fonctions f vérifiant certaines propriétés [Schölkopf, 2001] existent. Le clustering spectral peut ensuite être appliqué à la matrice de similarité à noyau, une méthode connue sous le nom de *clustering spectral à noyau*. Il est montré dans [Von Luxburg et al., 2008] que la classification spectrale utilisant des matrices d'affinité bien connues (matrice Laplacienne) entre un nombre de vecteurs aléatoires

$\mathbf{x}_i \in \mathbb{R}^p, 1 \leq i \leq n$ ($n \to \infty$ avec p fixe) est consistante dans le sens que les vecteurs propres dominants de ces matrices d'affinités convergent asymptotiquement vers des vecteurs limites comportant l'informant sur les classes. Par contre, comme nous le montrons dans cette thèse, le clustering de très grandes données ($n \to \infty, p \to \infty$ avec p/n constant) est moins trivial et d'importantes différences existent par rapport au clustering en petites dimensions .

L'intuition originelle de [Ng et al., 2002] appuyant le clustering spectral n'est plus valide pour des données à grandes dimensions. Pour voir cela, commençons par décrire l'intuition de [Ng et al., 2002] sur le clustering spectral en petites dimensions. Considérons n vecteurs de données $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ (p fixe) que nous souhaitons classifier en 2 classes distinctes en utilisant une fonction à noyau f de telle sorte que l'affinité $K_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$ entre vecteurs $\mathbf{x}_i$ et $\mathbf{x}_j$ est grande lorsque les vecteurs appartiennent à la même classe et est faible lorsqu'ils appartiennent à des classes distinctes. En supposant sans perte de généralité que les vecteurs $\mathbf{x}_i$ sont ordonnés par classes c'est-à-dire, $\mathbf{x}_1, \cdots, \mathbf{x}_{n/2}$ constituent la première classe tandis que $\mathbf{x}_{n/2+1}, \cdots, \mathbf{x}_n$ forment la deuxième classe, nous notons que les vecteurs $[\mathbf{1}_{\frac{n}{2}}, \mathbf{0}_{\frac{n}{2}}]$ et $[\mathbf{0}_{\frac{n}{2}}, \mathbf{1}_{\frac{n}{2}}]$ que nous appelons *vecteurs canoniques des classes* sont les vecteurs propres de la matrice Laplacienne $\mathbf{L} = \mathbf{D} - \mathbf{K}$ (avec $\mathbf{D} = \text{diag}(\left\{\sum_{j=1}^n K_{ij}\right\}_{i=1}^n)$) associés à la plus petite valeur propre 0, et donc un algorithme de classification spectrale utilisant ces deux vecteurs propres attribuera sans erreur la bonne classe à chaque donnée. Dans des situations presque idéales, $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$ sera relativement importante pour les données de la même classe et donc un algorithme spectral utilisant les vecteurs propres associés aux plus petites valeurs propres de $\mathbf{L}$ sera en mesure de récupérer les classes avec une performance raisonnable. Ce raisonnement n'est plus valide dans le régime "big data" où la dimension p peut être très grande et comparable au nombre n de données. Pour illustrer ce fait, supposons que les vecteurs $\mathbf{x}_i$ sont des vecteurs Gaussiens indépendants ayant la même covariance $\mathbf{I}_p$ mais avec une moyenne $\boldsymbol{\mu}_1$ lorsqu'ils appartiennent à la première classe et une moyenne $\boldsymbol{\mu}_2$ lorsqu'ils appartiennent à la deuxième classe. Nous avons donc $\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{p} \simeq \frac{\boldsymbol{\mu}_1^\mathsf{T}\boldsymbol{\mu}_2}{p}$ quand $\mathbf{x}_1 \neq \mathbf{x}_2$ tandis que $\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_i}{p} \simeq \frac{\|\boldsymbol{\mu}_1\|^2}{p} + 1$ quand $\mathbf{x}_1 = \mathbf{x}_2$. Comme nous le verrons plus tard, les moyennes de classe doivent satisfaire $\|\boldsymbol{\mu}_a\| = \mathcal{O}(1)$ (par rapport à p) afin d'éviter: **i)** que les vecteurs soient asymptotiquement très lointains de telle sorte que leur classification devient triviale ou **ii)** les vecteurs sont très proches de telle sorte que la classification est impossible. Donc, d'après $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = \mathcal{O}(1)$, l'affinité K_{ij} (pour $i \neq j$) converge asymptotiquement vers la même valeur $f(0)$ peu importe les classes auxquelles appartiennent $\mathbf{x}_i$ et $\mathbf{x}_j$. Il ne semble donc pas possible de récupérer les classes en utilisant la procédure de l'algorithme spectral décrite au-dessus car la notion de "proximité" entre les données de la même classe n'est plus valide en grandes dimensions. Cependant, il s'avère que l'algorithme spectral fonctionne quand même bien comme on peut le voir sur la figure 1.3 où le vecteur propre dominant de $\mathbf{K}$ est composé de "plateaux" bruités avec chaque plateau représentant une classe et donc un k-means serait en mesure de récupérer l'information sur les classes en utilisant le vecteur propre dominant, avec des erreurs non triviales. Il est alors essentiel de comprendre les raisons pour lesquelles les méthodes spectrales à base de noyau en grandes dimensions marchent bien malgré le fait que cette notion de "proximité" des données de même classe (au coeur de la classification

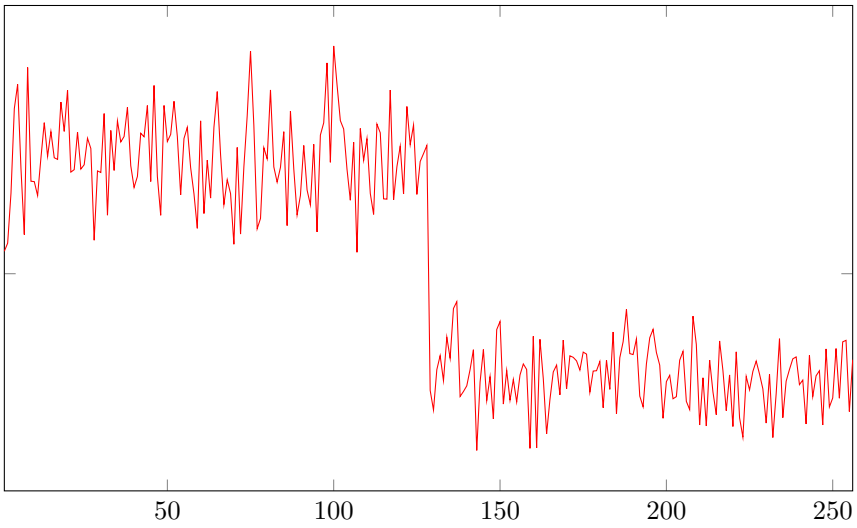


Figure 1.3: $p = 512$, $n = 256$. Dominant eigenvector (associated with the largest eigenvalue) of $K_{ij} = f\left(\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{p}\right)$, $f(x) = x^2$, 2 balanced classes $\mathcal{C}_1, \mathcal{C}_2$, $x_i \in \mathcal{C}_1 \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}\left(3/\sqrt{p}\mathbf{1}_p, (1+3/\sqrt{p})\mathbf{I}_p\right)$ and $\mathbf{x}_i \in \mathcal{C}_2 \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}\left(-3/\sqrt{p}\mathbf{1}_p, (1-3/\sqrt{p})\mathbf{I}_p\right)$.

spectrale) n'est plus valide.

Afin de comprendre les différents mécanismes du clustering spectral en grandes dimensions, il est important de comprendre la structure des valeurs propres et vecteurs propres des grandes matrices d'affinité aléatoire structurées $\mathbf{K}$. Il ressort de la structure de $K_{ij} = f\left(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j\right)$ que $\mathbf{K}$ a des entrées non linéaires et des colonnes dépendantes, une structure peu commune en théorie classique des matrices aléatoires. Les premiers travaux de [El Karoui et al., 2010] ont montré que lorsque des vecteurs de grandes dimensions $\mathbf{x}_i$ sont gaussiens avec une moyenne nulle et une covariance $\mathbf{C}$ (aucune supposition de classe ici), la matrice $\mathbf{K} = \left\{f\left(\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{p}\right)\right\}_{i,j=1}^n$ est asymptotiquement équivalente à une matrice aléatoire du type "matrice de covariance empirique perturbée". Plus précisément, il a été montré dans [El Karoui et al., 2010] que

$$\|\mathbf{K} - \hat{\mathbf{K}}\| \to 0$$

en probabilité où

$$\hat{\mathbf{K}} = f'(0)\mathbf{X}\mathbf{X}^\mathsf{T} + \left(f(0) + f''(0)\frac{\operatorname{tr}\mathbf{C}^2}{2p^2}\right)\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \left(f\left(\frac{\operatorname{tr}\mathbf{C}}{p}\right) - f(0) - f'(0)\frac{\operatorname{tr}\mathbf{C}}{p}\right)\mathbf{I}_n$$

avec $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. La matrice $\hat{\mathbf{K}}$ a une structure simple d'un point de vue "matrice aléatoire" : elle est essentiellement composée de la matrice de covariance empirique

$f'(0)\mathbf{X}\mathbf{X}^\mathsf{T}$ ajoutée à la matrice de rang faible $\mathbf{1}\mathbf{1}^\mathsf{T}$ et à la matrice identité. L'étude des valeurs propres et vecteurs propres de $\mathbf{K}$ peut donc être réalisée comme dans la RMT classique. En s'appuyant sur les travaux de [El Karoui et al., 2010], [Couillet and Benaych-Georges, 2016] généralise l'étude de $\mathbf{K} = \{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2)\}_{i,j=1}^n$ aux données $\mathbf{x}_i$ issues d'un modèle de mélange gaussien (GMM) avec k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ de telle sorte que $\mathbf{x}_i$ soit Gaussien de moyenne $\boldsymbol{\mu}_a$ et de covariance $\mathbf{C}_a$ lorsqu'il appartient à la classe $\mathcal{C}_a$. Dans le régime big data où la dimension p grandit linéairement avec la taille n des données ($n, p \to \infty$ with $n/p = \mathcal{O}(1)$), [Couillet and Benaych-Georges, 2016] montre que $\mathbf{K}$ est asymptotiquement égale à la somme d'une "matrice de covariance empirique déformée" avec des matrices de perturbation (de rang fini) contenant les informations sur les classes (moyennes, covariances) et donc, se comporte asymptotiquement comme un modèle de matrices aléatoires connu dit "spike". Ce type de structure matricielle permet d'effectuer une étude approfondie des valeurs propres et vecteurs propres de $\mathbf{K}$ dans ce régime à grandes dimensions. L'étude de cette matrice aléatoire révèle en particulier que le choix de la fonction noyau f a un fort impact sur la discrimination des données en fonction des statistiques (moyennes, covariances) des classes auxquelles elles appartiennent.

Dans cette thèse, nous considérons également des données en grandes dimensions issues d'un mélange de k distributions gaussiennes avec moyennes $\boldsymbol{\mu}_a$ et covariances $\mathbf{C}_a$ ($a = 1, \cdots, k$). Dans ce régime, en supposant un scénario supervisé où les moyennes et covariances de classes sont connues et l'objectif est d'estimer les classes de chaque donnée, les auteurs dans [Couillet et al., 2018] ont déterminé les régimes de croissance des distances minimales entre les moyennes de classes et entre les covariances de classes, requis pour asymptotiquement arriver à faire une classification non-triviale (c'est-à-dire ni parfaite ni impossible). Ces régimes minimaux de croissance que nous appellerons par la suite *taux optimaux oracle* sont donnés ci-dessous dans le cas $k = 2$ et constituent donc une référence de comparaison de différentes méthodes de classification (semi-supervisée ou non-supervisée).

- $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = \mathcal{O}(1)$.
- $\|\mathbf{C}_1 - \mathbf{C}_2\| = \mathcal{O}\left(\frac{1}{\sqrt{p}}\right)$.
- $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2) = \mathcal{O}(\sqrt{p})$.
- $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(1)$.

L'objectif de cette partie de la thèse est de proposer des méthodes de clustering spectral (non supervisées) à base de noyaux capables de discriminer les données avec des taux de distance proches des *taux optimaux oracle*[5]. Pour ce faire, nous commençons par analyser la matrice aléatoire à noyau $\mathbf{K} = f\left(\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{p}\right)$ avec f une fonction noyau générique supposée être au moins 3 fois dérivable autour de 0 (car $\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{p} \to 0$ dans le régime que l'on

[5]Le terme *oracle* est utilisé pour souligner le fait que ces taux optimaux sont obtenus dans un cas supervisé où les moyennes et covariances de classe sont parfaitement connues

considère). Cette étude révèle qu'estimer les labels de classes par le clustering spectral utilisant $\mathbf{K}$ ne fait pas mieux qu'une estimation aléatoire lorsqu'une fonction générique f est utilisée sur des données à moyennes égales et à covariances de classes telles que $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 \ll \mathcal{O}(p)$, on est donc loin des taux optimaux oracle. En utilisant à la place une fonction à noyau f telle que $f'(0) = 0$, on peut mieux faire qu'une estimation aléatoire quand $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(\sqrt{p})$ induisant ainsi de meilleures performances que des fonctions f génériques. Cependant, cette dernière fonction à noyau présente le côté négatif d'annuler l'effet des moyennes sur la classification, c'est-à-dire que l'utilisation de ce noyau ne sera pas capable de différencier des données qui ont les mêmes covariances mais des moyennes fortement différentes. Ce cas est étudié en détails dans [Kammoun and Couillet, 2017] pour des noyaux appliqués au produit scalaire entre les données $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ avec la fonction noyau f telle que $f'(0) = 0$ (0 étant la valeur limite de $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p$). Dans cette thèse, nous montrons qu'une analyse minutieuse de la matrice aléatoire à noyau $\mathbf{K}$ révèle que $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$ au lieu de $f'(0) = 0$ permet un traitement équitable entre les moyennes de classe et les covariances de classe lors de classification des données et est capable de les discriminer avec la condition $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(\sqrt{p})$ comme pour le noyau $f'(0) = 0$. Cette nouvelle matrice à noyau qui prend en compte l'équilibre entre les moyennes et les covariances des données est capable de discriminer les données dans des meilleurs régimes de croissance par rapport aux noyaux précédemment étudiés. De tels choix de noyau sont importants en pratique puisque des choix spécifiques de f sont plus adaptés aux ensembles de données contenant soit des différences prononcées entre les moyennes de classe comme dans le jeu de données populaire MNIST [LeCun, 1998] alors que d'autres fonctions sont plus efficaces sur des jeux de données avec des classes ayant des moyennes similaires mais, des covariances fortement différentes [Andrzejak et al., 2001].

1.3 Plan et contributions

Travaux liés au présent manuscrit

Comme indiqué dans les sections précédentes, cette thèse couvre deux applications principales de clustering: *détection de communautés dans les graphes* et *clustering (à base de noyau) de données à grandes dimensions*. Comme première contribution principale concernant la détection de communautés dans les graphes, en partant d'une étude complète des matrices de similarités, nous avons développé un algorithme spectral amélioré pour des modèles aléatoires de grands graphes *à simple couche, denses et hétérogènes* et qui généralise des méthodes spectrales existantes. La deuxième contribution principale concerne la détection de communautés dans les graphes *multi-couches* où nous avons développé un algorithme qui permet de déterminer les communautés de noeuds entre les différentes couches. En ce qui concerne le *clustering à base de noyau*, à travers une étude complète de la matrice d'affinité (produit scalaire) aléatoire utilisant des fonctions à noyau, nous avons proposé des choix de noyaux appropriés pour discriminer les données suivant leurs statistiques les plus dominantes.

Dans le Chapitre 3, nous présentons les outils mathématiques nécessaires pour suivre les différents résultats de cette thèse. Le levier principal des Chapitres 4 et 6 est l'analyse spectrale (valeurs propres et vecteurs propres) de grandes matrices aléatoires tandis que dans le Chapitre 5, nous utilisons une méthode Bayes variationnelle pour établir un algorithme de détection de communautés pour graphes à plusieurs couches. Ces deux outils mathématiques complètement différents (théorie des matrices aléatoires et inférence variationnelle) sont introduits afin de constituer une base pour les chapitres suivants.

Dans le Chapitre 4, nous considérons l'analyse des méthodes spectrales pour la détection de communautés dans des réseaux denses qui peuvent avoir une distribution de degrés hétérogène. Nous considérons une forme généralisée $\mathbf{L}_\alpha$ (paramétrée par α) des matrices de similarité les plus utilisées pour la classification spectrale sur les graphes denses, sous le modèle de graphe statistique réaliste le MBSDC. Les performances de ces méthodes spectrales dépendant de la position des valeurs propres de $\mathbf{L}_\alpha$ ainsi que du contenu des vecteurs propres correspondants, nous étudions les valeurs propres et vecteurs propres de $\mathbf{L}_\alpha$. La matrice $\mathbf{L}_\alpha$ n'étant pas une matrice aléatoire présentant visuellement la structure de communautés du graphe, une première étape consiste à approximer $\mathbf{L}_\alpha$ (pour un nombre de nœuds $n \to \infty$) par une matrice aléatoire théoriquement analysable $\tilde{\mathbf{L}}_\alpha$ qui appartient à la famille des modèles matriciels aléatoires "spike" et qui permet une étude approfondie des valeurs propres et des vecteurs propres de $\mathbf{L}_\alpha$. Les matrices aléatoires "spike" (qui seront introduites dans le Chapitre 3) présentent généralement une transition de phase au-delà de laquelle les informations utiles peuvent être extraites des vecteurs propres associés aux valeurs propres dominantes (et en dessous de laquelle aucune information utile ne peut être tirée). Dans notre contexte, cette transition de phase correspond à un *seuil de détectabilité des communautés*, commun dans l'analyse des algorithmes de détection de communautés. Nous caractérisons exactement cette transition de phase pour chaque valeur de α . Nous prouvons l'existence et obtenons une expression pour une valeur "optimale" α_{opt} de α pour laquelle le seuil de détectabilité des communautés [6] est facilement atteignable. Cette valeur n'est pas toujours 0 ou 1 et son bon choix est très important pour la classification de graphes très hétérogènes. Nous établissons un estimateur consistant $\hat{\alpha}_{\text{opt}}$ de α_{opt} en fonction des degrés $\mathbf{d}$. Nous montrons que pour obtenir un clustering consistant dans le modèle MBSDC, les vecteurs propres dominants utilisés pour le clustering doivent être pré-multipliés par $\mathbf{D}^{\alpha-1}$ avant la classification, ce qui permet ainsi de retrouver l'algorithme SCORE [Jin et al., 2015] pour $\alpha = 0$ et l'algorithme dans [Gulikers et al., 2015] pour $\alpha = 1$ comme cas spéciaux. Une étude approfondie des vecteurs propres régularisés nous permet d'améliorer l'initialisation de l'algorithme EM (dans la dernière étape de l'algorithme spectral décrit ci-dessus) en comparaison à une initialisation aléatoire. Des simulations numériques montrent que notre méthode fait mieux que des approches de l'état de l'art à la fois sur des données simulées et sur des données réelles. Ces travaux ont commencé avec l'article ci-après où nous avons procédé à l'étude spectrale de la matrice $\mathbf{L}_1 \propto \mathbf{D}^{-1}\left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\intercal}{\mathbf{d}^\intercal \mathbf{1}_n}\mathbf{D}^{-1}\right]$ dans les modèles MBSDC de graphes denses à grandes dimensions. Une étude des valeurs propres et vecteurs propres

[6] Le seuil de détectabilité des communautés est le point au-delà duquel il existe un algorithme d'estimation des communautés qui peut faire mieux qu'une estimation aléatoire.

de $\mathbf{L}_1$ nous a permis de déterminer le seuil de transition de phase aussi bien que le taux asymptotique d'erreur de classification d'un algorithme spectral utilisant $\mathbf{L}_1$.

[[Tiomoko Ali and Couillet, 2016b]] Tiomoko Ali, H. and Couillet, R. (2016). Performance analysis of spectral community detection in realistic graph models. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)

L'étude des matrices généralisées $\mathbf{L}_\alpha$ suivie de la caractérisation de la transition de phase et de l'estimation du α *optimal* a été effectuée dans

[[Tiomoko Ali and Couillet, 2016a]] Tiomoko Ali, H. and Couillet, R. (2016a). community detection in heterogeneous networks. In Signals, Systems and Computers, 2016 50th Asilomar Conference on, pages 1385–1389. IEEE.

L'article suivant publié dans "Journal of Machine Learning Research" englobe tous les travaux susmentionnés avec les preuves des différents résultats. Par ailleurs, l'étude des vecteurs propres a aussi été faite suivie par l'amélioration de l'algorithme EM utilisé dans la dernière étape de la procédure de classification spectrale.

[[Tiomoko Ali and Couillet, 2018]] Tiomoko Ali, H. and Couillet, R. (2018). Improved spectral community detection in large heterogeneous networks. Journal of Machine Learning Research, 18:1–49.

Dans le Chapitre 5, nous considérons le problème de détection de communautés sur des graphes à plusieurs couches. Nous proposons une nouvelle méthode qui permet de détecter simultanément des communautés partagées et non partagées entre les différentes couches du graphe. Nous définissons un MBS étendu à des graphes multi-couches avec des arêtes pondérées et une structure de communautés hétérogène. En raison de la forme non explicite de la distribution a posteriori des variables latentes du modèle, nous établissons un algorithme Bayes variationnel pour approximer cette distribution a posteriori qui nous permet d'estimer les paramètres de communautés voulues. Nous montrons que l'algorithme proposé est plus précis et robuste que de précédentes approches pour la détection de communautés dans les graphes à plusieurs couches. Nous illustrons enfin notre méthode sur une base de données génomique permettant la découverte de structures hétérogènes dans les différents types d'interaction entre les gènes. Ce travail, en collaboration avec l'Université du Michigan, a été soumis à la conférence NIPS 2018

[[Tiomoko Ali et al., 2018c]] Ali, H. T., Liu, S., Yilmaz, Y., Hero, A., Couillet, R., and Rajapakse, I. (2018b). Latent heterogeneous multilayer community detection.

Dans le Chapitre 6, nous effectuons une étude des matrices d'affinité (produit scalaire) aléatoires à noyau en grandes dimensions sous une hypothèse de données provenant d'un modèle de mélange gaussien (GMM) avec des taux de croissance des moyennes et covariances de classes fixés de manière à induire une classification non triviale de ces données (c'est-à-dire ni parfaite, ni impossible). La première étude avec une fonction à noyau générique f a été effectuée dans

[[Tiomoko Ali et al., 2018a]] Tiomoko Ali, H., Kammoun, A., and Couillet, R. (2018a). Random matrix asymptotic of inner product kernel spectral clustering. In Inter-

national Conference on Acoustics, Speech and Signal Processing (ICASSP'18). IEEE.

Dans le travail précédent, on remarque que les fonctions génériques à noyau f sont sous-optimales par rapport aux taux de discrimination de référence. On observe aussi que prendre des fonctions à noyau telles que $f'(0) = 0$ permet d'améliorer la condition de $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(p)$ à $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(\sqrt{p})$ mais présente l'effet négatif de complètement ignorer l'effet des moyennes de classes. Dans le travail suivant [Tiomoko Ali et al., 2018b], nous proposons une nouvelle famille de fonctions à noyau qui permettent de mieux exploiter les statistiques des données.

[[Tiomoko Ali et al., 2018b]] Tiomoko Ali, H., Kammoun, A., and Couillet, R. (2018b). Random matrix-improved kernels for large dimensional spectral clustering. In Statistical Signal Processing Workshop (SSP), 2016 IEEE, pages 1–4. IEEE.

Autres publications

Au cours de cette thèse, à travers une série de travaux, nous avons effectué une analyse quantitative des performances de réseaux de neurones dits "echo-state" (ESN). Bien que les ESN ne soient pas vraiment utilisés en pratique, les outils théoriques que nous avons développés pour analyser leurs performances permettent d'avoir des intuitions sur les fonctionnalités de mémoire des réseaux de neurones dits "récurrents" (RNN). Plus précisément, nous avons établi une première analyse théorique de l'*erreur quadratique moyenne* des phases d'apprentissage et de test, dans le régime à grandes dimensions où le nombre de neurones et la durée de l'apprentissage sont du même ordre de grandeur. Nous nous sommes appuyés sur des effets de concentrations de matrices aléatoires pour trouver des équivalents déterministes aux mesures de performances (erreur quadratique moyenne). Ceux travaux ont été développés dans les publications ci-après

[[Couillet et al., 2016b]] Couillet, R., Wainrib, G., Sevi, H., and Tiomoko Ali, H. (2016c). Training performance of echo state neural networks. In Statistical Signal Processing Workshop (SSP), 2016 IEEE, pages 1–4. IEEE.

[[Couillet et al., 2016c]] Couillet, R., Wainrib, G., Tiomoko Ali, H., and Sevi, H. (2016d). A random matrix approach to echo-state neural networks. In International Conference on Machine Learning, pages 517–525.

[[Couillet et al., 2016a]] Couillet, R., Wainrib, G., Sevi, H., and Tiomoko Ali, H. (2016b). The asymptotic performance of linear echo state neural networks. Journal of Machine Learning Research, 17(178):1–35.

Vu que ces travaux ne sont pas directement liés au sujet de ce manuscrit, nous ne les détaillerons pas plus dans la suite.

Chapter 2

Introduction

Machine learning (ML) is probably the most popular field nowadays in technological sciences reaching both the industrial and academic sectors. The power of ML lies in its capability of automatically processing any kind of data which can be texts, numerics, images, sounds, videos, etc., from which practically relevant data abstractions can be retrieved. The huge amount of data arising from different industrial fields calls for the development of innovative and efficient automated algorithms for their processing. The general idea behind ML is to learn a *blackbox fonction* $\mathbf{y} = f(\mathbf{x})$ mapping some inputs $\mathbf{x}$ to some outputs $\mathbf{y}$ by using a set of training examples $\mathbf{x}_i$'s for which the outputs $\mathbf{y}_i$'s are either known (supervised learning) or unknown (unsupervised learning). The learning problem is known as *classification* when the set of outputs $\mathbf{y}$ is discrete and as *regression* when the outputs are real-valued. In the case of supervised learning, the parameters of the function are learned and then used to predict the output of new data (test set) while in the unsupervised learning case, the input features are directly used to categorize the inputs. *Clustering*, one of the most popular unsupervised learning approaches, is the main subject of this thesis.

In general, clustering objects consists in putting them into groups in such a way that the objects are very similar inside each group. The objects of interest can either be sets of data (which can be seen as vectors) or graphs (representing interactions between a set of nodes). Data vectors are the objects of interest in the so-called *data clustering* [Shalev-Shwartz and Ben-David, 2014] which consists in a grouping based on the most similar objects where similarity values (for each pair of data) are computed using a "proximity measure". The clustering problem is known as *community detection* or *graph clustering* [Newman, 2010, Goldenberg et al., 2010, Fortunato, 2010] when sets of interactions (weighted or not) between pairs of nodes are provided and the task consists in finding groups of nodes with the most compact interactions. In most machine learning datasets, the vectors are usually not linearly separable and instead of using pair-wise similarities in the input space, the data are first projected into a feature space (where the data become linearly separable) before pair-wise similarity computation. Due to the kernel-trick [Schölkopf, 2001], the computation of the aforementioned pair-wise similarity in the feature space (usually expensive) can be performed in the input space by applying an

appropriate kernel function on the data similarity in the input space, and the clustering task is known as *kernel-based clustering*. Obviously, *data clustering* or *kernel spectral clustering* can be seen as *community detection* tasks on weighted graphs.

2.1 Community detection in graphs

Community detection in graphs (networks) is one of the most fundamental problems in data mining as it enables to explore and analyze graphs representing real-world interactions in many fields including but not limited to sociology [Goldenberg et al., 2010], biology and medecine [Chen and Yuan, 2006, Marcotte et al., 1999, Cline et al., 2007], transportation [Guimera et al., 2005], Internet of Things networks [Linden et al., 2003, Clauset et al., 2004]. A graph $\mathcal{G}$ is defined as a pair of sets $(\mathcal{V}, \mathcal{E})$ with $\mathcal{V}$ denoting the set of nodes and $\mathcal{E}$ is a subset of all pairwise interactions $\mathcal{V} \times \mathcal{V}$. The graph is said to be *directed* when the elements of the set $\mathcal{E}$ are ordered and *undirected* otherwise. In many applications, real value weights can be assigned to each element of $\mathcal{E}$ indicating the strength of the connection resulting in a *weighted graph* while binary weights correspond to an *unweighted graph*. The nodes of the graph might either be involved in different types of interactions (e.g., actors involved in different social networks), or the interactions might evolve over time (dynamic graphs), leading to the so-called *multi-layer graph*. The terminology *single-layer graph* will be employed when only one type of interactions characterizes the graph.

For a given graph, community detection consists in extracting hidden or latent communities of nodes where there are many edges inside the communities and few edges across them. Research in single-layer community detection has and continues to be a very active field both from theoretical and algorithmic aspects. In current real-world networks, individuals might be involved in different types of interactions thus leading to multi-layer/multi-relational/multiplex/multi-dimensional networks/graphs. For example, large company employees may be connected to each other based on their similar activities on the one hand and based on their social activities on the other hand [Oselio et al., 2015]. In genomics, the genes might be related either by their transcriptional interactions (function relations), e.g., measured by RNA-seq profile similarity, or by chromatin interactions (spatial relations), e.g., measured by chromatin conformation capture (Hi-C) of promter-enhancer ligations [Dixon et al., 2015, Boulos et al., 2017, Dekker et al., 2017]. The recent growth of those dynamic networks with heterogeneous interactions has thus called for the development of new community detection methods for multi-layer graphs. Two main classes of methods are generally used to solve the community detection problem (for single or multi-layer graphs): *statistical inference* and *metric-based* optimization. The community detection problem described in Section 2.1.1 is specific to single-layer graphs where the graph nodes are only involved in one type of interaction.

2.1.1 Single-layer graph community detection

As stated above, statistical inference methods are one of the big classes of approaches to community detection. They consist in fitting the observed graph to a statistical model taking into account the latent community structure and then inferring the parameters of the model (among which the assignment of the nodes to the communities). The most basic statistical model with community structure which is often used for inference is the so-called Stochastic Block Model (SBM). Denoting $\mathcal{G}$ a k-class graph of n vertices with communities $\mathcal{C}_1, \ldots, \mathcal{C}_k$ with g_i the group assignment of node i, the SBM assumes an adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ with A_{ij} independent Bernoulli random variables with parameter $P_{g_i g_j}$ where P_{ab} represents the probability that any node of class $\mathcal{C}_a$ is connected to any node of class $\mathcal{C}_b$. The main limitation of this model is that it is only suited to homogeneous graphs where all nodes have the same average degree in each community (besides, class sizes are often taken equal). A more realistic model, the Degree-Corrected SBM (DCSBM), was proposed in [Coja-Oghlan and Lanka, 2009, Karrer and Newman, 2011] to account for degree heterogeneity inside communities. For the same graph $\mathcal{G}$ defined above, by letting q_i, $1 \leq i \leq n$, be some intrinsic weights which affect the probability for node i to connect to any other network node, the adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ of the graph generated by the DCSBM is such that A_{ij} are independent Bernoulli random variables with parameter $q_i q_j C_{g_i g_j}$, where $C_{g_i g_j}$ is a class-wise correction factor.

The second class of community detection methods, *metric-based*, consist in first defining a metric consistent with the formal definition of communities (e.g., modularity [Newman, 2006b], ratio-cut[Wei and Cheng, 1989]) and then maximizing the chosen metric over all possible partitionings of the network. Formulated as such, the community detection problem is proven to be NP-hard [Brandes et al., 2007]. The literature in single-layer community detection has been focused on finding polynomial-time approximation methods to the original problem by proposing greedy approaches [Newman, 2004], simulated annealing [Guimera et al., 2004], extremal optimization [Duch and Arenas, 2005] and spectral methods [Ng et al., 2002, Newman, 2006b, Newman, 2016]. The relaxation of modularity or ratio-cut optimization from discrete community memberships to continuous values leads to spectral methods (summarized in Algorithm 2 in Section 2.2) which consist in retrieving the nodes' communities from the eigenvectors associated with the dominant eigenvalues of similarity matrices (adjacency matrix, modularity matrix, Laplacian matrix) representing the graph. The similarity matrix in use for spectral methods depends on the metric under consideration; for example, the optimization of the modularity metric leads to a spectral method using the modularity matrix, the ratio-cut metric corresponds to the adjacency matrix while optimizing the normalized ratio-cut induces the use of the normalized Laplacian matrix. The community detectability threshold of spectral methods using the adjacency matrix (and its variants) is shown [Nadakuditi and Newman, 2012] to match the Bayes optimal threshold on *dense random graphs* (such that the average degree grows with the size of the graph) generated according to the Stochastic Block Model (SBM). However, there is an important gap between the community detectability threshold of those spectral methods [Kawamoto and Kabashima, 2015] and the Bayes optimal threshold [Decelle et al., 2011b, Decelle et al., 2011a] in the sparse

regime (i.e., when the maximum nodes' degree does not grow with the number of nodes) and thus showing the sub-optimality of those spectral methods in the sparse regime [Krzakala et al., 2013]. In the simple "sparse" SBM case with 2 communities of same sizes, the positive part of the phase transition conjecture [Decelle et al., 2011a] is proven in [Mossel et al., 2013, Massoulié, 2014] while the negative part is proven in [Mossel et al., 2015]. It was conjectured that spectral method using the Non-backtracking [Krzakala et al., 2013] and Bethe Hessian [Saade et al., 2014] matrices fill the gap with the Bayes optimal in the sparse regime. A study of the leading eigenvalues and eigenvectors of the Non-backtracking matrix in [Bordenave et al., 2015] led to prove the conjecture in [Krzakala et al., 2013]. However, the convergence of the eigenvalue distribution of the Non-backtracking matrix (a non-symmetric matrix) of random SBM graphs is still an open problem. From a random matrix theory perspective, the reason behind the failure of spectral methods using variants of the adjacency matrix in the sparse regime, lies in the fact that the spectrum (eigenvalue distribution) of those matrices does not concentrate; there might be some outlying eigenvalues, the eigenvectors of which are localized around few vertices (called *hubs*) and thus the global community structure is lost in those eigenvectors which are normally used in spectral clustering. Getting back to the "dense" regime (where the spectrum of the *adjacency family of matrices* is well behaved), [Nadakuditi and Newman, 2012] study the spectrum of the modularity matrix in a symmetric SBM, which led to the characterization of the detectability threshold which matches the Bayes optimal one. Through an explicit characterization of the dominant eigenvector in this simpler regime and by conjecturing on the Gaussianity of the eigenvector entries, [Nadakuditi and Newman, 2012] derive the asymptotic misclassification rate. Although the analysis of [Nadakuditi and Newman, 2012] is only based on the modularity matrix, the performances (and also the detectability threshold) of spectral methods using all matrices in the *adjacency family* are asymptotically the same for graphs generated according to the SBM. However, as we will show in the course of this thesis, the performances of the different spectral methods might be different in the dense DCSBM regime.

Community detection in DCSBMs has recently been studied, providing "consistent"[1] algorithms ranging from modularity/likelihood-based approaches to spectral clustering methods. Sufficient conditions under which likelihood-based approaches [Karrer and Newman, 2011] and modularity optimization methods [Newman, 2006b] are weakly and strongly consistent, have been provided in [Zhao et al., 2012]. The so-called CMM (Convexified Modularity Maximization) algorithm was proposed in [Chen et al., 2015b] to cope with the computational expensiveness of modularity/likelihood methods [Karrer and Newman, 2011,Newman, 2006b] by solving a convex programming relaxation of the modularity optimization. Asymptotic minimax risks for misclassification loss under the DCSBM have been established in [Gao et al., 2016]. There a consistent algorithm achieving the minimax optimal rates was derived, which is similar to spectral methods but proceeds without the explicit computation of eigenvectors and is hence computationally less expensive. As far

[1]Consistency is mainly defined in two forms. Informally, a community detection algorithm is **weakly** consistent whenever the fraction of misclassified nodes vanishes asymptotically with high probability and a community detection algorithm is **strongly** consistent whenever the labels estimated by the algorithm match exactly the true labelling asymptotically with high probability.

as spectral clustering methods are concerned, [Lyzinski et al., 2014] and [Lei et al., 2015] show the consistency of the classical spectral clustering procedure for community detection applied to the adjacency matrix of moderately sparse DCSBM (for not too irregular degree distributions) where the expected degree is as small as $\log n$. Later, it has been shown [Coja-Oghlan and Lanka, 2009, Qin and Rohe, 2013, Jin et al., 2015, Gulikers et al., 2015] that when the degrees are highly heterogeneous, the classical spectral methods fail to detect the genuine communities. To illustrate the limitations of spectral methods under the DCSBM, the two graphs of Figure 1.1 provide 2D representations of dominant eigenvector 1 versus eigenvector 2 for the standard modularity matrix and the Bethe Hessian matrix[2], when three quarters of the nodes connect with low weight $q_{(1)}$ and one quarter of the nodes with high weight $q_{(2)}$. For both methods, it is clear that k-means or EM alike would erroneously induce the detection of extra communities and even a confusion of genuine communities in the Bethe Hessian approach. Those extra communities are produced by some biases created by the intrinsic weights q_i's; intuitively, nodes sharing the same intrinsic connection weights tend to create their own sub-cluster inside each community, thereby forming additional sub-communities inside the genuine communities. To overcome this issue, a number of regularized spectral clustering techniques have been proposed to normalize either the adjacency matrix or the leading eigenvectors by the degrees. In [Coja-Oghlan and Lanka, 2009, Gulikers et al., 2015], the authors have proposed to cluster the nodes based on the eigenvectors of a normalized adjacency matrix $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}$ with $\mathbf{D}$ the diagonal matrix containing the observed degrees on the main diagonal. We precisely perform in [Tiomoko Ali and Couillet, 2016b] a spectral analysis of the normalized modularity matrix $\mathbf{D}^{-1}\left(\mathbf{A}-\frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{\mathbf{d}^\mathsf{T}\mathbf{1}_n}\right)\mathbf{D}^{-1}$ (which is not different from $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}$ as far as spectral clustering is concerned) in a dense DCSBM regime where classification is not asymptotically trivial. We derive there the phase transition where classification becomes asymptotically possible and establish in simple toy random graph models the asymptotic misclassification rate. Another approach to overcoming the biases induced in the DCSBM consists instead in using the leading eigenvectors (pre-normalized by the inverse degree matrix $\mathbf{D}^{-1}$) of the adjacency matrix, which is proposed as the SCORE algorithm in [Jin et al., 2015]. [Qin and Rohe, 2013] proposed to use the eigenvectors of the Laplacian matrix $\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$. As will become clear in the course of this manuscript, some of the stated methods have the disadvantage of having a more spread out spectrum leading to difficulty in reaching the phase transition. We propose in this thesis, a generic method that encompasses the latter in the dense DCSBM regime which identifies the best matrix (in the class of "adjacency" matrices) allowing to reach the phase transition in difficult scenarios. We also retrieve the normalization to perform on the eigenvectors in order to avoid biases.

As previously stated, the aforementioned works [Coja-Oghlan and Lanka, 2009, Gulikers et al., 2015, Jin et al., 2015] have shown that under some regularity (or regularization)

[2]The Bethe Hessian (BH) spectral method [Saade et al., 2014] is based on the union of the eigenvectors associated to the negative eigenvalues of $H(r_c)$ and $H(-r_c)$ respectively where $H(r) = (r^2-1)\mathbf{I}_n - r\mathbf{A} + \mathbf{D}$ for $r_c = \frac{\sum_i d_i^2}{\sum_i d_i} - 1$ with d_i the degree of node i ($\mathbf{D}$ and d_i are defined subsequently).

conditions, an almost perfect or perfect reconstruction of the node's labels can be achieved asymptotically. Our motivation in this first part of the thesis is to go beyond mere consistency results by understanding the performances of the different regularized spectral clustering algorithms for large but finite network sizes n. For this, we place ourselves in a regime where communities are too close to induce perfect reconstructions, so that the different spectral algorithms do not all lead to the same asymptotic trivial classification. As stated above, in order to encompass most aforementioned methods, we study here a generalized regularization of the adjacency matrix[3] given, for any $\alpha \in \mathbb{R}$, by

$$\mathbf{L}_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} \mathbf{D}^{-\alpha} \left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{2m} \right] \mathbf{D}^{-\alpha}$$

where $\mathbf{d}$ is the vector of degrees ($d_i = \sum_{j=1}^n A_{ij}$), $\mathbf{D}$ is the diagonal matrix of degrees (containing $\mathbf{d}$ on the main diagonal) and $m = \frac{1}{2}\mathbf{d}^\mathsf{T}\mathbf{1}_n$ is the number of edges in the network. In particular, $\mathbf{L}_0$ is the modularity matrix [Newman, 2006b, Jin et al., 2015], $\mathbf{L}_{\frac{1}{2}}$ is a modularity equivalent to the normalized Laplacian matrix [Qin and Rohe, 2013, Chung, 1997] and $\mathbf{L}_1$ is the form used in [Coja-Oghlan and Lanka, 2009, Gulikers et al., 2015, Tiomoko Ali and Couillet, 2016c].

We consider here a *dense* DCSBM model where $q_i = \mathcal{O}(1)$ (with respect to n)[4]. In this regime, when the correction factors $C_{g_i g_j}$ differ by $\mathcal{O}(1)$, consistency (i.e., vanishing error rates are guaranteed asymptotically) is shown for all regularized spectral algorithms. In order to compare those algorithms, we consider a regime where the communities are barely separable from the community matrix $\mathbf{C}$, but still identifiable. This regime is ensured for $C_{g_i g_j} = \mathcal{O}(1)$ individually but for the $C_{g_i g_j}$'s differing by $\mathcal{O}(n^{-\frac{1}{2}})$. Under this regime, we have studied in [Tiomoko Ali and Couillet, 2016a, Tiomoko Ali and Couillet, 2018] the dominant eigenvalues and associated eigenvectors (used for classification) of $\mathbf{L}_\alpha$ for large dimensional dense graphs following the DCSBM in the aforementioned regime. The study of the eigenvalues allows us to prove that there exists a non-trivial *optimal* $\mathbf{L}_\alpha$ where we can estimate on-line α from the observed graph. In [Tiomoko Ali and Couillet, 2018], we characterize the asymptotic content of the dominant eigenvectors of $\mathbf{L}_\alpha$ which led us to provide limiting values of the means and covariances of the centroids (corresponding to each community) that an Expectation Maximization (EM) algorithm (assuming Gaussian Mixture Model distribution of the eigenvectors' entries) would find after convergence. Those findings are then used in practice to initialize EM instead of the classically used random initialization.

[3] The leading term $\frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{2m}$ (not providing any information about the communities) is shown in simulations to have asymptotically no impact on the clustering performance. It is discarded here mostly for mathematical simplicity. Note in passing that $\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{2m}$ corresponds to the modularity matrix [Newman, 2006a], therefore $\mathbf{L}_\alpha$ may be seen as a "α-normalized" modularity matrix.

[4] In light of the previous spectral clustering methods, the same analysis is valid for q_i in order as small as $\frac{\log n}{n}$.

2.1.2 Multilayer graph community detection

The extraction of community structures shared by the different layers as well as the structure distinct between the layers might be a good way to understand the interplay between the different interactions in multilayer graphs. There is thus a need for new community detection methods to cope with such a heterogeneous structure in the multilayer graphs. The recovery of the communities independently in each layer (using classical single-layer community detection methods) is suboptimal as this approach does not exploit the common information across the layers. The current research effort aims at developing joint inference methods by multilayer aggregation [De Domenico et al., 2015a, Nicosia and Latora, 2015, De Bacco et al., 2017]. The simplest aggregation approach is to collapse the multilayer network to a single-layer network on which classical community detection methods can be applied [Tang et al., 2009, Tang et al., 2012, Comar et al., 2012, Zhang et al., 2013, Amelio and Pizzuti, 2014, De Domenico et al., 2015b, Taylor et al., 2016, Kim et al., 2017]. Alternatively, some researchers have suggested performing community detection separately in each layer followed by consensus aggregation of the communities across layers [Mucha et al., 2010, Xiang et al., 2012, Oselio et al., 2014, Amelio and Pizzuti, 2014, Paul and Chen, 2016]. Another approach is to extend single-layer SBMs to multilayer networks [Sweet et al., 2014, Han et al., 2015, Paul and Chen, 2015, Peixoto, 2015, Stanley et al., 2016, Valles-Catala et al., 2016, Reyes and Rodriguez, 2016, Barbillon et al., 2017], and use statistical inference approaches specific to the multi-layer architecture.

As stated above, in practice, some communities might be shared between the different layers, while others might not be (see Figure 1.2). However, few methods in the literature explicitly consider this general scenario. The Multilayer Extraction algorithm proposed in [Wilson et al., 2017] allows identification of heterogeneous multilayer communities where the communities might be shared between a subset of layers. [Wilson et al., 2017] minimize a cost function and take into account the similarities and dissimilarities between the layers' communities. While the model used in [Wilson et al., 2017] is realistic when considering multilayer graph connections, the method is only limited to unweighted graphs. The approach proposed in [Boden et al., 2012] extend [Zeng et al., 2006] to weighted multilayer graphs and allows the extraction of coherent dense subgraphs (cliques) shared by subsets of layers. However, those methods are limited to identification of dense communities and might fail when the communities are connected but with only a few edges.

In a joint work with the University of Michigan, we propose in [Tiomoko Ali et al., 2018c], a method that can simultaneously detect shared and unshared communities between heterogeneous weighted networks. We take here a model-based approach (statistical inference) where joint weighted stochastic block models (WSBM) that share "a part" of their community structures are proposed. Due to the intractable form of the posterior distribution of the hidden latent community membership variables given the observed adjacency matrices, we make use of a variational Bayes approach to approximate that posterior. The approximate mean field variational distribution is then used to infer the latent shared and private communities from the proposed multilayer WSBM. This extends

the works in [Aicher et al., 2014, Zhang and Zhou, 2017] devised for WSBM in single-layer graphs.

2.2 Kernel Spectral clustering

The so-called *curse of dimensionality* in ML increases exponentially the need for the important amount of training data when the data dimension increases in order to avoid overfitting phenomena. Also, a huge number of features (dimensions) makes the training process very slow. The most natural solution to the curse of dimensionality problem is to reduce the number of features in such a way to keep the most important information in the data, an operation called *dimensionality reduction*. For example, pixels in images tend to be highly correlated and thus, retaining only a summary of the representative pixels tremendously reduces the data dimensionality. The most important dimensionality reduction techniques are Principal Component Analysis (PCA) and manifold learning approaches. PCA, by far the most popular dimensionality reduction technique, consists in finding the hyperplane summarizing most of the data and then project the data on it. The hyperplane is chosen is such a way that it minimizes the mean squared distance between the original dataset and its projection on the hyperplane axes. The hyperplanes axes are called the *principal components* and correspond to the largest eigenvectors in the Singular Value Decomposition (SVD) of the data matrix. Once the dimensionality is reduced with the most important information kept, classical ML algorithms can be applied with less complexity and more efficiency. *Spectral clustering* approaches can be seen as variants of PCA and consist in clustering the data using only a few Principal Components. We are interested here in the investigation of *spectral clustering* on *high dimensional data*, the behavior of which will be seen to be completely different from the original intuition of spectral clustering in small dimensions. We consider spectral clustering algorithms

Algorithm 2: Spectral algorithm

1: Compute the, say, ℓ eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_\ell \in \mathbb{R}^n$ corresponding to the dominant (largest or smallest) eigenvalues of one of the matrix representations of the network (adjacency, modularity, Laplacian) of size $n \times n$.
2: Stack the vectors $\mathbf{u}_i$'s columnwise in a matrix $\mathbf{W} = [\mathbf{u}_i, \ldots, \mathbf{u}_\ell] \in \mathbb{R}^{n \times \ell}$.
3: Let $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^\ell$ be the rows of $\mathbf{W}$. Cluster $\mathbf{r}_i \in \mathbb{R}^\ell$, $1 \leq i \leq n$ in one of k groups using any low-dimensional classification algorithm (e.g., k-means [Hartigan and Wong, 1979] or Expectation Maximization (EM) [Ng et al., 2012]). The label assigned to $\mathbf{r}_i$ then corresponds to the label of node i.

(Algorithm 2) of n data vectors say $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ that we wish to classify into distinct classes. A basic approach would consist in finding hyperplanes (in the $\mathbb{R}^p$-space) separating the different vectors into distinct classes. This method can be shown to be equivalent to performing a Principal Component Analysis (PCA) or spectral method on the affinity matrix containing the inner-product between the data. However, this separation is only

valid when the data are linearly separable which is not the case in most ML datasets. To cope with this difficulty, kernel methods have been introduced and consist in projecting the non linear objects into a higher dimensional feature space where hyperplanes can be found to linearly separate the data. The idea is thus to move $\mathbf{x}_i$ to $\Phi(\mathbf{x}_i)$ with $\Phi : \mathbb{R}^p \to \mathbb{R}^P$ (with $P \gg p$) and use a spectral method on the empirical covariance matrix in the feature space $\mathbb{R}^P$. The former operation being obviously prohibitive for large P, the kernel-trick [Schölkopf, 2001] was introduced to consider instead the similarity $f(\mathbf{x}_i, \mathbf{x}_j)$ between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ where f is a function such that $\Phi(\mathbf{x}_i)^\mathsf{T}\Phi(\mathbf{x}_j) = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ or $\Phi(\mathbf{x}_i)^\mathsf{T}\Phi(\mathbf{x}_j) = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$. Such functions f verifying certain properties [Schölkopf, 2001] are shown to exist. Spectral clustering can then be applied on the kernel matrix, a method known as *kernel spectral clustering*. It is shown in [Von Luxburg et al., 2008] that spectral clustering on well known affinity matrices (Laplacian) between growing number of random samples $\mathbf{x}_i \in \mathbb{R}^p, 1 \leq i \leq n$ ($n \to \infty$ with p fixed) is consistent in the sense that the dominant eigenvectors of those similarity matrices asymptotically converge to some limiting eigenfunctions exhibiting "reasonable partitioning" of the samples. However, as we will show in the sequel, clustering high dimensional data ($n \to \infty, p \to \infty$ with constant p/n) is less trivial and important differences exist compared to small dimensional data clustering.

The original intuition of [Ng et al., 2002] behind spectral clustering is no longer valid with high dimensional data. To see this, we will start by describing the kernel spectral clustering mechanism in small dimensions. Let us consider n small dimensional data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ that we wish to classify into two distinct classes using a kernel fonction f such that the affinity $K_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$ between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is large when those vectors belong to the same class and vanishes when they belong to distinct classes. Assuming without loss of generality that the vectors $\mathbf{x}_i$ are ordered by classes i.e., $\mathbf{x}_1, \cdots, \mathbf{x}_{n/2}$ constitute the first class while $\mathbf{x}_{n/2+1}, \cdots, \mathbf{x}_n$ form the second class, we note that the vectors $[\mathbf{1}_{\frac{n}{2}}, \mathbf{0}_{\frac{n}{2}}]$ and $[\mathbf{0}_{\frac{n}{2}}, \mathbf{1}_{\frac{n}{2}}]$ that we shall call in the sequel *the canonical vectors of the classes*, are approximately eigenvectors of the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{K}$ (with $\mathbf{D} = \text{diag}(\left\{\sum_{j=1}^n K_{ij}\right\}_{i=1}^n)$) associated with the smallest eigenvalue 0, and thus a spectral clustering algorithm using those eigenvectors will assign the right class to each data without error. In close-to-ideal situations, $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$ will be relatively large for data of the same class so that a spectral algorithm using the eigenvectors associated with the smallest eigenvalues of $\mathbf{L}$ will be able to retrieve the classes with a reasonable performance. This reasoning is no longer valid in the "big data" regime where the dimension p is very large. To illustrate this fact, let us assume that the vectors $\mathbf{x}_i$ are independent Gaussian with the same covariance $\mathbf{I}_p$ but with mean $\boldsymbol{\mu}_1$ when they belong to the first class and mean $\boldsymbol{\mu}_2$ when they belong to the second class. We thus have $\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{p} \simeq \frac{\boldsymbol{\mu}_1^\mathsf{T}\boldsymbol{\mu}_2}{p}$ when $\mathbf{x}_1 \neq \mathbf{x}_2$ while $\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_i}{p} \simeq \frac{\|\boldsymbol{\mu}_1\|^2}{p} + 1$ when $\mathbf{x}_1 \neq \mathbf{x}_2$. As will be shown later, the class means should satisfy $\|\boldsymbol{\mu}_a\| = \mathcal{O}(1)$ (with respect to p) in order to avoid having **i)** asymptotically very distant vectors such that clustering becomes trivial or **ii)** very closed vectors such that clustering is impossible. As a consequence of $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = \mathcal{O}(1)$, the affinity K_{ij} (for $i \neq j$) asymptotically converges to the same value $f(0)$ no matter the class membership of

$\mathbf{x}_i$ and $\mathbf{x}_j$. It thus seems not possible to retrieve the classes using the spectral algorithm procedure described above since the notion of "closeness" between data of the same class is no longer valid in high dimensions. However, it turns out that the spectral algorithm still works as can be seen in Figure 1.3 where the dominant eigenvectors of $\mathbf{K}$ for large dimensional vectors are noisy plateaus with each plateau representing a class and thus a k-means method would be able to retrieve the class information with non-trivial errors. It is thus essential to understand the reasons why large dimensional kernel spectral clustering methods still work despite the fact that the notion of "closeness" at the heart of spectral clustering is no longer valid.

In order to understand the different mechanisms of large dimensional spectral clustering, it is important to understand the eigenstructure of large structured random affinity matrices $\mathbf{K}$. It is clear from the structure of $K_{ij} = f\left(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j\right)$ that $\mathbf{K}$ has non linear entries and dependent columns, a structure not common in classical random matrix theory. Early works by [El Karoui et al., 2010] have shown that when the high dimensional vectors $\mathbf{x}_i$ are Gaussian with zero mean and covariance $\mathbf{C}$ (thus no class assumption here), the matrix $\mathbf{K} = \left\{f\left(\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{p}\right)\right\}_{i,j=1}^n$ is asymptotically equivalent to a random matrix of the type "perturbed empirical covariance matrix". Specifically, it was shown in [El Karoui et al., 2010] that

$$\|\mathbf{K} - \hat{\mathbf{K}}\| \to 0$$

in probability where

$$\hat{\mathbf{K}} = f'(0)\mathbf{X}\mathbf{X}^\mathsf{T} + \left(f(0) + f''(0)\frac{\operatorname{tr}\mathbf{C}^2}{2p^2}\right)\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \left(f\left(\frac{\operatorname{tr}\mathbf{C}}{p}\right) - f(0) - f'(0)\frac{\operatorname{tr}\mathbf{C}}{p}\right)\mathbf{I}_n$$

with $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. The matrix $\hat{\mathbf{K}}$ has a simple structure from a "random matrix" point of view: it is essentially a scaled version of the empirical covariance matrix $f'(0)\mathbf{X}\mathbf{X}^\mathsf{T}$ added to the low rank matrix $\mathbf{1}\mathbf{1}^\mathsf{T}$ and a scaled identity matrix. The eigenstructure of $\mathbf{K}$ can thus be performed as in classical RMT. Leveraging on [El Karoui et al., 2010], the work [Couillet and Benaych-Georges, 2016] generalizes the study of $\mathbf{K} = \{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2)\}_{i,j=1}^n$ to data $\mathbf{x}_i$'s arising from a Gaussian Mixture Model (GMM) with k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ such that $\mathbf{x}_i$ is Gaussian with mean $\boldsymbol{\mu}_a$ and covariance $\mathbf{C}_a$ when it belongs to class $\mathcal{C}_a$. In the big data regime where the data dimension p scales linearly with the number n of data ($n, p \to \infty$ with $n/p = \mathcal{O}(1)$), [Couillet and Benaych-Georges, 2016] show that $\mathbf{K}$ asymptotically behaves as a so-called spiked random matrix, i.e., the sum of a "deformed sample covariance matrix" with perturbation matrices containing the information about the classes (means, covariances). This spiked model structure allows to perform a thorough study of the eigenvalues and eigenvectors of $\mathbf{K}$ in this large dimensional regime. The random matrix analysis of the dominant eigenvectors reveals in particular that the choice of the kernel function f has a strong impact on the discrimation of the data based upon their statistical class means and/or class covariances.

In this thesis, we also consider high dimensional data arising from a mixture of k Gaussian distribution with means $\boldsymbol{\mu}_a$ and covariance $\mathbf{C}_a$ ($a = 1, \cdots, k$). In this high

dimensional setting, assuming a supervised scenario where the means and covariances are known and the objective is to retrieve the classes of each data. The authors in [Couillet et al., 2018] derive the minimal rates of the distances between the class means and class covariances required to asymptotically achieve a non trivial clustering (i.e., neither perfect nor impossible). Those minimal rates (that we shall call in the sequel *oracle optimal rates*), provided below in the case $k = 2$ thus constitutes a baseline of comparison between different clustering methods (semi-supervised or unsupervised).

- $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = \mathcal{O}(1)$.
- $\|\mathbf{C}_1 - \mathbf{C}_2\| = \mathcal{O}\left(\frac{1}{\sqrt{p}}\right)$.
- $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2) = \mathcal{O}(\sqrt{p})$.
- $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(1)$.

The objective of this part of the thesis is to propose efficient kernel spectral clustering methods (unsupervised) capable of discriminating data with distance rates closed to the *oracle optimal rates*[5]. To this end, we start by analyzing the kernel matrix $\mathbf{K} = f\left(\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{p}\right)$ with f a generic kernel function supposed to be at least 3 times differentiable around 0 (since $\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{p} \to 0$ in the regime under study). This study reveals that estimating the class labels by spectral clustering on $\mathbf{K}$ does not perform better than a random guess for generic function f when class means are equal and $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 \ll \mathcal{O}(p)$ which is far from the optimal oracle rate. By using instead a kernel function f such that $f'(0) = 0$, one can perform better than a random guess when $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(\sqrt{p})$ thus inducing better performances than generic kernel functions. However, the latter kernel has the negative side of cancelling the effect of the data means on the clustering i.e., using this kernel will not be able to differentiate data with same covariances but with strongly differing means. This case is carefully studied in [Kammoun and Couillet, 2017] for inner product kernels $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ with kernel function f such that $f'(0) = 0$ (0 being the limiting value of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p$). In this thesis, we show that a careful random matrix analysis of the inner product kernel matrix $\mathbf{K}$ reveals that setting $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$ instead of $f'(0) = 0$ allows for a fair treatment between class means and class covariances in the clustering procedure and is able to to discriminate the data with rates condition $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(\sqrt{p})$ as for the kernel with $f'(0) = 0$. This new kernel matrix which takes into account balance between statistical means and covariances in the data is able to discriminate the data with better rates compared to previously studied kernels. Such kernel choices are important in practice since specific choices of f are more adapted to datasets containing either pronounced differences between class means as in the popular MNIST dataset [LeCun, 1998] while other functions are more efficient on datasets with classes having similar means but strongly differing covariances [Andrzejak et al., 2001].

[5]The term *oracle* is employed to stress the fact that those optimal rates are obtained in a supervised case where the class means and class covariances are perfectly known.

2.3 Outline and Contributions

Outline and related publications

As stated in the previous sections, this thesis covers two main clustering applications: *community detection in graphs* and *high dimensional kernel data clustering*. As a first main contribution for community detection in graphs, leveraging a complete study of the spectrum of graph similarity matrices, we devise an improved spectral algorithm for large dense and heterogeneous *single-layer* graph models as a generalization of previous classical spectral methods. The second main contribution concerns *multi-layer* graph community detection where we devise a model-based algorithm to infer heterogeneous communities from the different layers. As for *data kernel spectral clustering*, through a complete study of the spectrum of large kernel similarity matrices, we provide some important insights on the appropriate kernel functions to use based on the statistics of the data.

In Chapter 3, we introduce the mathematical tools necessary to follow the different results of this thesis. The main lever of Chapters 4 and 6 is the spectral analysis (eigenvalues and eigenvectors) of large random matrices while in Chapter 5, we use variational Bayes inference methods to devise a multi-layer community detection algorithm with overlapping communities between the different layers. Those two completely different mathematical tools (Random Matrix theory and Variational Inference) are introduced to constitute a basis for the following chapters.

In Chapter 4, we consider the large dimensional spectral clustering dense community detection problem on realistic networks which can have heterogeneous degree distributions. We consider a generalized form $\mathbf{L}_\alpha$ (parametrized by α) of the most used similarity matrices (given for different values of α) used for spectral clustering on dense graphs, under the realistic statistical graph model, the DCSBM. The performances of those spectral methods depending on the position of the eigenvalues of $\mathbf{L}_\alpha$ as well as the content of the corresponding eigenvectors, we study the eigenstructure of $\mathbf{L}_\alpha$. The matrix $\mathbf{L}_\alpha$ not being a tractable random matrix clearly exhibiting the community structure of the graph, a first step is to approximate $\mathbf{L}_\alpha$ as the number of nodes $n \to \infty$ by a theoretically tractable random matrix $\tilde{\mathbf{L}}_\alpha$ which falls in the family of so-called spiked random matrix models and which allows for a thorough study of eigenvalues and eigenvectors of $\mathbf{L}_\alpha$. Spiked random matrices (which will be introduced in Chapter 3) generally exhibit a phase transition beyond which useful information can be extracted from the eigenvectors associated to outlying eigenvalues (and below which nothing can be said). In our context, this phase transition corresponds to a *community detectability threshold*, common in community detection algorithms analysis. We characterize exactly this phase transition for each value of α. We then prove the existence of and obtain an expression for an optimal value α_{opt} of α for which the community detectability threshold[6] is maximally achievable. This value needs not be either 0 or 1 and its proper choice is of utmost importance in highly het-

[6]The community detectability threshold is the point beyond which there exists a clustering algorithm which can do better than a random guess.

erogeneous graphs. We provide a consistent estimator $\hat{\alpha}_{\text{opt}}$ of α_{opt} based on $\mathbf{d}$ alone. We show that to achieve consistent clustering in the DCSBM model, the dominant eigenvectors used for clustering should be pre-multiplied by $\mathbf{D}^{\alpha-1}$ prior to the low dimensional classification (step 3 of Algorithm 2), thereby recovering the SCORE algorithm [Jin et al., 2015] for $\alpha = 0$ and the algorithm in [Gulikers et al., 2015] for $\alpha = 1$, as special cases. A deeper study of the regularized eigenvectors allows us to improve the initial setting of the EM algorithm (in the step 3 of the spectral algorithm described above) in comparison with a random setting. Numerical simulations show that our methods outperform state-of-the-art spectral methods both on synthetic graphs and on real-world networks. This line of work started with the following paper where we investigate the spectral analysis of the particular matrix $\mathbf{L}_1 \propto \mathbf{D}^{-1}\left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\intercal}{\mathbf{d}^\intercal \mathbf{1}_n}\mathbf{D}^{-1}\right]$ in the large dimensional dense DCSBM models. A study of the eigenvalues and eigenvectors of $\mathbf{L}_1$ allows us to derive the phase transition threshold as well as the asymptotic misclassification rates of spectral algorithms using $\mathbf{L}_1$.

[[Tiomoko Ali and Couillet, 2016b]] Tiomoko Ali, H. and Couillet, R. (2016). Performance analysis of spectral community detection in realistic graph models. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)

The study of the generalized scaled matrices $\mathbf{L}_\alpha$ along with the characterization of phase transition and estimation of the *optimal* α was then performed in

[[Tiomoko Ali and Couillet, 2016a]] Tiomoko Ali, H. and Couillet, R. (2016a). community detection in heterogeneous networks. In Signals, Systems and Computers, 2016 50th Asilomar Conference on, pages 1385–1389. IEEE.

The following article published in the Journal of Machine Learning Research encompasses all the aforementioned works with all proofs and calculation details. Also, the study of the eigenvectors is performed followed by the improvement of the classical EM algorithm in the last step of the spectral clustering procedure.

[[Tiomoko Ali and Couillet, 2018]] Tiomoko Ali, H. and Couillet, R. (2018). Improved spectral community detection in large heterogeneous networks. Journal of Machine Learning Research, 18:1–49.

In Chapter 5, we investigate the problem of multi-layer community detection with heterogeneous communities between the layers. We propose a new model-based method to simultaneously detect shared and unshared communities between heterogeneous weighted graphs. We define joint weighted stochastic block models (WSBM) that take into account similarities and dissimilarities between the community structures. Due to the intractable form of the posterior distribution of the latent community memberships given the observed similarity matrices of the different layers, we derive a variational Bayes algorithm for automatically inferring shared and unshared communities from multilayer weighted graphs. We establish that the proposed algorithm is more accurate and robust than previous approaches to community detection in multi-layer networks in extracting both shared and unshared communities from weighted graph benchmarks. We finally illustrate a real-world use of our method in multinomic molecular biology enabling the discovery of heteroge-

neous multilayer communities of gene-gene interactions in human fibroblast proliferation. The work presented in this chapter was submitted to the NIPS'2018 conference

[[Tiomoko Ali et al., 2018c]] Ali, H. T., Liu, S., Yilmaz, Y., Hero, A., Couillet, R., and Rajapakse, I. (2018b). Latent heterogeneous multilayer community detection.

In Chapter 6, we perform a random matrix study of high dimensional data clustering with inner product kernels under a Gaussian Mixture Model (GMM) assumption with growth rates on the means and covariances set in such a way to achieve non-trivial clustering (i.e., neither perfect clustering nor impossible clustering). The first study of the inner product kernels with generic kernel functions f is performed in

[[Tiomoko Ali et al., 2018a]] Tiomoko Ali, H., Kammoun, A., and Couillet, R. (2018a). Random matrix asymptotic of inner product kernel spectral clustering. In International Conference on Acoustics, Speech and Signal Processing (ICASSP'18). IEEE.

In the previous work, it is seen that generic kernel functions f are rate suboptimal compared to the oracle setting. It is also observed that taking $f'(0) = 0$ leads to improving the rate from $\mathrm{tr}\,(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(p)$ to $\mathrm{tr}\,(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(\sqrt{p})$ but completely annihilating the effect of the class means. In the following work [[Tiomoko Ali et al., 2018b]], we propose new kernels conciliating the previous works by balancing between the class means and class covariances in the data while allowing to achieve the same distance rate with the kernel f such that $f'(0) = 0$.

[[Tiomoko Ali et al., 2018b]] Tiomoko Ali, H., Kammoun, A., and Couillet, R. (2018b). Random matrix-improved kernels for large dimensional spectral clustering. In Statistical Signal Processing Workshop (SSP), 2016 IEEE, pages 1–4. IEEE.

Other publications

In the course of the PhD, we have also provided through a series of work, a quantitative analysis of the performance of *linear* echo-state networks (ESNs). While ESNs are not so used in practice, the theoretical tools that we have developed to analyze their performances convey a deeper understanding on the core mechanism underplay and provide some intuition on the memory functionality of Recurrent Neural Networks (RNN). Specifically, we provide a first theoretical analysis of the *mean-square error* performance of linear ESNs for both training and testing phases in the regime where the reservoir size and the training (or testing) duration are large and commensurable. We leverage on (random matrices) concentration of measure properties to provide a deterministic limit of the aforementioned performance (mean-square error). Those works were conducted in the following publications

[[Couillet et al., 2016b]] Couillet, R., Wainrib, G., Sevi, H., and Tiomoko Ali, H. (2016c). Training performance of echo state neural networks. In Statistical Signal Processing Workshop (SSP), 2016 IEEE, pages 1–4. IEEE.

[[Couillet et al., 2016c]] Couillet, R., Wainrib, G., Tiomoko Ali, H., and Sevi, H. (2016d). A random matrix approach to echo-state neural networks. In International Conference on Machine Learning, pages 517–525.

[[Couillet et al., 2016a]] Couillet, R., Wainrib, G., Sevi, H., and Tiomoko Ali, H. (2016b). The asymptotic performance of linear echo state neural networks. Journal of Machine Learning Research, 17(178):1–35.

As these works cover topics not directly within the scope of the present manuscript, we do not further elaborate on them here.

Chapter 3

Mathematical background

3.1 Introduction

This chapter introduces theoretical concepts necessary to understand the results in the subsequent chapters. As stated in Chapter 2, the performance analysis of spectral clustering methods (community detection in graphs, kernel spectral clustering) requires the understanding of the behavior of the eigenvalues and eigenvectors of large random matrices representing similarities of the data analyzed through their closest probabilistic model. A large field of Random Matrix Theory has covered throughout the years the analysis of the eigenvalues and eigenvectors of large covariance matrices or large Hermitian matrices as well as some transformations of those. Related to clustering are the so-called *spiked random matrices* (equivalent to low rank information + high rank noise matrices) which will be used throughout this thesis, the dominant eigenvectors of which contain the relevant information about the data model. We discuss in particular in this chapter simple introductory results on spiked random matrix analysis along with sketches of proofs. This will form a basis to follow the main results of this thesis which involve more general spiked random matrix models. In the second part of this thesis, we use variational Bayes inference methods to devise a new method for multi-layer community detection. The end of this chapter will introduce the overall idea behind variational Bayes inference methods in order to follow the results in Chapter 5.

3.2 Random matrix theory

Random matrix theory (RMT) dates back to 1928 with the work of the statistician John Wishart [Wishart, 1928] who was interested in studying the behavior of sample covariance matrices in the form $\hat{\mathbf{C}}_p = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{H}}$ of i.i.d random processes $\mathbf{x}_i, \ldots, \mathbf{x}_n \in \mathbb{C}^p$. Wishart provided the joint distribution of the eigenvalues of $\hat{\mathbf{C}}_p$ for independent vectors identically distributed as standard Gaussian ($\hat{\mathbf{C}}_p$ is called Wishart matrix since then). Later then, Eugene Wigner [Wigner, 1993] introduced the eigenspectrum analysis of large random Her-

mitian matrices by postulating that the spacing between the lines in the spectrum of heavy atom nucleus resemble the spacing between the eigenvalues of symmetric matrices with independent entries (matrices known since then as *Wigner matrices*) uniformly distributed in $\{-1, 1\}$. He then proved the convergence of the eigenvalue distribution of the latter random matrix towards the *semi-circle* law for growing matrix dimensions. The works of Wigner then triggered enormous research on the properties of the joint eigenvalue distribution, the distribution of extreme eigenvalues and eigenvectors of random matrices with growing size. Those findings have gained more attention in many domains involving large random matrices models such as in physics [Mehta, 2004], finance [Laloux et al., 2000], evolutionary biology [Arnold et al., 1994], wireless communications [Telatar, 1999,Couillet and Debbah, 2011] and recently the growing field of machine learning [El Karoui et al., 2010,Benaych-Georges and Couillet, 2016,Couillet and Benaych-Georges, 2016]. The goal of this chapter is not an exhaustive development of the different concepts in the spectral analysis of large random matrices but the introduction of the necessary tools to follow our results in large dimensional data clustering (kernel spectral clustering and spectral community detection).

3.2.1 Limiting eigenvalue distribution of large dimensional random matrices

As stated above, the performance analysis of many real-world applications appeals to the knowledge of the eigenvalues distribution of large Wishart matrices introduced above. As introductory example, let us consider a sequence of vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$ randomly drawn from a p-variate zero mean random process. For the dimension p fixed, as $n \to \infty$, the sample covariance matrix $\hat{\mathbf{C}}_p = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\mathsf{T}$ converges almost surely to the *population covariance matrix* $\mathbf{C} = \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^\mathsf{T}]$ in the sense that $\|\hat{\mathbf{C}}_p - \mathbf{C}\| \to 0$ on a set of probability one and the convergence is true for any norm. This result follows from Markov inequality on sufficiently high order moments, the union bound, and the Borel Cantelli lemma leading to a "uniform" law of large numbers. However, this convergence result is no longer true for large p-large n with n, p growing at the same rate. To give an illustrative counter example, let us consider i.i.d vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$ with $\mathbf{x}_1 \sim \mathcal{CN}(0, \mathbf{I}_p)$ such that $\frac{p}{n} \to c > 1$. Denoting $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$, from the law of large numbers, we have the *joint point-wise convergence*

$$\max_{1\leq i,j\leq p} \left| \left[\hat{\mathbf{C}}_p - \mathbf{I}_p\right]_{ij} \right| = \max_{1\leq i,j\leq p} \left| \frac{1}{n}\sum_{k=1}^n X_{ik}X_{jk}^\star - \delta_{ij} \right| \xrightarrow{\text{a.s.}} 0. \tag{3.1}$$

However, the convergence in spectral norm is not true since $\hat{\mathbf{C}}_p$ has at least $p - n$ zero eigenvalues while all the eigenvalues of the population cavariance matrix $\mathbf{C} = \mathbf{I}_p$ are all equal to one. More generally, for $p/n \to c > 0$, the eigenspectrum of $\hat{\mathbf{C}}_p$ tends to spread far from 1 (the eigenvalue of the population covariance matrix), as illustrated in Figure 3.1 (where we take $n = 2000$ draws of $\mathbf{x}_1 \sim \mathcal{CN}(0, \mathbf{I}_p)$ with $p = 500$). This important observation has led to an important field of RMT dedicated to the study of limiting eigen-

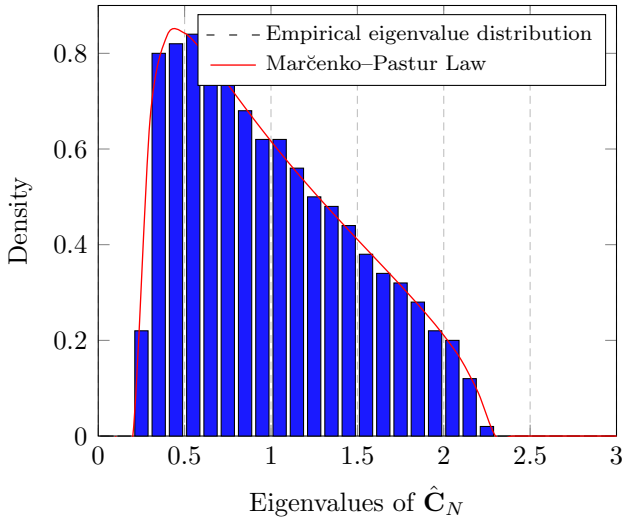


Figure 3.1: Histogram of the eigenvalues of $\hat{\mathbf{C}}_p$ for $p = 500$, $n = 2000$, $\mathbf{C}_p = \mathbf{I}_p$.

spectrum distribution of large sample covariance matrices and large Hermitian random matrices in general. Before stating the prior results on the eigenvalues distributions of those random matrices, we first define the marginal density of their eigenvalues.

Definition 1 (Empirical Spectral Density)**.** *Let us consider a symmetric matrix* $\mathbf{X}_n \in \mathbb{R}^{n\times n}$*. We define its empirical spectral density (e.s.d.)* $\mu_n^{\mathbf{X}_n}$ *as*

$$\mu_n^{\mathbf{X}_n} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\delta}_{\lambda_i(\mathbf{X}_n)}.$$

with $\lambda_i(\mathbf{X}_n)$ *the* i*-th eigenvalue of* $\mathbf{X}_n$*.*

An important class of large random Hermititian matrices has the property that their random e.s.d. converges in some sense to a deterministic limiting distribution that we shall call the *limiting spectral distribution* (l.s.d.). In the following, we recall the well-known results which will also be the basis of our results in the subsequent chapters. The first (historical) result is due to Wigner [Wigner, 1993] and concerns the convergence of the e.s.d. of the aforementioned Wigner matrices.

Theorem 2 (Theorem 2.5 and Theorem 2.9 of [Bai and Silverstein, 2010])**.** *Consider an* $n \times n$ *symmetric matrix* $\mathbf{X}_n$*, with independent entries* $\frac{1}{\sqrt{n}}(\mathbf{X}_n)_{ij}$ *such that* $\mathbb{E}[(X_n)_{ij}] = 0$*,* $\mathbb{E}[|(X_n)_{ij}|^2] = 1$ *and there exists* ϵ *such that the* $(X_n)_{ij}$ *have a moment of order* $2+\epsilon$*. Then* $\mu_n^{\mathbf{X}_n} \Rightarrow \mu$ *almost surely, where* μ *has a density* f *defined as*

$$f(x) = \frac{1}{2\pi}\sqrt{(4-x^2)^+}. \tag{3.2}$$

Moreover, if the $(X_n)_{ij}$ *are identically distributed, the result holds without the need for existence of a moment of order* $2+\epsilon$*.*

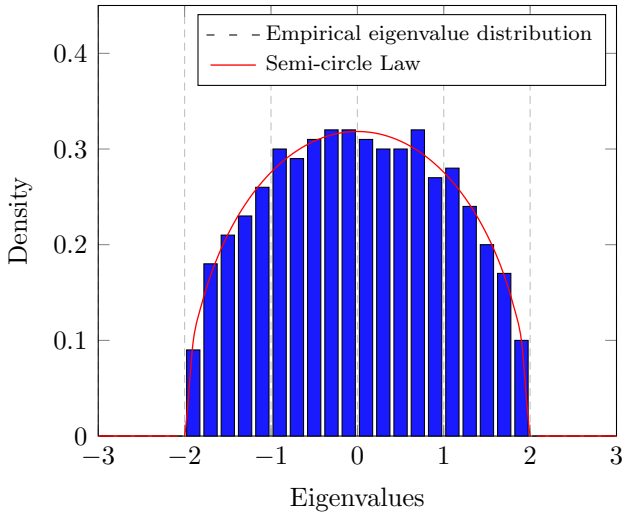


Figure 3.2: Histogram of the eigenvalues of Wigner matrices and the semi-circle law, for $n = 500$.

An example of l.s.d of a symmetric matrix with independent upper-diagonal $X_{ij} \sim \mathcal{N}(0, 1/n)$ entries, is shown in Figure 3.2. This result is proved in [Bai and Silverstein, 2010] using the method of moments (Section 30 of [Billingsley, 1995]) but also using the Stieltjes transform method in [Bai and Silverstein, 2010]. A similar result is obtained in our community detection problem on heterogeneous graph models where we deal with Wigner matrices with independent but not identically distributed adjacency entries (having a variance profile). We get in that case a *deformed* semi-circle law; this will be detailed in Chapter 4.

A more involved result concerns non-symmetric matrices where the l.s.d. is this time a *full circle* law, the eigenvalues of those matrices lying in the complex plane. The first proof of this result is by Girko but we give here the more general result due to Bai in 1997 [Bai, 2008]

Theorem 3 (Full circle law [Bai, 2008])**.** *Let $\mathbf{X}_n \in \mathbb{R}^{n\times n}$ have i.i.d entries $\frac{1}{\sqrt{n}}(X_n)_{ij}$, $1 \leq i, j \leq n$, such that $(X_n)_{11}$ has zero mean, unit variance. Additionally, assume that the joint distribution of the real and imaginary parts of $\frac{1}{\sqrt{n}}(X_n)_{11}$ has bounded density. Then, with probability one, the e.s.d. of $\mathbf{X}_n$ tends to the uniform distribution on the unit complex disc. This distribution is referred as the circular law, or full circle law.*

This result is given for completeness but does not intervene in the applications of this thesis. Figure 3.3 illustrates the circular law for a non-symmetric matrix with i.i.d. $X_{ij} \sim \mathcal{N}(0, 1/n)$ entries.

Let us get back to the sample covariance matrix. We recall that in the large n - large p regime, the sample covariance matrix does not converge to the population covariance matrix. However, the e.s.d. of a large class of sample covariance matrix converges to a well-known deterministic distribution. This well-known result with important applications

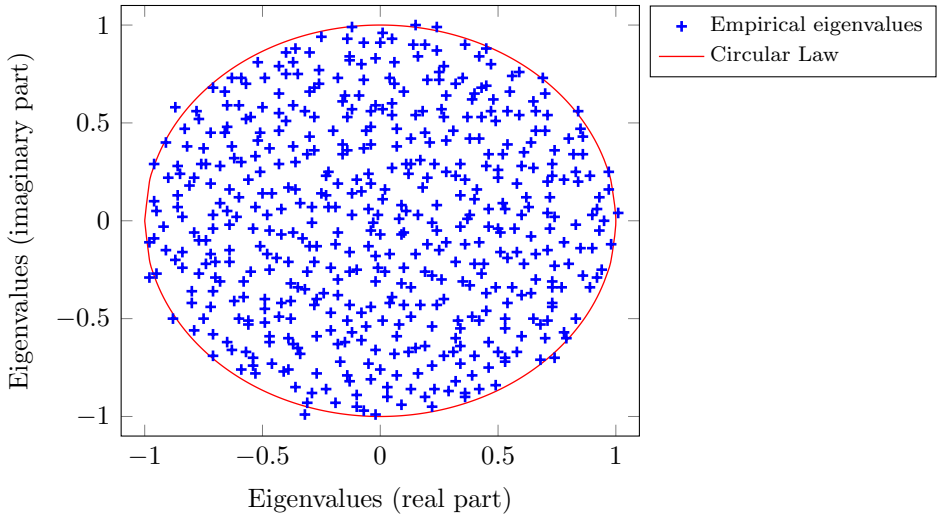


Figure 3.3: Eigenvalues of $\mathbf{X}_n$ with i.i.d. standard Gaussian entries, for $n = 500$.

in wireless communications and now in machine learning concerns the convergence of the Gram matrix (empirical covariance matrix) of a random matrix with i.i.d entries of zero mean and normalized variance to the so-called *Marčenko-Pastur law* (MP) [Marčenko and Pastur, 1967].

Theorem 4 (Marčenko-Pastur law [Marčenko and Pastur, 1967])**.** *Consider a matrix* $\mathbf{X} \in \mathbb{R}^{p\times n}$ *with i.i.d. entries* $\left(\frac{1}{\sqrt{n}}(X_p)_{ij}\right)$, *independent for all* i, j, n *such that* $(X_p)_{ij}$ *has zero mean, unit variance. As* $n, p \to \infty$ *with* $\frac{p}{n} \to c \in (0, \infty)$, *the e.s.d. of* $\hat{\mathbf{C}}_p = \mathbf{X}\mathbf{X}^\mathsf{T}$ *converges weakly and almost surely to a non-random distribution function* μ *with density* f_c *given by:*

$$f_c(x) = (1 - c^{-1})\delta(x) + \frac{1}{2\pi c x}\sqrt{(x-a)^+(b-x)^+}. \tag{3.3}$$

where $a = (1 - \sqrt{c})^2$, $b = (1 + \sqrt{c})^2$ *and* $\delta(x) = 1_0(x)$.

Figure 3.4 exhibits the MP distribution for different limit ratios c. We can notice that in particular, as $c \to 0$ (when the number of samples n is much larger than the fixed dimension p for example) the distribution shrinks towards 1 since as evidenced previously, the spectral norm difference between the sample covariance matrix and the population covariance matrix converges to 0 for large n.

The proofs of those convergence results rely either on the moment method (Section 30 of [Billingsley, 1995]) or on the Sieltjes transform method. We use in this thesis Stieltjes transform based approaches to prove the convergence of e.s.d. of the random similarity matrices involved in our clustering problems. Let us thus introduce the Stieltjes transform.

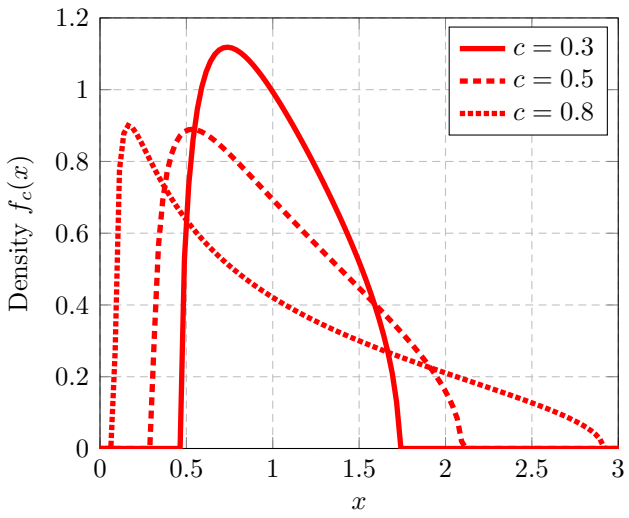


Figure 3.4: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

3.2.2 The Stieltjes transform and its properties

Definition 5 (The Stieltjes transform)**.** *Let μ be a real-valued bounded measurable function over $\mathbb{R}$. Then, the Stieltjes transform $m_\mu(z)$ of μ, for $z \in Supp(\mu)^c$, the complex space complementary to the support*[1] *of μ, is defined as*

$$m_\mu(z) \triangleq \int_{-\infty}^{\infty} \frac{1}{t-z} \mathrm{d}\mu(t). \tag{3.4}$$

The inverse mapping is defined for all distributions μ admitting a Stieltjes transform m_μ as follows

Theorem 6 (Inverse transformation)**.** *If x is a continuous point of μ, then:*

$$\mu(x) = \frac{1}{\pi} \lim_{y\to 0^+} \int_{-\infty}^{x} \mathcal{I}[m_\mu(x+iy)]\mathrm{d}x. \tag{3.5}$$

Working directly with the e.s.d. of large random matrices to find its limiting distribution turns out to be a hard task in most cases. As a workaround, the Stieltjes transform method is employed instead. It first consists in computing the Stieltjes transform m_{μ_n} of the e.s.d μ_n, finding its limit m_μ which in most cases is also the Stieltjes transform of a distribution function μ. Using the following theorem, one can then show that the l.s.d. of μ_n is μ.

Theorem 7 (Theorem $B.9$ of [Bai and Silverstein, 2010])**.** *Let $\{\mu_n\}$ be a set of bounded real functions such that $\lim_{x\to-\infty} \mu_n(x) = 0$. Then, for all $z \in \mathbb{R}$*

$$\lim_{n\to\infty} m_{\mu_n}(z) = m_\mu(z) \tag{3.6}$$

[1]The support of $Supp(\mu)$ of a distribution function μ with density f is defined as the closure of the set $\{x \in \mathbb{R}, f(x) > 0\}$

if and only if there exists μ *such that* $\lim_{x\to-\infty}\mu(x)=0$ *and* $|\mu_n(x)-\mu(x)|\to 0$ *for all* $x\in\mathbb{R}$.

Due to the following property, working with the Stieltjes transform m_{μ_n} of the symmetric matrix $\mathbf{X}$ e.s.d 's μ_n boils down to working with a tractable functional of the matrix $\mathbf{X}$ itself.

Remark 8. *For a symmetric matrix* $\mathbf{X}\in\mathbb{R}^{n\times n}$, *the e.s.d.* $\mu_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i(\mathbf{X})}$ *has Stieltjes transform given by*

$$m_{\mu_n}(z)=\frac{1}{n}\operatorname{tr}\left(\mathbf{X}-z\mathbf{I}_n\right)^{-1}.$$

To see this, we have

$$\begin{aligned} m_{\mu_n}(z) &= \int\frac{1}{t-z}\mathrm{d}\mu_n(t)\\ &= \frac{1}{n}\operatorname{tr}\left(\mathbf{\Omega}-z\mathbf{I}_n\right)^{-1}\\ &= \frac{1}{n}\operatorname{tr}\left(\mathbf{X}-z\mathbf{I}_n\right)^{-1} \end{aligned}$$

where $\mathbf{\Omega}$ *is the diagonal matrix containing the eigenvalues of* $\mathbf{X}$ *and in the last line we use the spectral decomposition of the matrix* $\mathbf{X}$ *along with the commutative property of the trace. The matrix* $(\mathbf{X}-z\mathbf{I}_n)^{-1}$ *is the so-called* resolvent *of the random matrix* $\mathbf{X}$.

As stated above, finding the limiting distribution (l.s.d.) of the e.s.d. μ_n consists in finding a limit to its Stieltjes transform $m_{\mu_n}(z)=\frac{1}{n}\operatorname{tr}\left(\mathbf{X}-z\mathbf{I}_n\right)^{-1}$. For simpler models (e.g., the Marčenko-Pastur or the semi-circle distributions), the normalized trace of the resolvent $\frac{1}{n}\operatorname{tr}\left(\mathbf{X}-z\mathbf{I}_n\right)^{-1}$ admits a limit but in more involved models, the limit does not exist and one has to resort on other approximations in the limit of growing size. One such approximation known as *deterministic equivalents* is introduced in the following.

3.2.3 Deterministic equivalents and Gaussian methods

Definition 9. *Consider a series of matrices* $\mathbf{A}_1,\mathbf{A}_2,\cdots$ *with* $\mathbf{A}_n\in\mathbb{R}^{n\times n}$ *(functions of symmetric random matrices* $\mathbf{X}_1,\mathbf{X}_2,\cdots$, *with* $\mathbf{X}_n\in\mathbb{R}^{n\times n}$*). A* deterministic equivalent *of* $\mathbf{A}_n$ *is a series* $\mathbf{B}_1,\mathbf{B}_2,\ldots$ *where* $\mathbf{B}_n\in\mathbb{R}^{n\times n}$, *of* deterministic *matrices, such that for a deterministic matric* $\mathbf{C}$ *of bounded spectral norm and deterministic vectors* $\mathbf{a},\mathbf{b}$ *of bounded norms, we have*

$$\begin{aligned} \frac{1}{n}\operatorname{tr}\mathbf{C}\mathbf{A}_n-\frac{1}{n}\operatorname{tr}\mathbf{C}\mathbf{B}_n &\xrightarrow{\text{a.s.}} 0\\ \mathbf{a}^*\left(\mathbf{A}_n-\mathbf{B}_n\right)\mathbf{b} &\xrightarrow{\text{a.s.}} 0. \end{aligned}$$

We will call $\mathbf{B}_n$ *the* deterministic equivalent *of* $\mathbf{A}_n$ *and denote* $\mathbf{A}_n\leftrightarrow\mathbf{B}_n$.

There are two main techniques to find deterministic equivalents: the so-called *Bai and Silverstein technique* and the *Gaussian methods*. Our results being derived using Gaussian methods, we will be restricted to that particular approach. Readers interested in *Bai and Siverstein* approach are referred to Section 6.2 of [Couillet and Debbah, 2011]. While Gaussian methods are particularly designed for random matrix models with Gaussian entries, we use this technique for deriving deterministic equivalent for the e.s.d. of the adjacency matrix in the community detection problem. This is motivated by the *universality* property (see e.g., [Silverstein and Bai, 1995]) stating that for certain random matrices (having entries with zero mean, unit variance and bounded first order moments), the limiting law of their e.s.d. does not change no matter the distribution of the entries of the matrices.

Gaussian techniques rely on two main ingredients:

- an integration by parts formula given in its simpler form

  Lemma 10. *Let x be a standard real Gaussian random variable and $f : \mathbb{R} \to \mathbb{R}$ be a $\mathbb{C}^1$ function with first derivative $f^{'}(x)$ having at most polynomial growth. Then,*

  $$\mathbb{E}[xf(x)] = \mathbb{E}[f^{'}(x)].$$

- the Nash-Poincaré inequality

  Lemma 11. *Let x be a standard real Gaussian random variable and $f : \mathbb{R} \to \mathbb{R}$ be a $\mathbb{C}^1$ function with first derivative $f^{'}(x)$. Then, we have*

  $$\mathrm{Var}[f(x)] \leq \mathbb{E}[|f^{'}(x)|^2].$$

The proofs of those lemma can be found in [Pastur and Shcherbina, 2011].

The *Gaussian technique* approach to evaluate let's say the deterministic equivalent of the normalized trace of the resolvent $\frac{1}{n}\operatorname{tr}(\mathbf{X} - z\mathbf{I}_n)^{-1}$ consists in

- evaluating $\mathbb{E}\left[\frac{1}{n}\operatorname{tr}(\mathbf{X} - z\mathbf{I}_n)^{-1}\right]$ using the integration by part formula. Here, the Nash-Poincaré inequality is also used to bound the small variations.

- Showing that $\frac{1}{n}\operatorname{tr}(\mathbf{X} - z\mathbf{I}_n)^{-1}$ converges almost surely to its expectation computed above, using Nash-Poincaré inequality along with the Borell Cantelli Lemma.

We provide here the main calculation steps for the derivation of deterministic equivalents of the resolvent associated with a Wigner matrix. The more rigorous proofs with full control of the vanishing terms can be found in [Hachem et al., 2007,Pastur and Shcherbina, 2011]. Specifically, let us consider a symmetric matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ having Gaussian i.i.d. entries of zero mean and variance σ^2/n. We want to find a deterministic equivalent $\bar{\mathbf{Q}}$ for

the resolvent matrix $\mathbf{Q} = (\mathbf{X} - z\mathbf{I}_n)^{-1}$ such that for any deterministic bounded vectors $\mathbf{a}$, $\mathbf{b}$, we have that

$$\mathbf{a}^*\mathbf{Q}\mathbf{b} - \mathbf{a}^*\bar{\mathbf{Q}}\mathbf{b} \xrightarrow{\text{a.s.}} 0$$
$$\frac{1}{n}\operatorname{tr}\mathbf{Q} - \frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}} \xrightarrow{\text{a.s.}} 0.$$

As explained above, the strategy of the proof consists in working directly with the expectations of the above resolvent functional since it can be shown using the Nash-Poincaré inequality that those quantities converge towards their expectations. Using the above integration by part formula, we can compute $\mathbb{E}(\mathbf{XQ})_{ij}$ which is related to $\mathbb{E}[\mathbf{Q}]$ by the following

$$\begin{aligned}\mathbb{E}[\mathbf{XQ}] &= \mathbb{E}[\mathbf{X}(\mathbf{X} - z\mathbf{I})^{-1}] \\ &= \mathbb{E}[(\mathbf{X} - z\mathbf{I} + z\mathbf{I})(\mathbf{X} - z\mathbf{I})^{-1}] \\ &= \mathbb{E}[\mathbf{I} + z\mathbf{Q}]. \end{aligned} \tag{3.7}$$

Denoting $X_{il} = \frac{\sigma}{\sqrt{n}} Z_{il}$ with $Z_{il} \sim \mathcal{N}(0,1)$, we can write

$$\mathbb{E}(\mathbf{XQ})_{ij} = \sum_{l=1}^{n} \frac{\sigma}{\sqrt{n}} \mathbb{E}(Z_{il} Q_{lj}).$$

Using the integration by part formula, we have

$$\begin{aligned}\mathbb{E}(Z_{il}Q_{lj}) &= \mathbb{E}\left(\frac{\partial (\mathbf{X} - z\mathbf{I})^{-1}_{lj}}{\partial Z_{il}}\right) \\ &= \mathbb{E}\left(-(\mathbf{X} - z\mathbf{I})^{-1}\frac{\partial \mathbf{X}}{\partial Z_{il}}(\mathbf{X} - z\mathbf{I})^{-1}\right)_{lj} \\ &= \mathbb{E}\left(-(\mathbf{X} - z\mathbf{I})^{-1}\frac{\sigma}{\sqrt{n}}(\mathbf{E}_{il} + \mathbf{E}_{li})(\mathbf{X} - z\mathbf{I})^{-1}\right)_{lj}\end{aligned}$$

where $\mathbf{E}_{il}$ is the matrix with all entries equal to 0 but the entry (i,l) which is equal to 1 and thus

$$\mathbb{E}(\mathbf{XQ})_{ij} = \sum_{l=1}^{n} -\frac{\sigma^2}{n}\left(\mathbb{E}\left[Q_{li}Q_{lj}\right] + \mathbb{E}\left[Q_{ll}Q_{ij}\right]\right).$$

Using Equation (3.7), we then have

$$\mathbb{E}[Q_{ij}] = -\frac{1}{z}\delta_{ij} - \frac{\sigma^2}{z}\frac{1}{n}\mathbb{E}[(Q^2)_{ij}] - \frac{\sigma^2}{z}\frac{1}{n}\mathbb{E}[(\operatorname{tr}\mathbf{Q})Q_{ij}]. \tag{3.8}$$

By applying the Cauchy-Shwartz inequality, we can show that

$$\left|\mathbb{E}[\frac{\operatorname{tr}\mathbf{Q}}{n}Q_{ij}] - \mathbb{E}[\frac{\operatorname{tr}\mathbf{Q}}{n}]\mathbb{E}[Q_{ij}]\right| = \mathcal{O}(n^{-1}). \tag{3.9}$$

To see this, we have

$$\mathbb{E}\left[\left(\frac{\operatorname{tr}\mathbf{Q}}{n}-\mathbb{E}\frac{\operatorname{tr}\mathbf{Q}}{n}\right)(Q_{ij}-\mathbb{E}Q_{ij})\right]\leq\sqrt{Var\left(\frac{\operatorname{tr}\mathbf{Q}}{n}\right)Var\left(Q_{ij}\right)}.$$

Nash-Poincaré inequality allows to show that $Var\left(\frac{\operatorname{tr}\mathbf{Q}}{n}\right)=\mathcal{O}(n^{-2})$ (see [Hachem et al., 2007] for similar results). The result (3.9) is proved using the previous argument along with the fact that the entries of the resolvent matrix are uniformly bounded in the positive complex plane i.e., $\forall i,j$ and $z\in\mathbb{R}$, $|Q_{ij}|\leq\frac{1}{|\Im(z)|}$.

Using (3.9), we can write (3.8) as

$$\left(-z-\sigma^2\mathbb{E}\frac{\operatorname{tr}\mathbf{Q}}{n}\right)\mathbb{E}[Q_{ij}]=\delta_{ij}+\frac{\sigma^2}{n}\mathbb{E}[(Q^2)_{ij}]+o(1). \tag{3.10}$$

For $z\in\mathbb{R}$, we have that $\Im(-z-\mathbb{E}\operatorname{tr}(\frac{\mathbf{Q}}{n}))<-\Im(z)$ and thus $-z-\mathbb{E}\operatorname{tr}(\frac{\mathbf{Q}}{n})$ does not vanish asymptotically. We can thus write (3.8) as

$$\mathbb{E}(Q_{ij})=\frac{\frac{\sigma^2}{n}\mathbb{E}\left[\mathbf{Q}^2\right]_{ij}+\delta_{ij}}{-z-\sigma^2\mathbb{E}\frac{\operatorname{tr}\mathbf{Q}}{n}}+o(1). \tag{3.11}$$

Multiplying Equation (3.11) by $\frac{\sigma^2}{n}$, setting $i=j$ and summing over i, we get

$$\sigma^2\mathbb{E}\frac{\operatorname{tr}\mathbf{Q}}{n}=\frac{\frac{\sigma^4}{n^2}\mathbb{E}\left[\operatorname{tr}\mathbf{Q}^2\right]+\frac{\sigma^2}{n}}{-z-\sigma^2\mathbb{E}\frac{\operatorname{tr}\mathbf{Q}}{n}}+o(1). \tag{3.12}$$

Using again Nash-Poincaré inequality, one can show that $\frac{\sigma^4}{n^2}\mathbb{E}\left[\operatorname{tr}\mathbf{Q}^2\right]=\mathcal{O}(n^{-1})$ (see [Hachem et al., 2007]) and thus we have that

$$\mathbb{E}\frac{\operatorname{tr}\mathbf{Q}}{n}-m(z)\xrightarrow{\text{a.s.}}0$$

where $m(z)$ is the solution, for $z\in\mathbb{R}$ of

$$m(z)=\frac{1}{-z-\sigma^2m(z)}. \tag{3.13}$$

Similarly, for deterministic vectors $\mathbf{a}$ and $\mathbf{b}$ of bounded norms, we have using Equation (3.11)

$$\mathbb{E}\left[\mathbf{a}^*\mathbf{Q}\mathbf{b}\right]=\sum_{ij}a_i\mathbb{E}\left[Q_{ij}\right]b_j \tag{3.14}$$

$$=\mathbf{a}^*(m(z)\mathbf{I})\mathbf{b} \tag{3.15}$$

where again we use the Nash-Poincaré inequality to show that $\sum_{ij}a_i\mathbb{E}\left[\frac{(Q^2)_{ij}}{n}\right]b_j=\mathcal{O}(n^{-1})$. The following result summarizes those calculus to obtain deterministic equivalents of the resolvent matrix $\mathbf{Q}$.

Theorem 12 (Deterministic equivalents for Wigner matrices)**.** *Let* $\mathbf{Q} = (\mathbf{X} - z\mathbf{I}_n)^{-1}$. *Then, for all* $z \in \mathbb{R}$,

$$\mathbf{Q} \leftrightarrow \bar{\mathbf{Q}} = m(z)\mathbf{I}_n \tag{3.16}$$

where $m(z)$ *is the unique solution in* $\mathbb{R}$ *for* $z \in \mathbb{R}$, *of* $m(z) = \frac{1}{-z-\sigma^2 m(z)}$.

Theorem 12 can be used to show the convergence of the limiting eigenvalue distribution of Wigner matrices towards the semi-circle law. From Theorem 12, $m(z)$ is the limiting value of the Stieltjes transform $\frac{1}{n}\operatorname{tr}\mathbf{Q}$ of the e.s.d. From Equation (3.13), we can see that the solution to the quadratic equation satisfying the condition $\Im m(z)\Im z \geq 0$, $\Im z \neq 0$ (property of a Stieltjes transform) is unique and verifies

$$m(z) = \begin{cases} \frac{1}{2\sigma^2}\left(\sqrt{z^2-4\sigma^2} - z\right), & \mathcal{R}(z) > 0 \\ \frac{1}{2\sigma^2}\left(-\sqrt{z^2-4\sigma^2} - z\right), & \mathcal{R}(z) < 0 \end{cases} \tag{3.17}$$

where $\sqrt{z^2-4\sigma^2}$ is the branch that has the asymptotic behavior

$$\sqrt{z^2-4\sigma^2} = z + \mathcal{O}(|z|^{-1})$$

as $z \to \infty$.

Applying the inverse formula of the Stieltjes transform (Theorem 6), we get the semi-circle law

$$\mu(x) = \frac{2}{\pi\sigma^2}\sqrt{(\sigma^2 - x^2)^+}.$$

3.2.4 Spiked random matrices

Generally speaking, spiked random matrices are low-rank perturbations of large random matrices which can be seen as information+noise models, and are popular in many applications such as wireless communications, signal processing, statistics and machine learning. Most commonly used spiked random matrix models include

- **The perturbed sample covariance matrix** of the type $\mathbf{T}_n^{\frac{1}{2}}\mathbf{X}_n\mathbf{X}_n^\mathsf{T}\mathbf{T}_n^{\frac{1}{2}}$, with $\mathbf{X}_n \in \mathbb{R}^{n\times n}$ having random i.i.d entries of zero mean and variance $1/n$ and $\mathbf{T}_n = \mathbf{I}_n + \sum_{i=1}^r \omega_i \mathbf{v}_i\mathbf{v}_i^\mathsf{T}$ a perturbation of the identity matrix with the finite low-rank matrix $\sum_{i=1}^r \omega_i \mathbf{v}_i\mathbf{v}_i^\mathsf{T}$.
- **The additive model** of the type $(\mathbf{X}_n + \sum_{i=1}^r \omega_i \mathbf{v}_i\mathbf{v}_i^\mathsf{T})$ with $\mathbf{X}_n \in \mathbb{R}^{n\times n}$ a symmetric matrix having random i.i.d entries of zero mean and variance $1/n$.

As will become clear in the subsequent chapters, the additive model is the variant that appears to correspond to the asymptotic equivalent of the similarity matrices involved

in the clustering applications. In what follows, we provide asymptotic results of the eigenvalues and eigenvectors for elementary additive spiked models.

Let us consider the model $\mathbf{Y} = \frac{\mathbf{X}}{\sqrt{n}} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$ where $\mathbf{X}$ is a symmetric random matrix with entries of zero mean and unit variance, $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_k]$ with $\mathbf{v}_i$ unit norm eigenvectors, $\mathbf{\Omega} = \text{diag}\,(\omega_1, \cdots, \omega_k)$ with ω_i the eigenvalue associated with $\mathbf{v}_i$. We are interested in the asymptotic localization of the eigenvalues of $\mathbf{Y}$. We know from Theorem 2 that the e.s.d. of $\mathbf{Y}$ converges weakly to the limiting law of $\frac{\mathbf{X}}{\sqrt{n}}$ (which is the semi-circle law) since $\mathbf{Y}$ and $\mathbf{X}$ only differ by a low-rank matrix. So, most of the eigenvalues of $\mathbf{Y}$ fall within the support $\mathcal{S} = [-2, 2]$ of the semi-circle law. However some eigenvalues of $\mathbf{Y}$ might isolate from the semi-circle support due to the low-rank matrix $\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$. We follow classical random matrix approaches for the study of the eigenvalues and eigenvectors of spiked random matrices ([Benaych-Georges and Nadakuditi, 2012]). Knowing the localization of most of the eigenvalues of $\mathbf{Y}$ (in the support $\mathcal{S} = [-2, 2]$), we will now determine the position of the possibly remaining eigenvalues. We thus need to solve for large n and $\rho \notin \mathcal{S}$

$$0 = \det\left(\frac{\mathbf{X}}{\sqrt{n}} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T} - \rho\mathbf{I}_n\right). \tag{3.18}$$

Following ideas from [Bai and Silverstein, 1998], in addition to the convergence of the eigenvalues of Wigner matrices towards the semi-circle law, one can show a "No eigenvalue outside the support" result meaning that the random matrix $\frac{\mathbf{X}}{\sqrt{n}}$ does not have asymptotically eigenvalues outside the range $\mathcal{S} = [-2, 2]$. With this in mind, $\det\left(\frac{\mathbf{X}}{\sqrt{n}} - \rho\mathbf{I}_n\right)$ does not vanish asymptotically for $\rho \notin \mathcal{S}$ and thus Equation (3.18) is equivalent to

$$0 = \det\left(\mathbf{I}_n + \mathbf{Q}\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}\right) = \det\left(\mathbf{I}_k + \mathbf{V}^\mathsf{T}\mathbf{Q}\mathbf{V}\mathbf{\Omega} - \rho\mathbf{I}_n\right) \tag{3.19}$$

with $\mathbf{Q} = \left(\frac{\mathbf{X}}{\sqrt{n}} - \rho\mathbf{I}_n\right)^{-1}$. We can then use here our results on the deterministic equivalents to provide a deterministic limit of the quantity $\mathbf{V}^\mathsf{T}\mathbf{Q}\mathbf{V}\mathbf{\Omega}$. The matrix $\mathbf{V}$ being deterministic, we can readily apply Theorem 12 and we get

$$\mathbf{V}^\mathsf{T}\mathbf{Q}\mathbf{V}\mathbf{\Omega} \xrightarrow{\text{a.s.}} m(\rho)\mathbf{\Omega}. \tag{3.20}$$

Equation (3.18) is thus asymptotically equivalent to

$$0 = \prod_{i=1}^{k}(1 + \omega_i m(\rho)). \tag{3.21}$$

There thus exists an isolated eigenvalue of $\mathbf{Y}$ outside $\mathcal{S}$ when there exists ω_i such that the equation $1 + \omega_i m(\rho) = 0$ admits a solution for $\rho \notin [-2, 2]$. From the plot of the function $m(z)$ (Figure 3.5) in its defined domain, we see that a solution to the previous equation exists when $\frac{1}{\omega_i} < -\lim_{z\downarrow 2} m(z)$ or $\frac{1}{\omega_i} > -\lim_{z\uparrow -2} m(z)$ and in that case the limiting isolated eigenvalue ρ is equal to $m^{-1}\left(-\frac{1}{\omega_i}\right)$ [2]. Using the formula for $m(z)$ in

[2] $m^{-1}(z) = a \Leftrightarrow z = m(a)$

Equation (3.17), one has an explicit expression for the limiting isolated eigenvalue ρ. The following result provides the asymptotic localization of the eigenvalues of the spiked random matrix $\mathbf{Y}$.

Theorem 13. *Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be a symmetric random matrix having i.i.d. entries of zero mean and unit variance, for which the e.s.d. $\mu^{\mathbf{X}}$ converges almost surely toward the semi-circle law with compact support $\mathcal{S} = [-2, 2]$. Consider also a rank-k perturbation matrix $\mathbf{V\Omega V}^{\mathsf{T}}$ with ordered eigenvalues $\omega_1 \geq \cdots \geq \omega_k$. Denote $\mathbf{Y}$ the matrix defined as $\mathbf{Y} = \mathbf{X} + \mathbf{V\Omega V}^{\mathsf{T}}$ with ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Then, as n grows large, for $i = 1, \cdots, k$*

- *If $\omega_i > 1$ or $\omega_i < -1$,*

$$\lambda_i \xrightarrow{\text{a.s.}} \frac{1+\omega_i^2}{\omega_i} \tag{3.22}$$

- *If $\omega_i \in [-1, 0]$,*

$$\lambda_i \xrightarrow{\text{a.s.}} -2 \tag{3.23}$$

- *If $\omega_i \in [0, 1]$,*

$$\lambda_i \xrightarrow{\text{a.s.}} 2. \tag{3.24}$$

Theorem 13 asserts that there exists a phase transition phenomenon by which the appearance of an eigenvalue of $\mathbf{Y}$ outside the main support $\mathcal{S}$ of the noise matrix $\mathbf{X}$ depends on the value of the eigenvalues of the low-rank matrix $\mathbf{V\Omega V}^{\mathsf{T}}$ compared with a certain threshold. As stated in the result below, when this phase transition occurs (i.e., an eigenvalue of $\mathbf{Y}$ appears outside the main support $\mathcal{S}$), the associated eigenvector gets correlated to some extent to the eigenvector $\mathbf{v}_i$ associated with the eigenvalue ω_i of the low-rank matrix while there is a zero correlation when the phase transition does not occur.

Theorem 14. *Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be a symmetric random matrix having i.i.d. entries of zero mean and unit variance, for which the e.s.d. $\mu^{\mathbf{X}}$ converges almost surely towards the semi-circle law with compact support $\mathcal{S} = [-2, 2]$. Consider also a rank-k perturbation matrix $\sum_{i=1}^{k} \omega_i \mathbf{v}_i \mathbf{v}_i^{\mathsf{T}}$ with ordered eigenvalues $\omega_1 \geq \cdots \geq \omega_k > 0$. Denote $\mathbf{Y}$ the matrix defined as $\mathbf{Y} = \mathbf{X} + \sum_{i=1}^{k} \omega_i \mathbf{v}_i \mathbf{v}_i^{\mathsf{T}}$ with ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Then, as n grows large, for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ deterministic vectors and $\mathbf{u}_i$ eigenvector of $\mathbf{Y}$ associated with eigenvalue λ_i,*

$$\mathbf{a}^* \mathbf{u}_i \mathbf{u}_i^* \mathbf{b} - \frac{\omega_i^2 - 1}{\omega_i^2} \mathbf{a}^* \mathbf{v}_i \mathbf{v}_i^* \mathbf{b} \cdot 1_{w_i > 1} \xrightarrow{\text{a.s.}} 0. \tag{3.25}$$

In particular

$$|\mathbf{v}_i^* \mathbf{u}_i|^2 \xrightarrow{\text{a.s.}} \frac{\omega_i^2 - 1}{\omega_i^2} \cdot 1_{w_i > 1}. \tag{3.26}$$

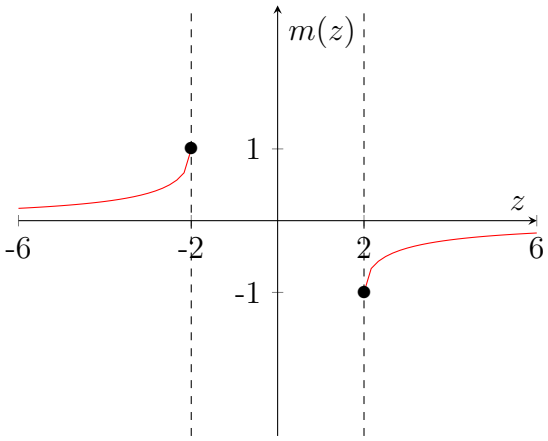


Figure 3.5: Representation of $m(z)$

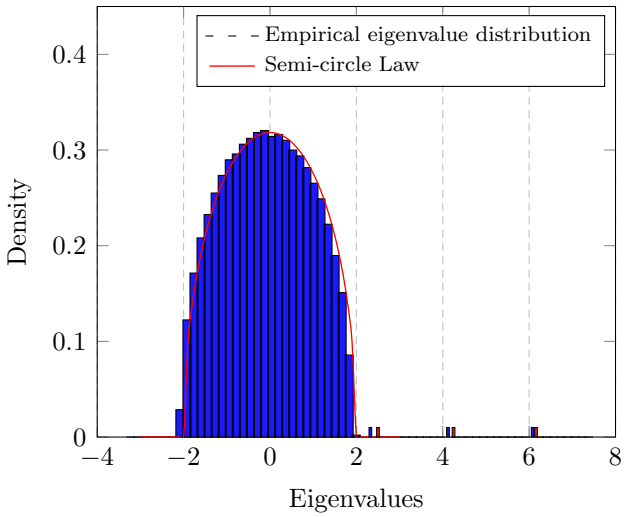


Figure 3.6: Histogram of the eigenvalues of $\mathbf{Y} = \mathbf{X} + \sum_{i=1}^{3} \omega_i \mathbf{j}_i \mathbf{j}_i^\mathsf{T}$, $X_{ij} \sim \mathcal{N}(0, 1/n)$, $\omega_1 = 2, \omega_2 = 4, \omega_3 = 6$ versus theoretical law in red, for $n = 3000$.

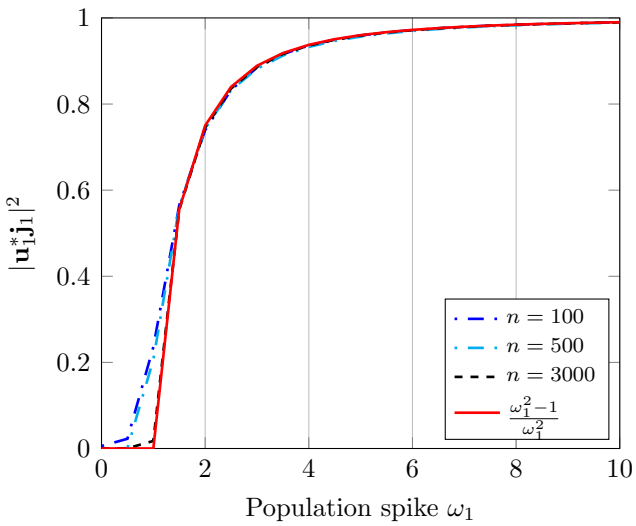


Figure 3.7: Simulated versus limiting $|\mathbf{u}_1^*\mathbf{j}_1|^2$ for $\mathbf{Y} = \mathbf{X} + \omega_1 \mathbf{j}_1\mathbf{j}_1^\mathsf{T}$, $\mathbf{j}_1 = \frac{2}{\sqrt{n}}[\mathbf{1}_{n/2}, \mathbf{0}_{n/2}]$, $X_{ij} \sim \mathcal{N}(0, 1/n)$, varying ω_1.

Proof of Theorem 14. The proof of Theorem 14 relies on a Cauchy integration approach. For $\mathbf{u}_i$ an eigenvector of $\mathbf{Y}$ associated with an isolated eigenvalue λ_i (i.e., when $\omega_i > 1$), we have from the Cauchy integration formula

$$\mathbf{a}^*\mathbf{u}_i\mathbf{u}_i^*\mathbf{b} = \frac{1}{2\pi i}\oint_{\Gamma_i} \mathbf{a}^* \left(\mathbf{Y} - z\mathbf{I}_n\right)^{-1} \mathbf{b} dz \tag{3.27}$$

for large n almost surely, where Γ_i is a complex (positively oriented) contour circling around the eigenvalue λ_i only. In order to work out the previous integral, we need to find a deterministic equivalent to the integrand. Applying the Woodburry identity [3] to $\left(\mathbf{Y} - z\mathbf{I}_n\right)^{-1}$, for $\mathbf{Q} = \left(\mathbf{X} - z\mathbf{I}_n\right)^{-1}$, we have

$$\mathbf{a}^* \left(\mathbf{X} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T} - z\mathbf{I}_n\right)^{-1} \mathbf{b} = \mathbf{a}^* \left[\mathbf{Q} - \mathbf{Q}\mathbf{V}\mathbf{\Omega}\left(\mathbf{I}_k + \mathbf{V}^\mathsf{T}\mathbf{Q}\mathbf{V}\mathbf{\Omega}\right)^{-1}\mathbf{V}^\mathsf{T}\mathbf{Q}\right]\mathbf{b} \tag{3.28}$$

Using the deterministic equivalent of the resolvent $\mathbf{Q}$ (Theorem 12), we have for large n,

$$\mathbf{a}^* \left(\mathbf{X} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T} - z\mathbf{I}_n\right)^{-1} \mathbf{b} \xrightarrow{\text{a.s.}} m(z)\mathbf{a}^*\mathbf{b} - \mathbf{a}^*\mathbf{V}\mathcal{D}\left(\frac{\omega_j m^2(z)}{1+\omega_j m(z)}\right)_{j=1}^{k} \mathbf{V}^\mathsf{T}\mathbf{b} \tag{3.29}$$

The first term in Equation (3.29) has asymptotically no residue inside the contour Γ_i while only the diagonal term $\frac{\omega_i m^2(z)}{1+\omega_i m(z)}$ survives in the residue calculus of the second right-hand term. We thus obtain

$$\mathbf{a}^*\mathbf{u}_i\mathbf{u}_i^*\mathbf{b} \xrightarrow{\text{a.s.}} -\lim_{z\to\lambda_i}(z-\lambda_i)\frac{\omega_i m^2(z)}{1+\omega_i m(z)}\mathbf{a}^*\mathbf{v}_i\mathbf{v}_i^\mathsf{T}\mathbf{b}. \tag{3.30}$$

[3] For invertible matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and $\mathbf{D}$, the following equality holds: $(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}+\mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}$

Since from Theorem 13, in the limit $z \to \lambda_i$, $m(z) = -\frac{1}{\omega_i}$, the numerator and the denominator of the Equation (3.30) are vanishing in the limit. By applying the l'Hopital rule, we get

$$\mathbf{a}^* \mathbf{u}_i \mathbf{u}_i^* \mathbf{b} \xrightarrow{\text{a.s.}} - \lim_{z \to \lambda_i} \frac{m^2(z)}{m'(z)} \mathbf{a}^* \mathbf{v}_i \mathbf{v}_i^\mathsf{T} \mathbf{b} \tag{3.31}$$

where $m'(z)$ is the first derivative of $m(z)$. We know from Theorem 13 that in the limit $z \to \lambda_i$, $m(z) \to -\frac{1}{\omega_i}$ and we can easily show from there that $m'(z) \to \frac{1}{1-\omega_i^2}$. This concludes the proof. □

Notice from Theorem 14 that for sufficiently large informative eigenvalue ω_i (say $\omega_i \to \infty$ or equivalently $\lambda_i \to \infty$,) the alignment between the eigenvector $\mathbf{u}_i$ of $\mathbf{Y}$ and the eigenvector $\mathbf{v}_i$ of the low-rank deterministic matrix, converges asymptotically to 1 meaning a perfect alignment between those two vectors. In contrast, under the phase transition i.e. for eigenvalue $\omega_i \in [-1, 1]$, the alignment between the two eigenvectors vanishes asymptotically. This has important applications in spectral clustering where the similarity matrix of the data will be shown to be of the form of the additive spike random matrix $\mathbf{Y}$ where the eigenvectors of the low-rank matrix will be proportional to the canonical vectors of the classes (i.e., with zero entries except in the support of the class) and thus the eigenvectors of $\mathbf{Y}$ (which are in fact used for spectral clustering) will be correlated to the class information the more they are far apart from the bulk (the limiting eigenvalue distribution of the noise matrix $\mathbf{X}$.)

These results on the additive spiked models are the basis for the understanding of the behavior of spectral clustering algorithms. As stated above, the intuition behind spectral clustering stems from the fact that the eigenvectors associated with the extreme eigenvalues of the data similarity matrices are correlated to some extent to the canonical class vectors that we denote $\mathbf{j}_a = 1_{i \in \mathcal{C}_a}$ (vectors of size n the number of nodes with values equal to 1 in the support of class $\mathcal{C}_a$ and zero otherwise). More precisely, denoting $\mathbf{Y}$ a given random similarity matrix associated with data grouped into k classes, it generally takes the form

$$\mathbf{Y} \propto \mathbf{J}\mathbf{J}^\mathsf{T} + \mathbf{X} \tag{3.32}$$

where $\mathbf{J} = [\mathbf{j}_1, \cdots, \mathbf{j}_k]$ and $\mathbf{X} = (\mathbf{Y} - \mathbb{E}[\mathbf{Y}])$ a zero-mean random matrix.

Equation (3.32) takes the form of the aforementioned additive spike model and thus when the phase transition occurs, the eigenvectors associated with the extreme eigenvalues of $\mathbf{Y}$ have some non-zero correlation with the eigenspace of $\mathbf{J}\mathbf{J}^\mathsf{T}$ and thus some class information can be recovered from those eigenvectors. Alternatively, in the absence of the phase transition occurence, the eigenvectors associated with the extreme eigenvalues are uncorrelated to the class eigenspace $\mathbf{J}\mathbf{J}^\mathsf{T}$ and no class information can be retrieved from those. However, the aforementioned projection between the eigenspaces of $\mathbf{Y}$ and $\mathbf{J}\mathbf{J}^\mathsf{T}$ (Theorem 14) is not enough to characterize the performances of spectral clustering algorithms. To get a finer analysis, [Couillet and Benaych-Georges, 2016] proposed to

write the dominant eigenvectors as "noisy" linear combinations of the canonical vectors $\mathbf{j}_a$ as

$$\mathbf{u}_i = \sum_{a=1}^{k} \alpha_i^a \frac{\mathbf{j}_a}{\sqrt{n_a}} + \sigma_i^a \mathbf{w}_i^a \tag{3.33}$$

with $\mathbf{j}_a \in \mathbb{R}^n$ canonical vector of class $\mathcal{C}_a$, $\mathbf{w}_i^a$ noise orthogonal to $\mathbf{j}_a$, and

$$\alpha_i^a = \frac{1}{\sqrt{n_a}} \mathbf{u}_i^\mathsf{T} \mathbf{j}_a \tag{3.34}$$

$$(\sigma_i^a)^2 = \left\| \mathbf{u}_i - \alpha_i^a \frac{\mathbf{j}_a}{\sqrt{n_a}} \right\|^2 = \mathbf{u}_i^\mathsf{T} \mathbf{j}_a \mathbf{u}_i - (\alpha_i^a)^2. \tag{3.35}$$

The approach used in [Couillet and Benaych-Georges, 2016] to estimate α_i^a consists in first finding an estimate for $\frac{1}{n_a} \mathbf{j}_a^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{j}_a$ which is exactly $(\alpha_i^a)^2$. To this end, a contour integral approach is used as in Theorem 14. Namely by residue calculus, we have that

$$\frac{1}{n_a} \mathbf{j}_a^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{j}_a = \frac{1}{2\pi i} \oint_{\Gamma_i} \frac{1}{n_a} \mathbf{j}_a^\mathsf{T} \left(\mathbf{Y} - z \mathbf{I}_n \right)^{-1} \mathbf{j}_a \mathrm{d}z \tag{3.36}$$

for large n almost surely, where Γ_i is a complex (positively oriented) contour circling arount the limiting isolated eigenvalue (outside the main support) associated with the eigenvector $\mathbf{u}_i$ of $\mathbf{Y}$. Using then the Woodburry identity, we have

$$\frac{1}{n_a} \mathbf{j}_a^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{j}_a \propto \frac{1}{2\pi i} \oint_{\Gamma_a} \frac{1}{n_a} \mathbf{j}_a^\mathsf{T} \left(\mathbf{X} + \mathbf{J}\mathbf{J}^\mathsf{T} - z I_n \right)^{-1} \mathbf{j}_a \mathrm{d}z \tag{3.37}$$

$$= \frac{1}{2\pi i} \oint_{\Gamma_a} \frac{1}{n_a} \mathbf{j}_a^\mathsf{T} \mathbf{Q} \mathbf{j}_a \mathrm{d}z + \frac{1}{2\pi i} \oint_{\Gamma_a} \frac{1}{n_a} \mathbf{j}_a^\mathsf{T} \mathbf{Q} \mathbf{J} \left(\mathbf{I} + \mathbf{J}^\mathsf{T} \mathbf{Q} \mathbf{J} \right)^{-1} \mathbf{J}^\mathsf{T} \mathbf{j}_a \mathrm{d}z \tag{3.38}$$

with $\mathbf{Q} = (\mathbf{X} - z\mathbf{I}_n)^{-1}$ the resolvent of the matrix $\mathbf{X}$. Contour integration on the first term of (3.38) gives 0 asymptotically since from the *No eigenvalues outside the support* results, there is asymptotically no eigenvalues of $\mathbf{X}$ outside its main support. The integrand of second term can be computed using standard calculus along with deterministic approximations of quantities of the type $\mathbf{a}^\mathsf{T} \mathbf{Q} \mathbf{b}$ for deterministic vectors of bounded norms $\mathbf{a}$ and $\mathbf{b}$ (see deterministic equivalents of functionals of the resolvent in Section 3.2.3). A final residue calculus on the obtained approximation allows to find deterministic approximation of $(\alpha_i^a)^2$ in the limit of large n.

The estimation of $(\sigma_i^a)^2$ follows similar arguments as the ones for $(\alpha_i^a)^2$. Namely, one needs an estimate for the quantity $\mathbf{u}_i^\mathsf{T} \mathbf{j}_a \mathbf{u}_i$ which can be obtained by dividing any entry, say (a, b) of the more involved object $\frac{1}{n} \mathbf{J}^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{D}_a \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{J}$ by α_i^a and α_i^b where $D_a = \mathrm{diag}(\mathbf{j}_a)$. The estimation of $\frac{1}{n} \mathbf{J}^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{D}_a \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{J}$ is done using the similar contour integration formula as in (3.36). Namely,

$$\frac{1}{n} \mathbf{J}^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{D}_a \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{J} = \frac{1}{2\pi i} \oint_{\Gamma_a} \frac{1}{n} \mathbf{J}^\mathsf{T} \left(\mathbf{Y} - z \mathbf{I}_n \right)^{-1} \mathbf{D}_a \left(\mathbf{Y} - \tilde{z} \mathbf{I}_n \right)^{-1} \mathbf{J} \mathrm{d}z \mathrm{d}\tilde{z}. \tag{3.39}$$

Detailed derivations of those estimators on more involved models are provided in the Appendix for specific application in graphs community detection and kernel spectral clustering.

3.3 Bayesian inference

We briefly introduce in this section variational Bayes approaches which are used to find approximation to the intractable posterior distribution in most applications. In general, the maximization of the posterior distribution of the model in use can be shown to be equivalent to an optimization problem involving some variational variables, the solutions of which take closed form expressions in some parameteric family of distributions. Such distributions are often members of the exponential family which covers most of the probabilistic models used in statistics and machine learning. The first sub-section introduces the exponential family and the second sub-section provides the mean-field variational inference methodology to solve the maximum-a-posterior problem involving likelihoods belonging to the exponential family.

3.3.1 Exponential family of distributions

Consider a set of variables x living in a fixed domain $\mathcal{X}$ and a set of parameters θ living in a domain Θ. An exponential family is defined as a collection of parametric distributions that can be written in the form

$$f(x|\theta) = h(x)\exp(T(x)\cdot\eta(\theta)) \tag{3.40}$$

for $x \in \mathcal{X}$ and $\theta \in \Theta$ where h, T, η are some fixed functions. The function $T(x)$ is called the sufficient statistic and the function $\eta(\theta)$ is called the natural parameter. The different distributions are distinct from the form of their functions h, T, η. Most commonly used distributions belonging to the exponential family are the normal, exponential, gamma, log-normal, Pareto binomial, multinomial, Poisson, Beta distributions.

An important feature of exponential family is that they have explicit conjugate prior distributions π given by

$$\pi(\theta) = \frac{1}{Z(\theta)}\exp(\tau\cdot\eta(\theta)), \tag{3.41}$$

where τ are the hyperparameters of the prior.

Under the prior distribution $\pi(\cdot|\tau)$, it is well known that the expectation of the natural parameter $\eta(\theta)$ is given by

$$\mathbb{E}_\pi\eta(\theta) = \frac{\partial \log Z(\tau)}{\partial\tau}, \tag{3.42}$$

a property that is useful in mean-field variational inference computations.

3.3.2 Mean field variational Bayes inference

Many statistical models are described by a set of observations $x_1, \cdots, x_n$, parametrized by a set θ of parameters and the goal is to infer some hidden variables $\mathbf{w} \triangleq \{w_1, \cdots, w_n\}$

involved in the structure of the observations. For this class of problems, a Bayesian approach would consist in finding the hidden variables $\mathbf{w}$ maximizing the posterior distribution $p(\mathbf{w}|x_1, \cdots, x_n, \theta)$. One such model is used in community detection problems that we describe as follows. The basic statistical model for generating graphs structured into communities is the so-called *Stochastic Block Model* (SBM) that we describe from a Bayesian perspective. Given some latent community label $g_i \in \{1, \cdots, k\}$ (with k denoting the number of communities) of each vertex i ($1 \leq i \leq n$) and a community-wise connectivity matrix $\mathbf{C} \in \mathbb{R}^{k \times k}$, an edge is placed between two vertices i and j with an adjacency weight A_{ij} such that

$$\mathbb{P}(A_{ij}|g_i, g_j, C_{g_i,g_j}) \propto \exp\left\{T(A_{ij})\eta(C_{g_i g_j})\right\}.$$

with T the sufficient statistic function defining the graph likelihood distribution and η the corresponding natural parameter function. Following a Bayesian approach, prior distributions are attributed to the labels g_i and the community-wise connectivity matrix $\mathbf{C}$. So here, the observations are the adjacency entries $A_{11}, \cdots, A_{nn}$, the latent variables are the community labels g_i's and the entries of the connectivity matrix $\mathbf{C}$, while the parameters are given by the priors assigned to g_i and $\mathbf{C}$. With this example in mind, which will be the focus of Chapter 5 in its multi-graph model form, let us get back to the general Bayesian statistical models.

In most of those models, the posterior distribution is intractable to compute and is not available in closed form. As an alternative, approximation techniques are used. Markov Chain Monte Carlo (MCMC) family of methods [Gamerman and Lopes, 2006] are such approximations. MCMC generally build samples from a Markov Chain in such a way that its stationary distribution is close to the posterior of interest. However, MCMCs might suffer from large computations and convergence issues especially when used on large datasets. Variational methods [Opper and Saad, 2001, Wainwright et al., 2008] provide a deterministic approximation to the posterior by solving an optimization problem consisting in minimizing the distance (with a well-chosen metric) between the true distribution and the chosen variational distribution. The variational distribution is generally chosen in such a way that the aforementioned optimization is tractable and fast. Most well known variational methods include Belief Propagation [Yedidia et al., 2003] based techniques, expectation propagation [Minka, 2001] and mean field approaches, with the latter being computationally faster but less precise than the former. In Chapter 5, we adopt a mean field variational approach for the multi-layer community detection problem with specific constraints on the communities between the different layers. We thus describe in what follows the mean field variational approach to general Bayesian statistical models.

We consider the problem of approximating the posterior distribution $p(\mathbf{w}|x_1, \cdots, x_n, \theta)$ with a variational distribution $q(\mathbf{w})$. We can decompose the marginal of the log of the

observed data as follows

$$\begin{aligned}
\log \mathbb{P}(x_1, \cdots, x_n) &= \int_{\mathbf{w}} q(\mathbf{w}) \mathrm{d}\mathbf{w} \log \mathbb{P}(x_1, \cdots, x_n) \\
&= \int_{\mathbf{w}} q(\mathbf{w}) \mathrm{d}\mathbf{w} \log \frac{\mathbb{P}(x_1, \cdots, x_n, \mathbf{w})}{\mathbb{P}(\mathbf{w}|x_1, \cdots, x_n)} \mathrm{d}\mathbf{w} \\
&= \int_{\mathbf{w}} q(\mathbf{w}) \mathrm{d}\mathbf{w} \log \frac{\mathbb{P}(x_1, \cdots, x_n, \mathbf{w})}{q(\mathbf{w})} \mathrm{d}\mathbf{w} - \int_{\mathbf{w}} q(\mathbf{w}) \mathrm{d}\mathbf{w} \log \frac{\mathbb{P}(\mathbf{w}|x_1, \cdots, x_n)}{q(\mathbf{w})} \mathrm{d}\mathbf{w} \\
&= \mathbb{E}_q \log \mathbb{P}(x_1, \cdots, x_n, \mathbf{w}) - \mathbb{E}_q \log q(\mathbf{w}) + D_{KL}(q(\mathbf{w})||\mathbb{P}(\mathbf{w}|x_1, \cdots, x_n)) \\
&= \underbrace{\mathbb{E}_q \log \mathbb{P}(x_1, \cdots, x_n|\mathbf{w}) + \mathbb{E}_q \log \frac{\mathbb{P}(\mathbf{w})}{q(\mathbf{w})}}_{\mathcal{G}(q)} + D_{KL}(q(\mathbf{w})||\mathbb{P}(\mathbf{w}|x_1, \cdots, x_n))
\end{aligned}$$

where $D_{KL}(q(\mathbf{w})||\mathbb{P}(\mathbf{w}|x_1, \cdots, x_n))$ is the Kullback-Leibler (KL) divergence between the variational approximation $q(\mathbf{w})$ and the true posterior $\mathbb{P}(\mathbf{w}|x_1, \cdots, x_n)$ and $\mathbb{P}(\mathbf{w})$ is the prior assigned to the latent variables (parametrized by θ). Since $D_{KL}(q(\mathbf{w})||\mathbb{P}(\mathbf{w}|x_1, \cdots, x_n)) \geq 0$, the above defined function $\mathcal{G}(q(\mathbf{w}))$ is a lower-bound of the observed data log-marginal. The variational approach consists then in choosing $q(\mathbf{w})$ maximizing $\mathcal{G}(q(\mathbf{w}))$ such that the KL divergence between the true posterior and its approximate is minimized. *Mean field variational Bayes* technique assumes $q(\mathbf{w})$ to be factorized over the latent variables i.e., the variables w_i are independent over the distribution $q(\mathbf{w})$. With this choice, the function $\mathcal{G}(q(\mathbf{w}))$ is generally easily computable and optimized for most used families of likelihood distributions and in particular exponential family.

Chapter 4

Improved Spectral community detection in heterogeneous single-layer graphs

4.1 Introduction

We are interested in spectral methods (Algorithm 2) for community detection in *dense* realistic graph models. The DCSBM described in Chapter 2 is such a model as it takes into account the degree heterogeneity of the nodes within classes, this being an important feature of real-world networks. Under the DCSBM, the classical spectral method (Algorithm 2) might fail as the dominant eigenvectors entries of the similarity matrix are not uniform within the support of each community. To overcome this, previous works have proposed different normalization techniques of the similarity matrix or of its dominant eigenvectors. The authors in [Coja-Oghlan and Lanka, 2009, Gulikers et al., 2015] have considered performing spectral clustering based on the dominant eigenvectors of $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}$ where $\mathbf{D}$ is the diagonal matrix containing the nodes' degrees on the diagonal and $\mathbf{A}$ is the adjacency matrix. This operation on the adjacency matrix can be seen as regularizing the nodes' connections in such a way that the eigenvector associated with $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}$ will look like *homogeneous* step vectors (with each step representing a class support instead of multiple steps within the class support). Alternatively, as proposed in [Jin et al., 2015], one might regularize instead the eigenvectors of $\mathbf{A}$ by performing spectral clustering on those eigenvectors premultiplied by $\mathbf{D}^{-1}$ to ensure the homogeneity of the eigenvectors' entries within class supports. Other methods [Qin and Rohe, 2013] consider using the eigenvectors of the well-known Laplacian matrix $\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$. Among all those methods, it is unclear which one performs better on given graphs. One might look into a general similarity matrix $\mathbf{D}^{-\alpha}\mathbf{A}\mathbf{D}^{-\alpha}$ and find the best α for a particular graph.

In this work, we propose to study similarity matrices of the form $\mathbf{L}_\alpha \propto \mathbf{D}^{-\alpha}[\mathbf{A} -$

$\frac{\mathbf{d}\mathbf{d}^\intercal}{\mathbf{d}^\intercal\mathbf{1}_n}]\mathbf{D}^{-\alpha}$ where we use $\mathbf{A}-\frac{\mathbf{d}\mathbf{d}^\intercal}{\mathbf{d}^\intercal\mathbf{1}_n}$ the so-called modularity matrix [1] instead of the adjacency matrix, the latter being directly related to the most used cost-function (modularity) in community detection tasks. It turns out that the leading eigenvector of the adjacency matrix $\mathbf{A}$ (a noisy constant vector) does not provide any information about the classes and thus spectral clustering based on the adjacency matrix use the dominant eigenvectors but the largest. On the other hand, the term $\frac{\mathbf{d}\mathbf{d}^\intercal}{\mathbf{d}^\intercal\mathbf{1}_n}$ corresponds to a noisy eigenspace (does not contain class information) and thus considering spectral clustering with the dominant eigenvectors of the modularity matrix $\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\intercal}{\mathbf{d}^\intercal\mathbf{1}_n}$ is equivalent to the spectral clustering using the adjacency matrix $\mathbf{A}$ (excluding the largest eigenvector). In graphs where the communities are well separated, methods based on $\mathbf{L}_\alpha$ for different values of α all achieve asymptotically perfect reconstruction. In order to compare the different methods and find an optimal one, we place ourselves in a setting where community detection is difficult. We will show that, for a DCSBM in the regime under study, the matrices $\mathbf{L}_\alpha$ are equivalent to an additive spiked random matrix (introduced in Section 3.2.4), and thus there is a phase transition beyond which the information can be retrieved (here the classes). We analyze the eigenvalues of this spiked random matrix from which we characterize exactly the phase transition, and we take the α for which the phase transition is maximally achieved thus allowing to best recover the classes in difficult settings. In addition, we provide a consistent estimate of α only based on the observed degrees. We also provide a characterization of the dominant eigenvectors of the spiked random matrix (as described in Section 3.2.4) from which we propose better initialization points to the EM algorithm in the last step of the spectral method.

4.2 Preliminaries

This section recalls the network model under study, which is based on the DCSBM defined in Chapter 2, and provides preliminary technical results.

We consider an n-node random graph with k classes $\mathcal{C}_1, \dots, \mathcal{C}_k$ of sizes $|\mathcal{C}_a| = n_a$. Each node is characterized by an intrinsic connexion weight q_i which affects the probability that this node gets attached to another node in the graph. A null model would consider that the existence of an edge between i and j has probability $q_i q_j$. In order to take into account the membership of the nodes to some group, we define $\mathbf{C} \in \mathbb{R}^{k\times k}$ as a matrix of class weights C_{ab}, independent of the q_i's, affecting the connection probability between nodes in $\mathcal{C}_a$ and nodes in $\mathcal{C}_b$. Following [Karrer and Newman, 2011], conditioned to the knowledge of the q_i's and the the C_{ab}'s, the adjacency matrix $\mathbf{A}$ of the graph generated from a DCSBM model has independent entries (up to symmetry) which are Bernoulli random variables with parameter $P_{ij} = q_i q_j C_{g_i g_j} \in (0,1)$ where g_i is the group assignment of node i. We set $A_{ii} = 0$ for all i. For convenience of exposition and without loss of generality, we assume that node indices are sorted by classes, i.e nodes 1 to n_1 constitute

[1] Approximate optimization of the modularity metric leads to a spectral clustering using the modularity matrix.

$\mathcal{C}_1$, nodes n_1+1 to n_1+n_2 form $\mathcal{C}_2$, and so on.

As motivated in Section 4.1, we proceed to the study of the matrix

$$\mathbf{L}_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} \mathbf{D}^{-\alpha} \left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{2m} \right] \mathbf{D}^{-\alpha} \tag{4.1}$$

where $\mathbf{d} = \mathbf{A}\mathbf{1}_n$, $\mathbf{D} = \mathrm{diag}(\mathbf{d})$ and $m = \frac{1}{2}\mathbf{d}^\mathsf{T}\mathbf{1}_n$.

We are mainly interested in a *dense* graph regime where clustering is not asymptotically trivial so that methods based on different matrices $\mathbf{L}_\alpha$ do not all lead to perfect classification. This regime is ensured by the following growth rate conditions.

Assumption 1. *As $n \to \infty$, k remains fixed and, for all $i, j \in \{1, \ldots, n\}$*

1. *$C_{g_i g_j} = 1 + \frac{M_{g_i g_j}}{\sqrt{n}}$, where $M_{g_i g_j} = \Omega(1)$; we shall denote $\mathbf{M} = \{M_{ab}\}_{a,b=1}^k$.*
2. *q_i are i.i.d. random variables with measure μ having compact support in $(0,1)$.*
3. *$\frac{n_i}{n} \to c_i > 0$ and we will denote $\mathbf{c} = \{c_a\}_{a=1}^k$.*

The goal being to understand the different mechanisms into play when using spectral methods based on $\mathbf{L}_\alpha$, it is essential to study its eigenstructure. As can be observed, $\mathbf{L}_\alpha$ has non independent entries as $\mathbf{D}$ (and $\mathbf{d}$) depend on $\mathbf{A}$, and it thus does not follow a standard random matrix model. We thus proceed in approximating $\mathbf{L}_\alpha$ by a more tractable random matrix $\tilde{\mathbf{L}}_\alpha$ which asymptotically preserves the eigenvalue distribution and isolated eigenvectors of $\mathbf{L}_\alpha$. We obtain the following approximation of $\mathbf{L}_\alpha$.

Theorem 15. *Let Assumption 1 holds and let $\mathbf{L}_\alpha$ be given by (4.1). Then, for $\mathbf{D}_q \triangleq \mathcal{D}(\mathbf{q})$, as $n \to \infty$, $\|\mathbf{L}_\alpha - \tilde{\mathbf{L}}_\alpha\| \to 0$ in operator norm, almost surely, where*

$$\begin{aligned}
\tilde{\mathbf{L}}_\alpha &= \frac{1}{\sqrt{n}} \mathbf{D}_q^{-\alpha} \mathbf{X} \mathbf{D}_q^{-\alpha} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}, \\
\mathbf{V} &= \begin{bmatrix} \frac{\mathbf{D}_q^{1-\alpha}\mathbf{J}}{\sqrt{n}} & \frac{\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{1}_n}{\mathbf{q}^\mathsf{T}\mathbf{1}_n} \end{bmatrix}, \\
\mathbf{\Omega} &= \begin{bmatrix} \left(\mathbf{I}_k - \mathbf{1}_k \mathbf{c}^T\right) \mathbf{M} \left(\mathbf{I}_k - \mathbf{c}\mathbf{1}_k^T\right) & -\mathbf{1}_k \\ -\mathbf{1}_k^T & 0 \end{bmatrix},
\end{aligned}$$

with $\mathbf{X} = \{X_{ij}\}_{i,j=1}^n$ symmetric with independent entries (up to the symmetry), X_{ij} having zero-mean and variance $q_i q_j (1 - q_i q_j)$, and $\mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_k] \in \{0,1\}^{n \times k}$ with $(\mathbf{j}_a)_i = \delta_{\{g_i = a\}}$.

Sketch of Proof. *The proof relies on the fact that we may write $A_{ij} = q_i q_j + q_i q_j \frac{M_{g_i g_j}}{\sqrt{n}} + X_{ij}$ where X_{ij} is a zero-mean random variable with variance $q_i q_j (1 - q_i q_j) + \Theta(n^{-\frac{1}{2}})$, since A_{ij} is a Bernoulli random variable with parameter $q_i q_j (1 + \frac{M_{g_i g_j}}{\sqrt{n}})$. From there, the terms: $\mathbf{d} = \mathbf{A}\mathbf{1}_n$, $\mathbf{d}^\mathsf{T}\mathbf{1}_n$, $\mathbf{d}\mathbf{d}^\mathsf{T}$ and $\mathbf{D} = \mathcal{D}(\mathbf{d})$ composing $\mathbf{L}_\alpha$ can be evaluated. Notably, $\mathbf{D}$ and*

$\mathbf{d}^\mathsf{T}\mathbf{1}_n$ can be decomposed as the sum of dominant terms (with higher spectral norms with respect to n) and trailing terms (vanishing spectral norms with respect to n), so that we can write a Taylor expansion of $\mathbf{D}^{-\alpha}$ and $(\mathbf{d}^\mathsf{T}\mathbf{1}_n)^\alpha$ for $\alpha \in \mathbb{R}$. By computing $\mathbf{L}_\alpha$ using the asymptotic approximations of $\mathbf{D}^{-\alpha}$, $(\mathbf{d}^\mathsf{T}\mathbf{1}_n)^\alpha$, $\mathbf{A}$, $\mathbf{d}\mathbf{d}^\mathsf{T}$, we obtain $\tilde{\mathbf{L}}_\alpha$. The complete proof is provided in Appendix A.1.

This result immediately implies the following Corollary.

Corollary 16. *Under Assumption 1, let $\lambda_i(\mathbf{L}_\alpha)$ (resp., $\lambda_i(\tilde{\mathbf{L}}_\alpha)$) be the eigenvalues of $\mathbf{L}_\alpha$ (resp., $\tilde{\mathbf{L}}_\alpha$) with associated eigenvectors $\mathbf{u}_i(\mathbf{L}_\alpha)$ (resp., $\mathbf{u}_i(\tilde{\mathbf{L}}_\alpha)$). We have*

$$\max_{1\leq i\leq n} \left|\lambda_i(\mathbf{L}_\alpha) - \lambda_i(\tilde{\mathbf{L}}_\alpha)\right| \xrightarrow{\text{a.s.}} 0$$

and, if $\liminf_n \min_{j\neq i} |\lambda_i(\mathbf{L}_\alpha) - \lambda_j(\tilde{\mathbf{L}}_\alpha)| > 0$,

$$\left\|\mathbf{u}_i(\mathbf{L}_\alpha) - \mathbf{u}_i(\tilde{\mathbf{L}}_\alpha)\right\| \xrightarrow{\text{a.s.}} 0.$$

The eigenstructure (eigenvalues and dominant eigenvectors) analysis of $\mathbf{L}_\alpha$ can thus be performed through that of $\tilde{\mathbf{L}}_\alpha$ for large enough n. The matrix $\tilde{\mathbf{L}}_\alpha$ is essentially a classical random matrix model and the study of its eigenvalues and dominant eigenvectors can be performed using standard random matrix theory (RMT) approaches [Benaych-Georges and Nadakuditi, 2012, Hachem et al., 2013].

4.3 Main results

4.3.1 Spiked model and dominant eigenvector regularization

Condition on the knowledge of the intrinsic weights q_i's, the matrix $\tilde{\mathbf{L}}_\alpha$ is equivalent to an additive spiked random matrix ([Baik et al., 2005] or Section 3.2.4) as it is the sum of the standard full rank symmetric random matrix $n^{-\frac{1}{2}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha}$ having independent zero-mean entries and a low-rank matrix $\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$. The difference between $\tilde{\mathbf{L}}_\alpha$ and classical additive spike random matrices lies in the fact that the low-rank term $\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$ is not totally independent of the noise random matrix, and the matrix $\mathbf{V}$ is not composed of orthonormal columns. However, asymptotically, those two differences turn out not having an impact on the spike analysis (the details are provided in Appendix A.3). As shown in Figure 4.1 and explained in Section 3.2.4, the spectrum (eigenvalue distribution) of spiked random matrices is generally composed of (one or several) bulks of concentrated eigenvalues and, when a phase transition is met, of additional eigenvalues which isolate from the aforementioned bulks. The eigenvectors corresponding to the isolated eigenvalues of the spiked random matrix become more correlated to the eigenvectors of the low-rank matrix when the corresponding eigenvalues are far away from the rest of the eigenvalues.

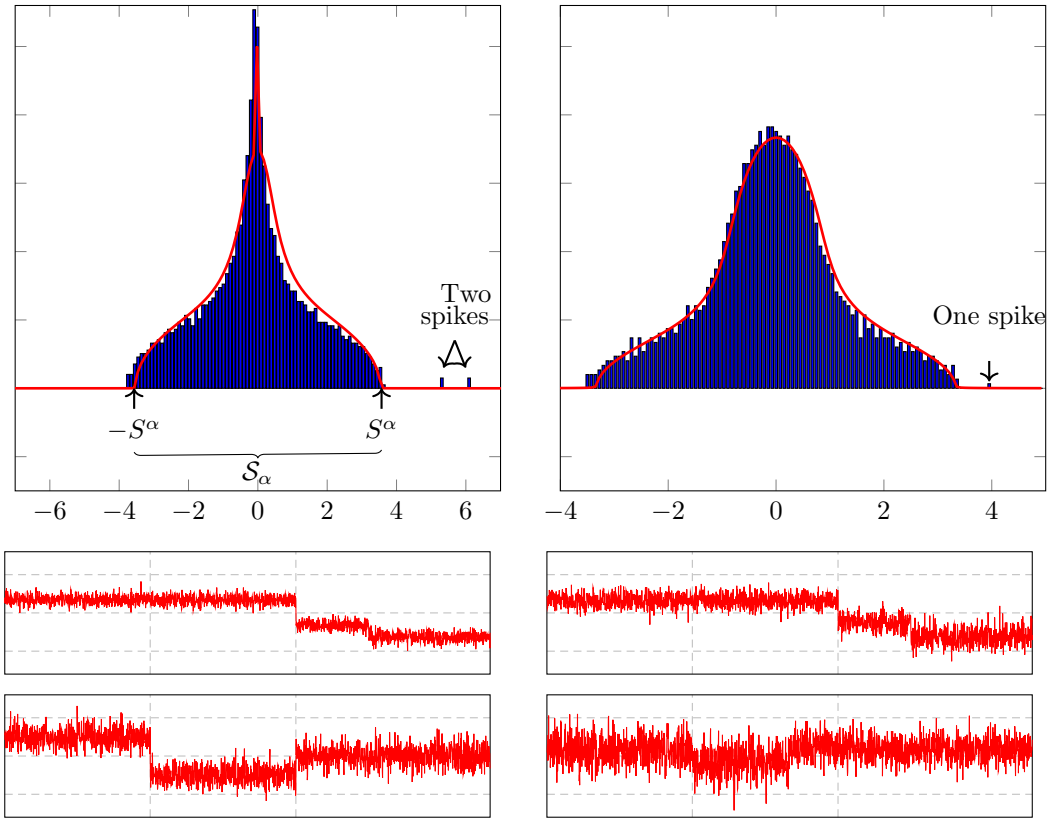


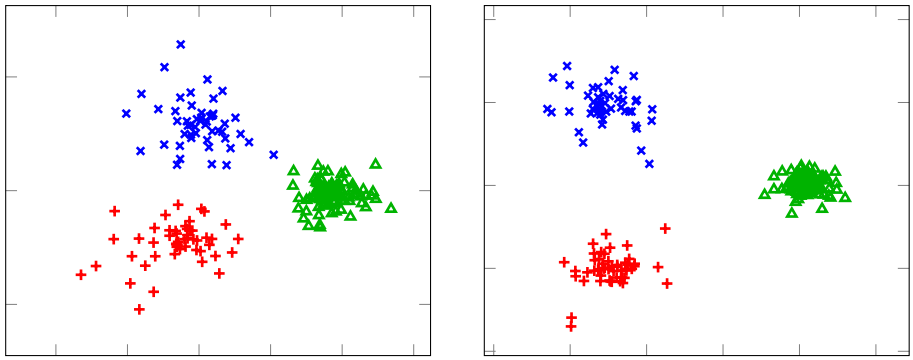
Figure 4.1: Two graphs generated upon the DCSBM with $k = 3$, $n = 2000$, $c_1 = 0.3, c_2 = 0.3, c_3 = 0.4$, $\mu = \frac{1}{2}\delta_{q_{(1)}} + \frac{1}{2}\delta_{q_{(2)}}$, $q_{(1)} = 0.4$, $q_{(2)} = 0.9$ and two different affinity matrices $\mathbf{M}$. **(Left)** $M_{ii} = 12$, $M_{ij} = -4, i \neq j$, **(Right)**: $M_{ii} = -3$, $M_{ij} = -10, i \neq j$, **(Top)**: Eigenvalue distribution of $\mathbf{L}_\alpha$, $\alpha = 0$. **(Bottom)**: First and second leading eigenvectors of $\mathbf{L}_\alpha$, $\alpha = 0$.

The low-rank matrix $\mathbf{V\Omega V}^\mathsf{T}$ in Theorem 15, contains the matrix $\mathbf{D}_q^{1-\alpha}\mathbf{J}$; so, when the phase transition is met, the eigenvectors of $\tilde{\mathbf{L}}_\alpha$ will be correlated to some extent to $\mathbf{D}_q^{1-\alpha}\mathbf{J}$ as long as the corresponding informative eigenvalues are isolated from the bulk of eigenvalues. This is well illustrated in Figure 4.1 where the eigenvectors associated with non-isolated eigenvalues are noisy, i.e., classes can be barely distinguished from those eigenvectors. On the other hand, the eigenvectors associated with isolated eigenvalues consist of noisy plateaus characterizing the classes and thus a consistent classification can be expected using those eigenvectors. However, for a better clustering, one expects instead the vectors used for classification to be correlated to the canonical vectors $\mathbf{j}_a$, $1 \leq a \leq k$, instead of $\mathbf{D}_q^{1-\alpha}\mathbf{j}_a$, the latter being the class step vector entries weighted by the intrinsic probabilities q_i's thus creating some biases when the q_i's are not uniform.

As a consequence, we claim that, letting $\mathbf{u}_1, \ldots, \mathbf{u}_l$ *be the eigenvectors associated with the* ℓ *isolated eigenvalues of* $\mathbf{L}_\alpha$*, the vectors* $\mathbf{n}_i = \mathbf{D}^{\alpha-1}\mathbf{u}_i$ *for* $1 \leq i \leq \ell$ *should be the ones*

used for the classification instead of the $\mathbf{u}_i$*'s.*[2]

This important observation helps correcting the biases (creation of artificial classes) introduced by the degree heterogeneity observed earlier in Figure 1.1. As shown in Figure 4.2, which assumes the same setting as Figure 1.1, when the aforementioned eigenvector regularization is performed prior to EM or k-means classification, the genuine communities are correctly recovered.



$(\alpha = 1)$ $(\alpha = \hat{\alpha}_{\text{opt}})$

Figure 4.2: Two dominant eigenvectors of $\mathbf{L}_\alpha$ pre-multiplied by $\mathbf{D}^{\alpha-1}$(x-y axes) for $n = 2000$, $k = 3$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$, $q_{(1)} = 0.1$, $q_{(2)} = 0.5$, $c_1 = c_2 = \frac{1}{4}$, $c_3 = \frac{1}{2}$, $\mathbf{M} = 100\mathbf{I}_3$ with $\hat{\alpha}_{\text{opt}}$ defined in Section 4.3.4. Same setting as Figure 1.1.

As mentionned earlier, the eigenvectors corresponding to eigenvalues in the bulk have vanishing correlation with the low-rank informative matrix and are thus of no use for clustering asymptotically. It is thus important to characterize the phase transition point beyond which eigenvalues isolate from the bulk and determine which α best ensures this transition. To this end, we will first determine the support $\mathcal{S}^\alpha$ of the limiting spectral distribution (l.s.d) of $\mathbf{L}_\alpha$. Then, following popular spiked model tools, we will find conditions for the existence of isolated eigenvalues. This is the objective of the next sections.

4.3.2 Limiting support

In this section, we characterize the l.s.d. of $\mathbf{L}_\alpha$ where most eigenvalues concentrate. This in turn shall allow to determine the transition point beyond which informative eigenvalues isolate from the main bulk of eigenvalues and consistent clustering can thus be achieved by using the corresponding eigenvectors associated with those eigenvalues. The limiting eigenvalue distribution of $\mathbf{L}_\alpha$ is given in the following result.

[2]As far as the eigenvectors are concerned, we may freely replace $\mathbf{D}_q$ (unknown in practice) by $\mathbf{D}$ (which can be computed from the observed graph) since, from Lemma 24 in the subsequent section 4.3.4, the vector of degrees $\mathbf{d}$ is, up to a scale factor β, a consistent estimator of the vector of intrinsic weights $\mathbf{q}$ and thus $\|\beta\mathbf{D} - \mathbf{D}_q\| \to 0$ almost surely.

Theorem 17 (Limiting spectrum)**.** *Let $\pi_n^\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(\mathbf{L}_\alpha)}$ be the empirical spectral distribution (e.s.d.) of $\mathbf{L}_\alpha$. Then, as $n \to \infty$, $\pi_n^\alpha \to \bar{\pi}^\alpha$ almost surely where $\bar{\pi}^\alpha$ is a probability measure with compact symmetric support $\mathcal{S}^\alpha = [-S_\alpha, S_\alpha]$ defined, for $z \in \mathbb{C}^+ \setminus \mathcal{S}^\alpha$, by its Stieltjes transform*

$$m^\alpha(z) \equiv \int (t-z)^{-1}\, \mathrm{d}\bar{\pi}^\alpha(t) = \int \frac{1}{-z - f^\alpha(z) q^{1-2\alpha} + g^\alpha(z) q^{2-2\alpha}} \mu(dq)$$

where $(f^\alpha(z), g^\alpha(z)) \in (\mathbb{C}^+)^2$ (resp., $(\mathbb{R}^-)^2$) is the unique solution for $z \in \mathbb{C}^+$ (resp., $\mathbb{R}^+$), of

$$\begin{aligned} f^\alpha(z) &= \int \frac{q^{1-2\alpha}\mu(dq)}{-z - f^\alpha(z) q^{1-2\alpha} + g^\alpha(z) q^{2-2\alpha}} \\ g^\alpha(z) &= \int \frac{q^{2-2\alpha}\mu(dq)}{-z - f^\alpha(z) q^{1-2\alpha} + g^\alpha(z) q^{2-2\alpha}}. \end{aligned} \tag{4.2}$$

Theorem 17 gives the l.s.d $\bar{\pi}^\alpha$ of $\mathbf{L}_\alpha$ through its stieltjes transform $m^\alpha(z)$ which is only function of the law μ of the intrinsic weights q_i's. As explained in Section 3.2.2, the stieltjes transform approach is used to overcome the difficulty of working directly with the e.s.d. Note here that $m^\alpha(z)$ does not have a closed form expression and is instead defined through a fixed point equation which can be solved numerically in a few iterations; the l.s.d. $\bar{\pi}^\alpha$ is then found by using the inverse transform (see Theorem 6). As we will see below, in the particular case of homogeneous q_i's, $m(z)$ has an explicit expression which corresponds to the stieltjes transform of the popular semi-circle law (see Theorem 2). We give below a proof's sketch of Theorem 17, the detailed proof being provided in Appendix A.2.

Sketch of Proof. *Since $\tilde{\mathbf{L}}_\alpha = \frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$ is a spiked random matrix, the e.s.d. π_n^α of $\mathbf{L}_\alpha$ converges weakly to the the e.s.d. $\tilde{\pi}_n^\alpha$ of $\frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha}$ (by Weyl interlacing lemma) since $\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$ is a low-rank matrix. We thus find an asymptotic limit $\bar{\pi}^\alpha$ for $\tilde{\pi}_n^\alpha$ so that $\pi_n^\alpha \to \bar{\pi}^\alpha$ almost surely. To do so, we show that the Stieltjes transform of $\tilde{\pi}_n^\alpha$ converges to $m^\alpha(z)$ for $z \in \mathbb{C}^+$, which is the Stieltjes transform of the probability measure $\bar{\pi}^\alpha$ so that the convergence also holds for the probability measures (the e.s.d.). The Stieltjes transform of the e.s.d. $\tilde{\pi}_n^\alpha$ is $n^{-1}\operatorname{tr}(\frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha} - z\mathbf{I}_n)^{-1}$ (where $(\frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha} - z\mathbf{I}_n)^{-1}$ is the so-called resolvent of the random matrix $\frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha}$), the deterministic limit of which gives $m^\alpha(z)$, computed using classical random matrix theory (RMT) tools [Pastur and Shcherbina, 2011]. The calculus details are provided in Appendix A.2.*

Remark 18 (Stochastic Block Model)**.** *Particularizing Theorem 17 to the Stochastic Block Model (SBM) (where $q_i = q_0$ for all i), the limiting probability measure $\bar{\pi}^\alpha$ is the popular semi-circle distribution (Theorem 2) with density $\bar{\pi}^\alpha(dt) = \frac{2}{\pi\sigma^2}\sqrt{(\sigma^2 - t^2)^+}dt$ with $\sigma^2 = q_0^{1-2\alpha}\sqrt{1-q_0^2}$. The associated Stieljes transform $m^\alpha(z)$ is explicit with in particular*

$$q_0^{1-2\alpha} m^\alpha(z q_0^{1-2\alpha}) = q_0^{\frac{1}{2}-\alpha} f^\alpha(z q_0^{\frac{1}{2}-\alpha}) = q_0^{-1} g^\alpha(z q_0^{1-2\alpha}) = -\frac{z}{2(1-q_0^2)} - \sqrt{\left(\frac{z}{2(1-q_0^2)}\right)^2 - \frac{1}{1-q_0^2}}.$$

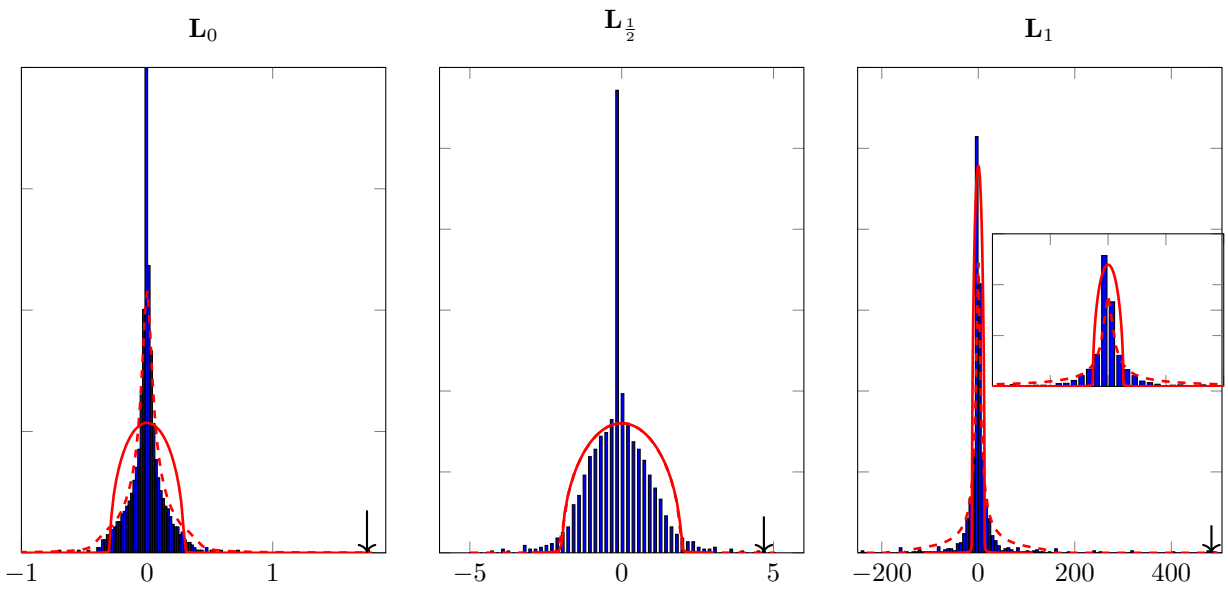


Figure 4.3: Political blogs [Adamic and Glance, 2005] network. Empirical versus Theoretical law of the eigenvalues of $\mathbf{L}_{\hat{\alpha}_{\text{opt}}}$ when fitting this network with the DCSBM (dashed) and the SBM (solid). Here $\hat{\alpha}_{\text{opt}} = 0$. The arrow shows the position of the largest eigenvalue.

The top of Figure 4.1, already discussed above, shows the density of the limiting $\bar{\pi}^\alpha$, for $\alpha = 0$, superimposed over the histogram of π_n^α. Figure 4.3 similarly displays the histogram π^α of the empirical eigenvalues of $\mathbf{L}_\alpha$ corresponding to the real network of Political blogs [Adamic and Glance, 2005] versus the theoretical limiting distribution $\bar{\pi}^\alpha$ obtained by fitting the network to the DCSBM (from Theorem 17, with μ the actual degree distribution of the graph) and the theoretical limiting distribution obtained by fitting the network to the SBM instead (in solid lines).[3] We note importantly that the DCSBM is a good fit for the political blogs network except possibly for $L_{\frac{1}{2}}$ while the SBM does not fit the network in any case. This suggests that the DCSBM is a more appropriate model when studying real-world networks.

4.3.3 Phase transition

We also observe in Figure 4.3 (and more obviously in the synthetic case of Figure 4.1) that different choices of α lead to different behaviors in the position of the dominant eigenvalues. We shall determine here when separation of one or several eigenvalues from the bulk occurs. To this end, we follow popular spiked model techniques [Benaych-Georges and Nadakuditi, 2012, Hachem et al., 2013] for phase transition characterization. This entails the following result.

Theorem 19 (Phase transition)**.** *Let Assumption 1 hold and let* $\lambda(\bar{\mathbf{M}})$ *be a non zero*

[3]The SBM assumes here $q_i = q_0$ for all i.

eigenvalue with multiplicity η of $\bar{\mathbf{M}} \equiv \left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}$. Then, for $\alpha \in \mathbb{R}$, there exists corresponding isolated eigenvalues $\lambda_i(\mathbf{L}_\alpha), \ldots, \lambda_{i+\eta-1}(\mathbf{L}_\alpha) \in \mathbb{R}\backslash\mathcal{S}^\alpha$ of $\mathbf{L}_\alpha$ all converging to $\rho \in \mathbb{R} \setminus \mathcal{S}^\alpha$, as $n \to \infty$, almost surely, if and only if [4]

$$\left|\lambda(\bar{\mathbf{M}})\right| > \tau^\alpha \triangleq -\lim_{x \downarrow S^\alpha} \frac{1}{g^\alpha(x)},$$

with $g^\alpha(x)$ defined in Theorem 17. In this case, ρ is defined by

$$\rho = (g^\alpha)^{-1}\left(-\frac{1}{\lambda(\bar{\mathbf{M}})}\right).$$

Note that Theorem 19 is similar to Theorem 13 on the eigenvalues characterization of additive spike models in that an isolated eigenvalue appears with a limit given by $\rho = (g^\alpha)^{-1}\left(-\frac{1}{\lambda(\bar{\mathbf{M}})}\right)$, when the phase transition is met (here $\left|\lambda(\bar{\mathbf{M}})\right| > \tau^\alpha \triangleq -\lim_{x\downarrow S^\alpha} \frac{1}{g^\alpha(x)}$). In practice, the level of separation between the communities is related to the strength of the eigenvalues of the connectivity matrix $\mathbf{M}$. The farther the communities are, the larger the dominant eigenvalues of $\mathbf{M}$ and thus the phase transition is most likely to occur.

Sketch of Proof. *From Theorem 17, the e.s.d. of $\mathbf{L}_\alpha$ converges weakly to the e.s.d. of $\frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha}$ with support $\mathcal{S}^\alpha$ (defined in Theorem 17) but since $\frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha}$ and $\mathbf{L}_\alpha$ only differ by a finite rank matrix $\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$, some eigenvalues of $\mathbf{L}_\alpha$ may isolate from the support $\mathcal{S}^\alpha$. To find those isolated eigenvalues, we solve for $\rho \notin \mathcal{S}^\alpha$, $\det(\mathbf{L}_\alpha - \rho\mathbf{I}_n) = 0$. This leads to find the ρ's for which $0 = \det(\mathbf{I}_{k+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V}\mathbf{\Omega})$ where $\mathbf{Q}_\rho^\alpha = (\frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha} - \rho\mathbf{I}_n)^{-1}$ is the resolvent of $\frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha}$. By using standard RMT calculus [Benaych-Georges and Nadakuditi, 2012], we obtain a deterministic approximation of $\mathbf{I}_{k+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}$ which leads to the phase transition condition in Theorem 19.*

Remark 20 (τ^α in SBM setting)**.** *From Remark 18, in the SBM setting, τ^α no longer depends on α and is given by $\tau^\alpha = \frac{\sqrt{1-q_0^2}}{q_0}$.*

Remark 21 (Number of isolated eigenvalues)**.** *From Theorem 19, there is a one-to-one mapping between the limiting isolated eigenvalues ρ of $\mathbf{L}_\alpha$ and non zero eigenvalues of $\bar{\mathbf{M}} = \left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}$. As $\mathbf{1}_k^\mathsf{T}\bar{\mathbf{M}} = 0$, $\bar{\mathbf{M}}$ has a maximum of $k-1$ non zero eigenvalues which means that at most $k-1$ eigenvalues of $\mathbf{L}_\alpha$ can be found at macroscopic distance from $\mathcal{S}^\alpha$. Thus, at most $k-1$ eigenvectors of $\mathbf{L}_\alpha$ can be used in the first step of the spectral algorithm described in the introduction.*

Remark 22 (The complete spectrum of $\mathbf{L}_\alpha$)**.** *Strictly speaking, the aforementioned statements are somewhat inaccurate. An exhaustive analysis of $\mathbf{L}_\alpha$ indeed reveals that, under some conditions on μ, and irrespective of the clustering matrix $\mathbf{M}$, extra isolated eigenvalues can be found in the spectrum of $\mathbf{L}_\alpha$, the eigenvectors of which do not contain any*

[4] The limit $\lim_{x\downarrow S^\alpha} g^\alpha(x)$ is well defined in $(-\infty, 0]$ as $x \mapsto g^\alpha(x)$ can be shown to be a continuous growing negative function on the right side of $\mathcal{S}^\alpha$.

structural information about the classes. This rather unfamiliar scenario has also been evidenced in the context of spectral kernel clustering in [Couillet and Benaych-Georges, 2016]. Since this hypothetical eigenvalue and eigenvector pair is of no value for the interest of clustering, it shall no longer be discussed in the following. Besides, most settings of practical interest do not present this singular behavior. A thorough discussion of this peculiarity is provided in Appendix A.5.

The value τ^α defined in Theorem 19 is a community detectability threshold which in the dense regime for the SBM case was shown to split the community detectability into two regions: a region where no algorithm can succed better than a random guess in classifying the nodes and a region where a non trivial detection is possible [Decelle et al., 2011b, Nadakuditi and Newman, 2012]. When the separability condition of Theorem 19 is ensured, the alignment between the properly normalized eigenvectors of $\mathbf{L}_\alpha$ and linear combinations of the class vectors $\mathbf{j}_a$'s (defined in Theorem 15) is away from zero, thus ensuring a non trivial classification performance. The larger $\lambda(\bar{\mathbf{M}})$, the closer are the vectors used for classification to the class vectors $\mathbf{j}_a$'s.

Theorem 23. *Under Assumption 1, let $\lambda(\bar{\mathbf{M}})$ and $\lambda(\mathbf{L}_\alpha)$ be an eigenvalue pair as defined in Theorem 19. We further assume $\lambda(\bar{\mathbf{M}})$ of unit multiplicity and denote $\mathbf{u}$ the eigenvector associated with the eigenvalue $\lambda(\mathbf{L}_\alpha)$. Then, letting $\bar{\mathbf{n}} = \frac{\mathbf{D}^{\alpha-1}\mathbf{u}}{\|\mathbf{D}^{\alpha-1}\mathbf{u}\|}$ and $\mathbf{\Pi} = \sum_{a=1}^k \frac{\mathbf{j}_a\mathbf{j}_a^\mathsf{T}}{n_a}$, for all $\epsilon > 0$, there exists $\gamma_-, \gamma_+ > 0$ such that, for all n large, almost surely,*

$$0 < \left|\lambda(\bar{\mathbf{M}})\right| - \tau^\alpha < \gamma_- \Rightarrow \bar{\mathbf{n}}^\mathsf{T}\mathbf{\Pi}\bar{\mathbf{n}} < \epsilon$$
$$\left|\lambda(\bar{\mathbf{M}})\right| - \tau^\alpha > \gamma_+ \Rightarrow \bar{\mathbf{n}}^\mathsf{T}\mathbf{\Pi}\bar{\mathbf{n}} > 1 - \epsilon.$$

This result, which is an extension of Theorem 14 to our model, is a direct corrolary of Theorem 28 in Appendix 4.3.5.

Figure 4.4 illustrates Theorem 23, which confirms that, below the phase transition threshold τ^α, there is asymptotically no correlation between the vectors $\bar{\mathbf{n}}$ and the class canonical vectors $\mathbf{j}_a$'s and thus no consistent clustering can be achieved in this regime. The theoretical curve is obtained by using the deterministic asymptotic approximation of $\bar{\mathbf{n}}^\mathsf{T}\mathbf{\Pi}\bar{\mathbf{n}}$ which is explicitly given in Appendix A.4.

4.3.4 Optimal α

In this section, we determine the values of α for which the community detectability threshold is maximally achieved. This, in turn, is expected to allow for the optimal extraction of information about the classes from the extreme eigenvectors although this is not easily proved.

From Theorem 19, since $\bar{\mathbf{M}}$ does not depend on α, the smaller τ^α the more likely the detectability condition $\left|\lambda(\bar{\mathbf{M}})\right| > \tau^\alpha$ is met. We then seek α for which τ^α is minimal. For any compact set $\mathcal{A} \subset \mathbb{R}$, we may thus define

$$\alpha_{\text{opt}} \triangleq \text{argmin}_{\alpha\in\mathcal{A}} \{\tau^\alpha\}$$

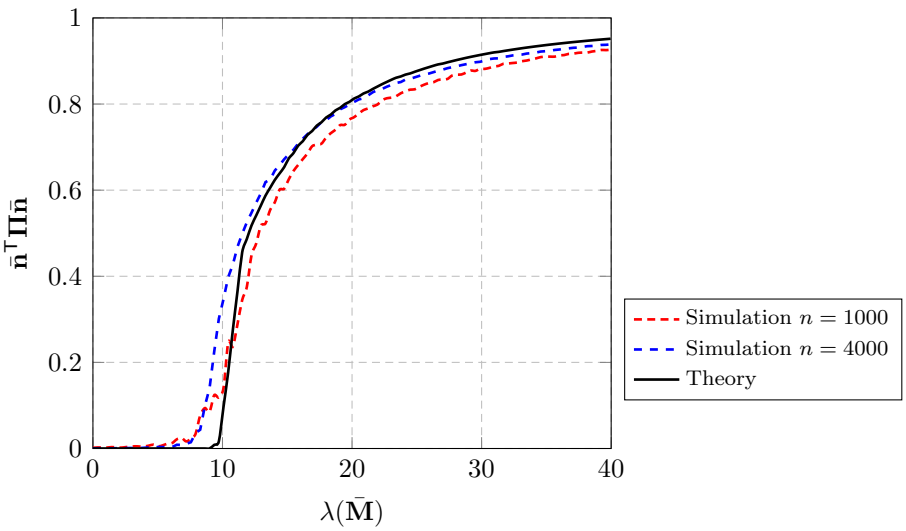


Figure 4.4: Simulated versus empirical $\bar{\mathbf{n}}^\mathsf{T}\mathbf{\Pi}\bar{\mathbf{n}}$ for $k = 3$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$, $q_{(1)} = 0.1$, $q_{(2)} = 0.2$, $c_1 = c_2 = \frac{1}{4}$, $c_3 = \frac{1}{2}$, $\mathbf{M} = \Delta\mathbf{I}_3$ with Δ ranging from 0 to 100.

which we shall assume is unique (if $q_i = q_0$ is constant, τ^α is constant across α; this case is thus excluded). The estimation of α_{opt} however requires the knowledge of $g^\alpha(x)$ for each $\alpha \in \mathcal{A}$. The estimation of $g^\alpha(x)$ can be done numerically by solving the fixed point equation defined in Theorem 17 provided μ is known. As a direct consequence of Assumption 1-(1), μ can in fact be estimated from the empirical graph degrees irrespective of the class matrix $\mathbf{C}$, according to the following result.

Lemma 24. *Let* $\hat{q}_i = \frac{d_i}{\sqrt{\mathbf{d}^\mathsf{T}\mathbf{1}_n}}$. *Then, under Assumption 1,*

$$\max_{1\leq i\leq n} |q_i - \hat{q}_i| \to 0 \tag{4.3}$$

almost surely.

We thus have all the ingredients to estimate α_{opt}.[5]

Proposition 25. *Define* $\hat{\mu} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{\hat{q}_i}$ *with* $\hat{q}_i = \frac{d_i}{\sqrt{\mathbf{d}^\mathsf{T}\mathbf{1}_n}}$ *and* $\hat{\mathcal{S}}^\alpha$, $\hat{f}^\alpha(z)$, $\hat{g}^\alpha(z)$, *as in Theorem 17 but for* μ *replaced by* $\hat{\mu}$. *Then, as* $n \to \infty$,

$$\hat{\alpha}_{\text{opt}} \to \alpha_{\text{opt}}$$

almost surely, where $\hat{\alpha}_{\text{opt}} \triangleq \text{argmin}_{\alpha\in\mathcal{A}}\{\hat{\tau}^\alpha\}$ *with*

$$\hat{\tau}_\alpha \equiv -\frac{1}{\lim_{x\downarrow\hat{S}^\alpha} \hat{g}^\alpha(x)}.$$

[5]Note here that imposing $\mathcal{A}$ to be a compact set ensures the uniform validity of Theorem 17.

Remark 26 (Numerical evaluation of S^α)**.** *Estimating $\hat{\tau}_\alpha$ requires to determine $\hat{\mathcal{S}}^\alpha$. To this end, we use the fact that $\hat{g}^\alpha(x)$ is only defined for $x \notin \hat{\mathcal{S}}^\alpha$. We thus evaluate $\hat{\mathcal{S}}^\alpha$ by an iterative dichotomic search in intervals of the type $[l, r]$ for which $\hat{g}^\alpha(l)$ is undefined (and thus the algorithm in Equation 4.2 does not converge) and $\hat{g}^\alpha(r)$ is defined (the algorithm converges), starting from e.g., $l = 0$ and r quite large.*

Remark 27 (Relevance of the choice of α)**.** *Following Remarks 18 and 20, note that the choice of α is only relevant to heterogeneous graphs, as in the SBM case, the phase transition threshold τ^α is constant irrespective of α. This suggests that the more heterogeneous the graph the more important an appropriate setting of α.*

The aforementionned importance of choosing $\alpha = \hat{\alpha}_{\text{opt}}$ along with the need to pre-multiply the dominant eigenvectors of $\mathbf{L}_\alpha$ by $\mathbf{D}^{\alpha-1}$ before classification, as discussed after exposing Theorem 15, naturally bring us to an improved version of Algorithm 2 provided below. The performances of Algorithm 3 mainly depend on the content of the eigenvectors

Algorithm 3: Improved spectral algorithm

1: Evaluate $\alpha = \hat{\alpha}_{\text{opt}} = \text{argmin}_{\alpha \in \mathcal{A}} \lim_{x \downarrow \hat{S}^\alpha} \hat{g}^\alpha(x)$ as per Proposition 25.
2: Retrieve the ℓ eigenvectors corresponding to the ℓ largest eigenvalues of $\mathbf{L}_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} \mathbf{D}^{-\alpha} \left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\intercal}{2m} \right] \mathbf{D}^{-\alpha}$. Denote $\mathbf{u}_1^\alpha, \ldots, \mathbf{u}_\ell^\alpha$ those eigenvectors.
3: Letting $\mathbf{n}_i^\alpha = \mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha$ and $\bar{\mathbf{n}}_i^\alpha = \frac{\mathbf{n}_i^\alpha}{\|\mathbf{n}_i^\alpha\|}$, stack the vectors $\bar{\mathbf{n}}_i^\alpha$'s columnwise in a matrix $\mathbf{N} = [\bar{\mathbf{n}}_1^\alpha, \ldots, \bar{\mathbf{n}}_\ell^\alpha] \in \mathbb{R}^{n \times \ell}$.
4: Let $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathbb{R}^\ell$ be the rows of $\mathbf{N}$. Cluster $\mathbf{r}_i \in \mathbb{R}^\ell$, $1 \leq i \leq n$ in one of the k groups using any low-dimensional classification algorithm (e.g., k-means or EM). The label assigned to $\mathbf{r}_i$ then corresponds to the label of node i.

$\bar{\mathbf{n}}_i^\alpha$'s. These regularized eigenvectors happen to be shapped like noisy "plateaus" (step functions), each plateau characterizing a class. The objective of the next section is to provide deterministic limits of the parameters of those noisy plateaus from which the asymptotic performances of Algorithm 3 unfold.

4.3.5 Eigenvectors and improvement of Expectation Maximization (EM) algorithm

In this section, we provide a precise characterization of the asymptotic class means and class covariances of the dominant eigenvectors entries (used for clustering) which in turn allows improving the classical EM algorithm used in the last step of spectral clustering procedures. The eigenvectors of $\mathbf{L}_\alpha$ have the property of remaining "stable" in the large dimensional limit, thereby allowing for a precise characterization of their content. This behavior (classical in the spike model analysis of random matrices) however only holds for eigenvectors associated with strictly isolated eigenvalues (in the sense that the latter

remain at a macroscopic distance of all other eigenvalues). In the remainder, we thus assume that the normalized eigenvector $\bar{\mathbf{n}}_i^\alpha$ under study is associated with such a strictly isolated eigenvalue.

As one can see in Figure 4.2, the different clusters of points (rows of $\mathbf{N}$ in Algorithm 3) have different dispersions (variances) in the DCSBM model under consideration. The most appropriate algorithm to use in step 4 of Algorithm 3 is the expectation maximization (EM) method. EM considers each point $\mathbf{r}_i \in \mathbb{R}^\ell$ arising from $[\bar{\mathbf{n}}_1^\alpha, \dots, \bar{\mathbf{n}}_\ell^\alpha]$ as a mixture of k Gaussian random vectors with means $\boldsymbol{\nu}_{EM}^a$ and covariances $\boldsymbol{\Sigma}_{EM}^a \in \mathbb{R}^{\ell\times\ell}$, $a \in \{1, \dots, k\}$. Starting from initial means and covariances, they are sequentially updated until convergence. To identify $\boldsymbol{\nu}_{EM}^a$, $\boldsymbol{\Sigma}_{EM}^a$ and thus understand the performance of Algorithm 3, we may write $\bar{\mathbf{n}}_i^\alpha$ [6] as the "noisy plateaus" vector

$$\bar{\mathbf{n}}_i^\alpha = \sum_{a=1}^{k} \nu_i^a \frac{\mathbf{j}_a}{\sqrt{n_a}} + \sqrt{\sigma_{ii}^a}\mathbf{w}_i^a \tag{4.4}$$

where $\mathbf{w}_i^a \in \mathbb{R}^n$ is a random vector orthogonal to $\mathbf{j}_a$, of norm $\sqrt{n_a}$ and supported on the indices of $\mathcal{C}_a$ and

$$\nu_i^a = \frac{1}{\sqrt{n_a}} (\bar{\mathbf{n}}_i^\alpha)^\mathsf{T} \mathbf{j}_a = \frac{1}{\sqrt{n_a}} \frac{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{j}_a}{\sqrt{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha}} \tag{4.5}$$

$$\sigma_{ij}^a = \frac{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathcal{D}_a\mathbf{D}^{\alpha-1}\mathbf{u}_j^\alpha}{\sqrt{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i}\sqrt{(\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_j^\alpha}} - \nu_i^a\nu_j^a \tag{4.6}$$

with $\mathcal{D}_a = \mathcal{D}(\mathbf{j}_a)$. The vector $\boldsymbol{\nu}^a = (\nu_i^a)_{i=1}^\ell \in \mathbb{R}^\ell$ and the matrix $\boldsymbol{\Sigma}^a = (\sigma_{ij}^a)_{i,j=1}^\ell \in \mathbb{R}^{\ell\times\ell}$ represent respectively the empirical means and empirical covariances of the points $\mathbf{r}_i$ (defined in Algorithm 3) belonging to class $\mathcal{C}_a$. Thus, provided that EM converges to the correct solution, $(\boldsymbol{\nu}_{EM}^a)_i$ and $(\boldsymbol{\Sigma}_{EM}^a)_{ij}$ shall converge asymptotically to the limiting values of $\nu_i^a \in \mathbb{R}$ and σ_{ij}^a respectively. Clearly, for small values of $\boldsymbol{\Sigma}^a$ compared to $\boldsymbol{\nu}^a$, clustering the vectors $\bar{\mathbf{n}}_i^\alpha$ shall lead to good performances.

We find the asymptotic limits of the class means ν_i^a and the class covariances σ_{ij}^a. The explicit expressions of those limits are provided in the proof section (Theorems 42 and 43) for readability reasons.

Theorem 28. *For ν_i^a, σ_{ij}^a defined in* (A.20), (A.21) *respectively, there exist deterministic limits $\nu_i^{a,\infty}$ and $\sigma_{ij}^{a,\infty}$ (explicitely defined in Theorems 42 and 43 in Appendix A.4) such that, as $n \to \infty$, almost surely*

$$\left|(\nu_i^a)^2 - (\nu_i^{a,\infty})^2\right| \to 0$$
$$\left|\sigma_{ij}^a - \sigma_{ij}^{a,\infty}\right| \to 0.$$

Sketch of Proof. *Technically, the standard tools used in spiked random matrix analysis do not allow for an immediate assessment of the quantities ν_i^a and σ_{ij}^a. As a workaround,*

[6] Recall that the graph nodes were assumed labeled by class, and thus the entries of $\bar{\mathbf{n}}_i^\alpha$ are similarly sorted by class.

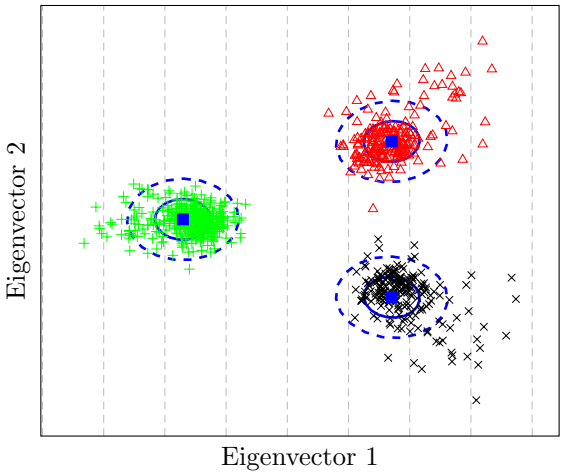


Figure 4.5: $n = 800$, $k = 3$ classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ of sizes $|\mathcal{C}_1| = |\mathcal{C}_2| = \frac{n}{4}$, $|\mathcal{C}_3| = \frac{n}{2}$, $\frac{3}{4}$ of the nodes having $q_i = 0.3$ and the others having $q_i = 0.8$, matrix of weights $\mathbf{C} = \mathbf{1}_3\mathbf{1}_3^\mathsf{T} + \frac{30}{\sqrt{n}}\mathbf{I}_3$. Two dimensional representation of the dominant eigenvectors 1 and 2 of $\mathbf{L}_\alpha$. In blue, theoretical means and one- and two- standard deviations.

we follow the approach used in [Couillet and Benaych-Georges, 2016] which relies on the possibility to estimate bilinear forms of the type $\mathbf{a}^\mathsf{T}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{b}$ *for given vectors* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ *and unit multiplicity eigenvectors* $\mathbf{u}_i^\alpha$ *of* $\mathbf{L}_\alpha$ *since we have from Cauchy formula, as* $n \to \infty$ *almost surely, (since* $\lambda_i(\mathbf{L}_\alpha) \to \rho$*)*

$$\mathbf{a}^\mathsf{T}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{b} = -\frac{1}{2\pi i}\oint_{\Gamma_\rho} \mathbf{a}^\mathsf{T}\left(\mathbf{L}_\alpha - z\mathbf{I}_n\right)^{-1}\mathbf{b}\mathrm{d}z$$

and for a given matrix $\mathbf{D}$

$$(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}\mathbf{u}_i^\alpha = \operatorname{tr}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D} = -\frac{1}{2\pi i}\oint_{\Gamma_\rho} \operatorname{tr}\left(\mathbf{L}_\alpha - z\mathbf{I}_n\right)^{-1}\mathbf{D}\mathrm{d}z$$

where Γ_ρ *is a positively oriented contour circling around the limiting eigenvalue* ρ *of* $\lambda_i(\mathbf{L}_\alpha)$ *associated with the eigenvector* $\mathbf{u}_i^\alpha$ *of* $\mathbf{L}_\alpha$*. The calculus details are provided in Appendix A.4.*

Using the asymptotic results in Theorem 28, we display in Figures 4.5 and 4.6 the theoretical means and standard deviations versus ground truths for each class-wise block of the eigenvectors entries. The good fit between the ground truths and the theoretical findings of the class means and class covariances calls for the improvement of the random initialization of the EM procedure in the last step of spectral clustering.

The performances of EM highly depend on the chosen starting parameters; a first natural choice is to set them randomly, which as we shall see leads to poor performances especially in cases where the clusters are not easily separable. Since the theoretical limiting means $\boldsymbol{\nu}^{a,\infty}$ and covariances $\boldsymbol{\Sigma}^{a,\infty}$ are respectively the limiting values of $\boldsymbol{\nu}_{EM}^a$ and

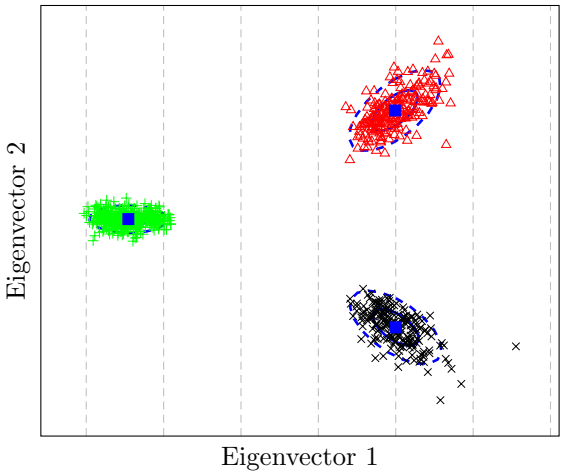


Figure 4.6: $n = 800$, $k = 3$ classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ of sizes $|\mathcal{C}_1| = |\mathcal{C}_2| = \frac{n}{4}$, $|\mathcal{C}_3| = \frac{n}{2}$, q_i's uniformly distributed over $[0.1, 0.9]$, matrix of weights $\mathbf{C} = \mathbf{1}_3\mathbf{1}_3^\mathsf{T} + \frac{100}{\sqrt{n}}\mathbf{I}_3$. Two dimensional representation of the dominant eigenvectors 1 and 2 of $\mathbf{L}_\alpha$. In blue, theoretical means and one- and two- standard deviations

covariances $\mathbf{\Sigma}_{EM}^a$ provided EM converges to the correct solution, we may set as initial parameters of EM our findings $\boldsymbol{\nu}^{a,\infty}$ (Theorem 42) and $\mathbf{\Sigma}^{a,\infty}$ (Theorem 43) for $a \in \{1, \ldots, k\}$ provided those can be estimated. In most scenarios, the many unknowns prevent such an estimation. Nonetheless, from Corollary 45 (Appendix A.4), provided the class proportions (or the sizes of each class) are (more or less) known, we can consistently estimate $\boldsymbol{\nu}^\infty$ and $\mathbf{\Sigma}^\infty$ in a 2-class scenario. As we shall see, this new setting of initial parameters is much better than other initializations approaches.

To show the effect of our setting of initial parameters of EM based on the findings $\boldsymbol{\nu}^\infty$ and $\mathbf{\Sigma}^\infty$, Figure 4.7 compares the empirical performances of our new spectral algorithm based on the regularized eigenvector of $\mathbf{L}_{0.5}$ for different initial settings of the EM parameters *i)* random setting (Random EM) *ii)* our theoretical setting (by assuming that the class proportions are known) and *iii)* the ground truth setting (oracle EM where we set the initial points to the empirically evaluated means and covariances of each cluster based on ground truth). Below the transition point, no consistent clustering can be achieved for large n using the eigenvectors associated with the highest eigenvalues since the classes are not separable and our theoretical limiting means and covariances are not defined since there are no isolated eigenvalues in that case. We have thus initialized EM at random in this non-interesting regime (as for Random EM). The EM algorithm may in that regime set all the nodes to the same class, which will then result in a classification rate close to the proportion of the nodes in the cluster of the largest size. In the interesting regime (after the transition point), we see that the performances (in terms of correct classification rate) of the algorithm using our theoretical setting of EM closely match the performances of an ideal setting with ground truth (oracle EM). The performances of the algorithm using a random initialization (Random EM) are completely degraded especially around critical cases (small values of Δ). Random EM becomes reliable only for very large values

of Δ where clustering is somewhat trivial.

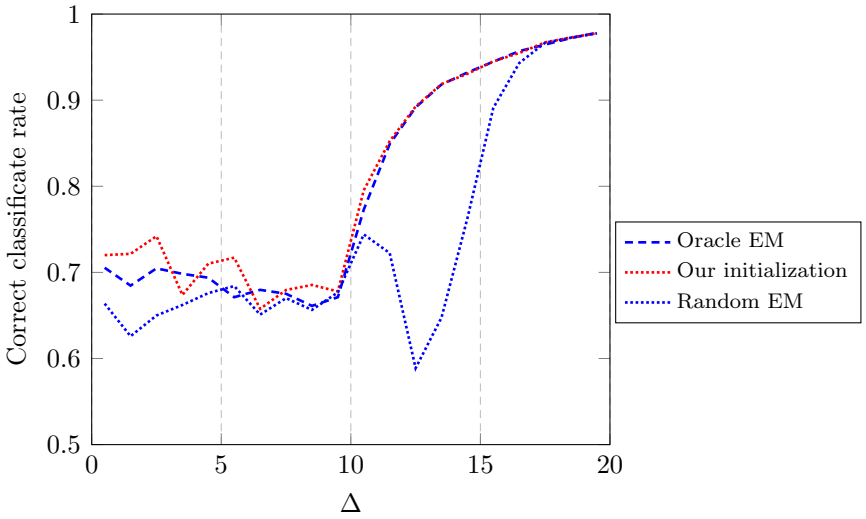


Figure 4.7: Probability of correct recovery for $\alpha = 0.5$, $n = 4000$, $k = 2$, $c_1 = 0.8$, $c_2 = 0.2$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.2$ and $q_{(2)} = 0.8$, $\mathbf{M} = \Delta\mathbf{I}_2$, for $\Delta \in [0, 20]$.

4.4 Numerical simulations

We restrict ourselves to $\alpha \in \mathcal{A} = [0, 1]$ for the numerical simulations. To illustrate the importance of the choice of α_{opt}, Figure 4.8 presents the theoretical (asymptotic) ratio between the limiting largest eigenvalue ρ of $\mathbf{L}_\alpha$ and the right edge S^α of the limiting support $\mathcal{S}^\alpha$ with respect to the amplitude of the eigenvalues of $\bar{\mathbf{M}}$. Although α_{opt} only ensures in theory to have the best isolation of the eigenvalues only in "worst cases scenarios"(i.e., when $\lambda(\bar{\mathbf{M}})$ is slighty larger than $\tau^{\alpha_{\text{opt}}}$), Figure 4.8 shows that taking $\alpha = \alpha_{\text{opt}}$ provides the largest gap $\frac{\rho}{S^\alpha}$ for all values of $\lambda(\bar{\mathbf{M}})$. This suggests (again, without any theoretical support) best performances with $\alpha = \alpha_{\text{opt}}$ in all cases (for any value of $\mathbf{M}$).

In the sequel, to compare the different algorithms, we will use the performance evaluation measure known as *overlap with ground truth communities*, defined in [Krzakala et al., 2013] as

$$\text{Overlap} \equiv \frac{\frac{1}{n}\sum_{i=1}^{n} \delta_{g_i \hat{g}_i} - \frac{1}{K}}{1 - \frac{1}{K}},$$

where g_i and $\hat{g}_i$ are the true and estimated labels of node i, respectively. Figure 4.9 subsequently shows the overlap performance under the setting of Figure 4.8 for a simulated graph of $n = 3000$ nodes. Note that the empirically observed phase transitions closely match the theoretical ones (drawn in circles and the same as in Figure 4.8). We then consider in Figure 4.10 a DCSBM graph where $\mathbf{M}$ is fixed and three quarters of the nodes connect with a fixed intrinsic low weight $q_{(1)} = 0.1$ and we vary the intrinsic weights $q_{(2)}$ of the remaining quarter of the nodes from low to high weights. We observe a sudden

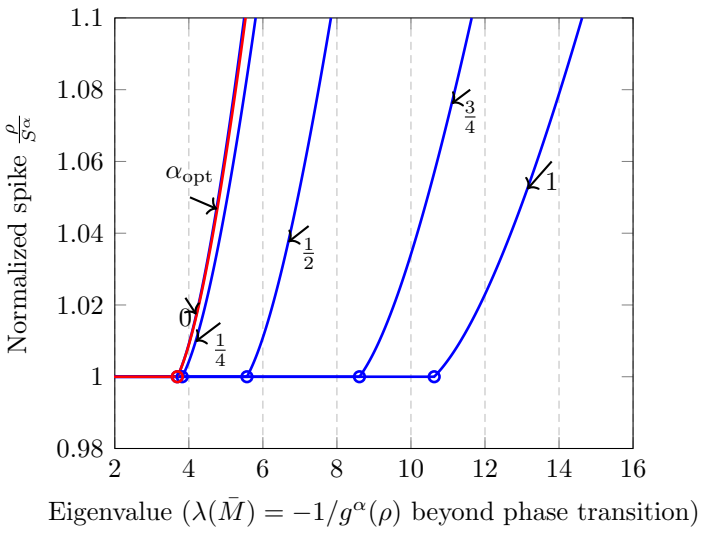


Figure 4.8: Ratio between the limiting largest eigenvalue ρ of $\mathbf{L}_\alpha$ and the right edge of the support $\mathcal{S}^\alpha$, as a function of the largest eigenvalue $\lambda(\bar{\mathbf{M}})$ of $\bar{\mathbf{M}}$, $\mathbf{M} = \Delta\mathbf{I}_3$, $c_i = \frac{1}{3}$, for $\Delta \in [10, 150]$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, for $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\text{opt}}\}$ (indicated on the curves of the graph). Here, $\alpha_{\text{opt}} = 0.07$. Circles indicate phase transition.

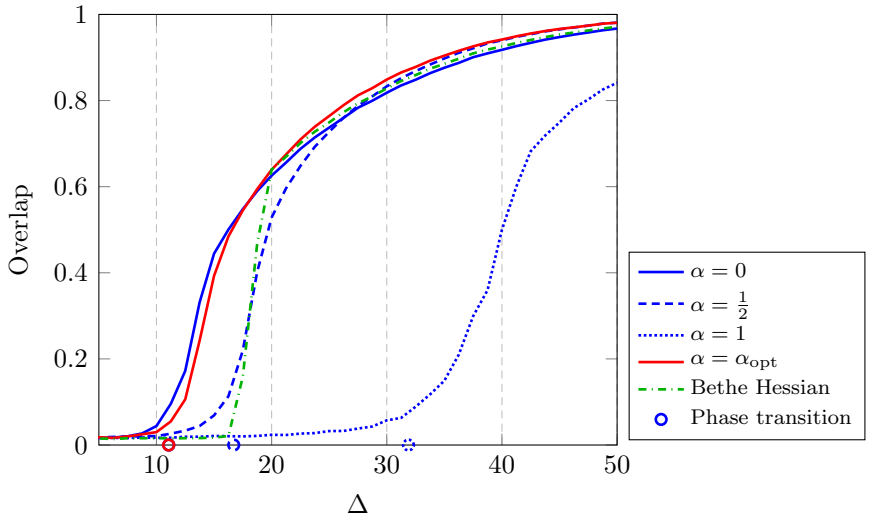


Figure 4.9: Overlap performance for $n = 3000$, $k = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $\mathbf{M} = \Delta\mathbf{I}_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\text{opt}} = 0.07$.

drop of the BH overlap for large $q_{(2)} - q_{(1)}$. This phenomenon is consistent with the fact, observed earlier in Figure 1.1, that BH creates artificial communities out of nodes with the same q_i parameter. This is a practical demonstration of the need for a proper eigenvector normalization to avoid degree biases. This observation has recently led [Newman, 2013] to consider a regularization for the non-backtracking operator on which the BH method is based, which still awaits for proper analysis.

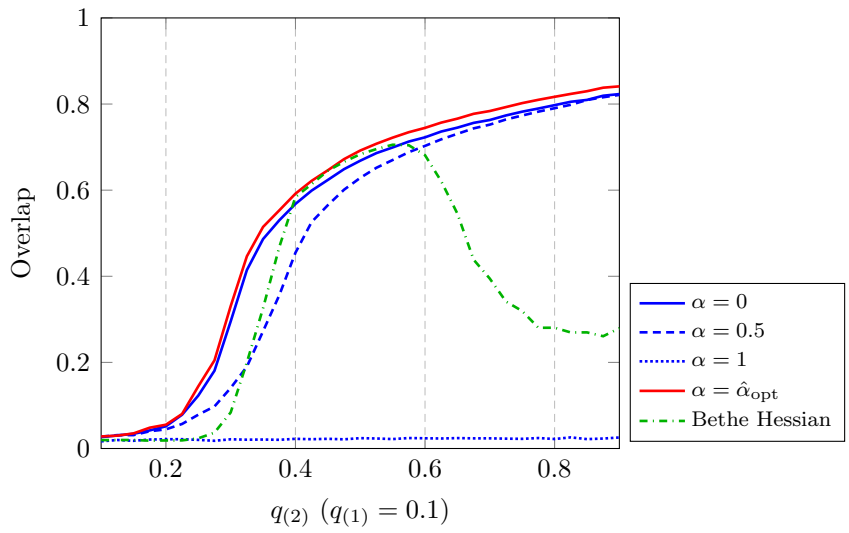


Figure 4.10: Overlap for $n = 3000$, $k = 3$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} \in [0.1, 0.9]$, $\mathbf{M}$ defined by $M_{ii} = 10$, $M_{ij} = -10, i \neq j$, $c_i = \frac{1}{3}$.

In Figure 4.11, we consider a more realistic synthetic graph where the q_i's assume a power law of support $[0.05, 0.3]$ which simulates a sparse graph characteristic of real-world networks. Although this is not the regime we study in this article, our method for $\alpha = \hat{\alpha}_{\text{opt}}$ still competes with the BH method which was developed for sparse homogeneous graphs. However, it is seen that the theoretical phase transitions do not closely match the empirical ones especially for the case $\alpha = 1$. This mismatch is likely due to the fact that our theoretical results in this article require $P_{ij} = \Omega(1)$ which is not always the case in this scenario.

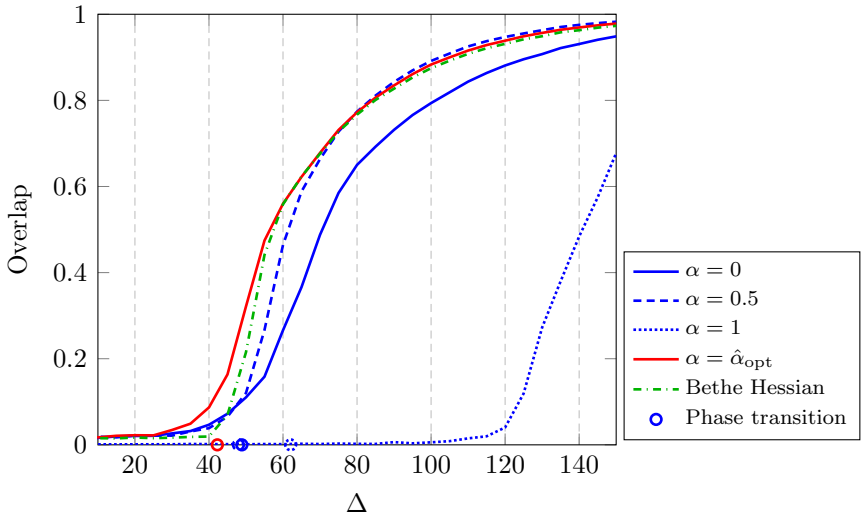


Figure 4.11: Overlap for $n = 3000$, $k = 3$, $c_i = \frac{1}{3}$, μ a power law with exponent 3 and support $[0.05, 0.3]$, $\mathbf{M} = \Delta\mathbf{I}_3$, for $\Delta \in [10, 150]$. Here $\hat{\alpha}_{\text{opt}} = 0.28$.

Algo	Overlap	Modularity
$\hat{\alpha}_{\text{opt}}$ ($\simeq 0$)	**0.897**	**0.4246**
$\alpha = 0.5$	0.035	$\simeq 0$
$\alpha = 1$	0.040	$\simeq 0$
BH	0.304	0.2723

Table 4.1: Overlap performance and Modularity after applying the different spectral algorithms on the Political blogs graph [Adamic and Glance, 2005].

We finally confront the performances (in terms of overlap and modularity [7]) of the different spectral algorithms on the Political blogs graph [Adamic and Glance, 2005] in Table 4.1. We should note that while $\alpha_{\text{opt}} = 0$ in this case, it achieves the best performance both in terms of the overlap to the ground truth and of the modularity. [8] Likely, the reason why $\alpha = 0$ is optimal on the Political blogs data set can be seen in Figure 4.3, where $\mathbf{L}_0$ is the similarity matrix for which the isolated eigenvalue is the farthest from the bulk of the other eigenvalues and thus the associated eigenvector is more aligned to the classes compared to the eigenvectors of $\mathbf{L}_{\frac{1}{2}}$ and $\mathbf{L}_1$.

4.5 Conclusion

In this chapter, we have studied a family of normalized modularity matrices $\mathbf{L}_\alpha \propto \frac{1}{\sqrt{n}}\mathbf{D}^{-\alpha}\left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\intercal}{2m}\right]\mathbf{D}^{-\alpha}$ which generalize the matrices (modularity, Laplacian) used for spectral community detection in dense networks. The main difficulty for the study of those random matrices comes from the dependency between their entries. We tackle this difficulty by establishing the approximation $\|\mathbf{L}_\alpha - \tilde{\mathbf{L}}_\alpha\| \to 0$ using similar techniques as in [Couillet and Benaych-Georges, 2016] where $\tilde{\mathbf{L}}_\alpha$ belongs to the class of so called "spiked" random matrices for which the study of eigenvalues and eigenvectors is classical. The study of eigenvalues and eigenvectors of $\mathbf{L}_\alpha$ used for the classification is thus performed using $\tilde{\mathbf{L}}_\alpha$.

We go further than the observation of [Gulikers et al., 2015] and [Newman, 2013] which states that it is important to use the eigenvectors of $\mathbf{L}_1$ rather than the classically

[7] The modularity Q for a given graph partition with class labels g_i's is defined as : $Q = \frac{1}{2m}\sum_{i,j=1}^{n}\left(A_{ij} - \frac{d_i d_j}{2m}\right)\delta_{g_i=g_j}$ where $\mathbf{d} = \mathbf{A}\mathbf{1}_n$ is the degree vector and $m = \frac{1}{2}\mathbf{d}^\intercal\mathbf{1}_n$ is the total number of edges.

[8] We should note here that the scores for the BH are different from the ones found in the article [Saade et al., 2014] since here we are running k-means algorithm in the last step of the spectral algorithm while the authors of [Saade et al., 2014] have instead used a sign classification of the eigenvector components for networks with two communities.

used $\mathbf{L}_0$ for the classification when the network has heterogeneous degree distributions to avoid some important misclassifications induced by degree biases. We saw in Figure 4.2 for example that the eigenvectors of $\mathbf{L}_1$ correct the degree biases but we show that it is better to use instead the eigenvectors of $\mathbf{L}_0$ premultiplied by $\mathbf{D}^{-1}$ (see Figures 4.8-4.11 for example). Better still, we show that there exists an optimal α called α_{opt} for which taking the eigenvectors of $\mathbf{L}_{\alpha_{\text{opt}}}$ pre-multiplied by $\mathbf{D}^{\alpha_{\text{opt}}-1}$ ensures best performance (or to be more precise best asymptotic cluster detectability).

We generalize the study in [Nadakuditi and Newman, 2012] concerning the evaluation of per-class means of the entries of the unique eigenvector used for the classification, which was limited to the symmetric stochastic block model with two classes of the same average size. Here we consider a more general model (a non necessarily symmetric stochastic block model with heterogeneous degree distribution (DCSBM), arbitrary number of classes, arbitrary class sizes) and we introduce new techniques to evaluate theoretically the limiting per-class means and covariances of the eigenvectors components which are used for the low-dimensional classification. Those aforementioned limiting per-class means and per-class covariances are the limiting values of the corresponding per-class means and per-class covariances that the Expectation Maximization (EM) algorithm shall find empirically (in the last step of the spectral method) provided it converges. One can then initialize the EM parameters with our theoretical findings instead of initializing them at random as classically done. However, those theoretical limiting quantities depend on the model parameters such as the class proportions, the eigenvectors of the affinity matrix $\bar{\mathbf{M}}$ which, except for particular cases ($K = 2$ classes for example), are not directly accessible from real world graphs. We may empirically estimate those parameters by applying Algorithm 3 for a fixed α, computing the empirical class-wise means ν_i^a's/covariances σ_{ij}^a's and using their theoretical formula (Theorems 42 and 43) to deduce the unknown parameters $\mathbf{v}$'s (eigenvectors of $\bar{\mathbf{M}}$) and $\mathbf{c}$ (class proportions vector) associated to the graph.

The results and methods in this article are all based on the strong assumption that the class-wise correction factors $C_{g_ig_j}$ differ by $\mathcal{O}(n^{-\frac{1}{2}})$ e.g., $\forall i, j \in \{1, \dots, n\}$, $C_{g_ig_j} = 1 + \frac{M_{g_ig_j}}{\sqrt{n}}$. Previous works [Lyzinski et al., 2014, Lei et al., 2015, Gulikers et al., 2015] suggest that the present analysis, which only considers "first order spectral statistics", should naturally extend to moderately sparse graphs (of as little as $\mathcal{O}(\log n)$ average degree). Under the sparse DCSBM graph assumption, strikingly different tools are required, opening up a challenging area of improved algorithm research. Similarly, if the $C_{g_ig_j}$'s differ at a rate $n^{-\frac{1}{2}} \ll r_n \ll 1$, mere refinements of our analysis ensure asymptotic weak consistency for all values of α based on the present tools. In passing, this shows that identifiability considerations are equivalent to those delineated for any α, as in [Gulikers et al., 2015] for $\alpha = 1$. Formally, the case where $r_n = \mathcal{O}(1)$ breaks Lemma 24 and therefore the validity of our present analysis but this scenario is also by and far covered by previous works.

Chapter 5

Multi-layer heterogeneous community detection

5.1 Introduction

In the previous chapter, the objects of interest were single-layer graphs for which research on community detection was and continues to be very active. However, the emergence of multiple types of relationships in current real-world networks leading to multilayer graphs, has called for the development of new methods for community detection in multilayer graphs. Current research efforts have been developping aggregation methods which consist in collapsing the different layers altogether to extract a common community structure. It is however more realistic in practice to have a community structure that is common to all the layers while a structure that is distinct between the different layers. In this chapter, we propose a new method that automatically detects shared and unshared communities between the different layers of a *multilayer weighted graph*. We follow a Bayesian generative model approach to generate multilayer weighted stochastic block models for which the labels of a subset of nodes are shared between all the layers while the labels of the complementary set of nodes are independently generated for each layer. Due to the intractable form of the posterior distribution of the nodes' labels given the graphs, under the constraints on the correlations between the nodes' labels, we derive a variational Bayes approximation to that posterior. The parameters of the variational approximation are determined to be the ones for which the variational distribution is the closest to the sought posterior (in terms of Kullback-Leibler divergence). We show through synthetic examples that the proposed method is more accurate than previous approaches to community detection in multilayer graphs in extracting both shared and unshared communities from weighted graphs. We also make use of our method on a multinomic molecular biology dataset where it enables the discovery of heterogeneous communities between *gene-gene functional network* and *gene-gene spatial network*.

5.2 Joint Weighted Stochastic Block Models

To start with, let us recall the definition of a single-layer Stochastic Block Model generated by a weighted distribution $\mathcal{D}$ with sufficient statistic function T and natural parameter function η. Given *latent* community label $g_i \in \{1, \ldots, k\}$ (with k denoting the number of communities) of each vertex i and a community-wise connectivity matrix $\mathbf{C} \in \mathbb{R}^{k \times k}$, an edge is placed between two vertices i and j with an adjacency weight A_{ij} such that

$$\mathbb{P}(A_{ij}|g_i, g_j, C_{g_i,g_j}) \propto \exp\left\{T(A_{ij})\eta(C_{g_i g_j})\right\}.$$

Following a Bayesian approach, prior distributions are attributed to the labels g_i and the community-wise connectivity matrix $\mathbf{C}$.

We denote a multilayer graph, $\mathcal{G}$, defining L as the number of layers and n as the number of vertices. The graph in the l-th layer is an undirected (possibly weighted) graph $\mathcal{G}^{(l)} = (\mathcal{V}, \mathcal{E}^{(l)})$ with $\mathcal{V}$ denoting the set of common vertices and $\mathcal{E}^{(l)}$ denoting the set of edges in graph $\mathcal{G}^{(l)}$. We denote by $\mathbf{A}^{(l)}$ the adjacency matrices containing the edge weights between each pair of vertices in graph $\mathcal{G}^{(l)}$. We propose the following generative heterogeneous community structure of the multilayer graph $\mathcal{G}$.

1. We assume that each layer is subdivided into $k^{(l)}$ non-overlapping communities among which the first $k \leq \min_l k^{(l)}$ are shared between the layers as described below.

2. We first generate the label $g_i^{(1)}$ of each vertex i in the first layer as $g_i^{(1)} \sim$ Multinomial $(\boldsymbol{\mu}_0^{(1)})$ where $\boldsymbol{\mu}_0^{(1)} \in \mathbb{R}^{k^{(1)}}$ contains prior probabilities that the vertices belong to one of the $k^{(1)}$ communities.

3. For each vertex i, if $g_i^{(1)} \in \{1, \ldots, k\}$ then set $g_i^{(l)} = g_i^{(1)}$ for each layer l. Otherwise, generate for each layer l, $g_i^{(l)} \sim$ Multinomial $(\boldsymbol{\mu}_0^{(l)})$.

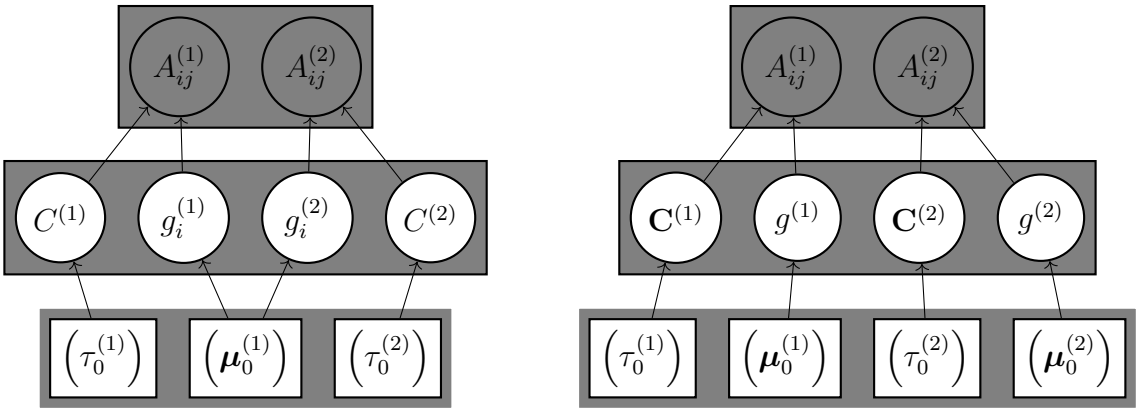
4. Given *latent* community labels $g_i^{(l)}$ (generated in steps 2 and 3) of each vertex i and community-wise connectivity matrices $\mathbf{C}^{(l)} \in \mathbb{R}^{k^{(l)} \times k^{(l)}}$ (the generation of which will be defined later), an edge is placed between two vertices i and j and it is assigned an adjacency weight $A_{ij}^{(l)}$ drawn according to

$$\mathbb{P}(A_{ij}^{(l)}|g_i^{(l)}, g_j^{(l)}, C^{(l)}_{g_i^{(l)} g_j^{(l)}}) \propto \exp\left\{T^{(l)}(A_{ij}^{(l)})\eta^{(l)}(C^{(l)}_{g_i^{(l)} g_j^{(l)}})\right\} \tag{5.1}$$

   where $T^{(l)}$ is the sufficient statistic function and $\eta^{(l)}$ is the natural parameter function of the weights distribution.

5. The community-wise connectivity matrices $\mathbf{C}^{(l)}$ are generated according to conjugate priors associated with the distribution characterized by $(T^{(l)}, \eta^{(l)})$ i.e.,

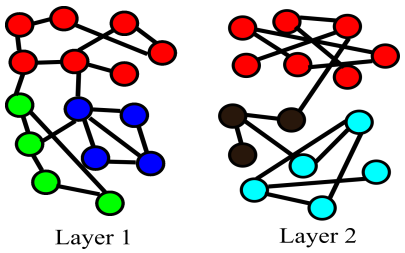
$$p^\star(C_{ab}^{(l)}) = \frac{1}{Z^{(l)}(\tau_0^{(l)})} \exp(\tau_0^{(l)} \eta^{(l)}(C_{ab}^{(l)}))$$



(a) $g_i^{(1)} \in \{1, \ldots, k\}$.

(b) $g_i^{(1)} \notin \{1, \ldots, k\}$.

Figure 5.1: Generative graphical model. Circles and rectangles represent random and deterministic (parameters) variables respectively. Observed variables are shaded.



(a) Heterogeneous multilayer network.

(b) Community affinity probability matrices $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$ decomposed into blocks.

Figure 5.2: **a)** Shared communities (in red) and unshared communities in different colors for each layer. **b)** Red blocks (common to $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$) used to update shared communities while blue blocks used to update private communities of $\mathbf{A}^{(1)}$ and green blocks for the update of $\mathbf{A}^{(2)}$'s private communities.

with $\tau_0^{(l)}$ denoting the associated hyperparameters and $Z^{(l)}(\tau_0^{(l)})$ the normalization constants.

For illustration, we specialize the presentation to two communities, for which each of the matrices $\mathbf{C}^{(l)}$ are decomposed into four blocks corresponding, respectively, to the shared-shared, shared-private, private-shared, private-private interconnections (see Figure 5.2b). As in [Aicher et al., 2014], we consider each sub-matrix $\mathbf{C}_1^{(l)}, \mathbf{C}_2^{(l)}, \mathbf{C}_3^{(l)}, \mathbf{C}_4^{(l)}$ as one-dimensional vectors where the elements are stacked. Let us denote by $r_1^{(l)}, r_2^{(l)}, r_3^{(l)}, r_4^{(l)}$ the indexing variables into each of the obtained vectors i.e., $r_1^{(l)} = 1, \ldots, k^2; r_2^{(l)} = 1, \ldots, k(k^{(l)} - k); r_3^{(l)} = 1, \ldots, k(k^{(l)} - k); r_4^{(l)} = 1, \ldots, (k^{(l)} - k)^2$. The overall prior distribution can thus be written as

$$p^\star(\mathbf{g}^{(l)}, \mathbf{C}^{(l)}, l = 1, \ldots, L) = \prod_{l=1}^{L} \prod_i (\mu_0^{(l)})_{i, g_i^{(l)}} \prod_{r^{(l)}} \frac{1}{Z^{(l)}(\tau_0^{(l)})} \exp(\tau_0^{(l)} \eta^{(l)}(C_{r^{(l)}}^{(l)})) \tag{5.2}$$

with $r^{(l)} \equiv \{r_1^{(l)}, r_2^{(l)}, r_2^{(l)}, r_4^{(l)}\}$.

Given $\mathcal{G}^{(l)}, l = 1, \ldots, L$ (equivalently their adjacency matrices $\mathbf{A}^{(l)}$), the goal is to infer the community labels $g_i^{(l)}$ for each node i in each layer l i.e., to find the most probable clustering $\mathbf{g}^{(l)}$ of the vertices in each layer in the set of all different possible partitioning

$$[(\mathbf{g}^{(1)})^\star, \ldots, (\mathbf{g}^{(l)})^\star] = \text{argmax}_{\mathbf{g}^{(l)}, l=1,\ldots,L} \mathbb{P}(\mathbf{g}^{(l)} | \mathbf{A}^{(l)}, \mathbf{C}^{(l)}, l = 1, \ldots, L) \tag{5.3}$$

with the correlations constraints on $\mathbf{g}^{(l)}$ defined in Point 3 of Section 5.2. The optimization problem (5.3) is N-P hard due to two main difficulties: the maximization is over all possible configurations of $\mathbf{g}^{(l)}$, the calculation of the posterior distribution $\mathbb{P}(\mathbf{g}^{(l)} | \mathbf{A}^{(l)}, \mathbf{C}^{(l)})$, which is intractable due to its high dimensional integral form. Our approach to the optimization (5.3) is the mean field variational Bayes approximation [Jordan et al., 1999, Blei et al., 2006] that uses a factorisable distribution as an approximation to the joint posterior $p(\mathbf{g}^{(l)}, \mathbf{C}^{(l)}) \equiv \mathbb{P}(\mathbf{g}^{(l)}, \mathbf{C}^{(l)} | \mathbf{A}^{(l)})$.

5.3 Variational inference

5.3.1 Mean field variational Bayes inference

Denote by $q(\mathbf{g}^{(l)}, \mathbf{C}^{(\mathbf{l})})$ an approximating (factorisable) distribution that depends on tunable shaping parameters $\boldsymbol{\mu}^{(l)}$ and $\boldsymbol{\tau}^{(l)}$. The variational Bayes algorithm fits the distribution q to the joint distribution by minimizing the KL-divergence, i.e., $q = \text{argmin}_r D_{KL}(r||p)$.

Here the distribution q is taken to have the same parametric form as the prior $p^\star$

$$q(\mathbf{g}^{(l)}, \mathbf{C}^{(l)}, l = 1, \ldots, L) = \prod_{l=1}^{L} \prod_i \mu_{i,g_i^{(l)}}^{(l)} \prod_{r^{(l)}} \frac{1}{Z^{(l)}(\tau_{r^{(l)}}^{(l)})} \exp(\tau_{r^{(l)}}^{(l)} \eta^{(l)}(C_{r^{(l)}}^{(l)})) \tag{5.4}$$

where $\tau_{r^{(l)}}^{(l)}$ and $\boldsymbol{\mu}^{(l)} \in \mathbb{R}^{n \times k^{(l)}}$ are variational parameters corresponding to the random variables $C_{r^{(l)}}^{(l)}$, $\mathbf{g}^{(l)}$ respectively. We can rewrite the original problem (5.3) as follows

$$\begin{aligned}
[(\mathbf{g}^{(\mathbf{1})})^\star, \cdots, (\mathbf{g}^{(\mathbf{L})})^\star] &= \text{argmax}_{\mathbf{g}^{(\mathbf{l})}} \int \mathbb{P}(\mathbf{g}^{(\mathbf{l})}, \mathbf{C}^{(\mathbf{l})} | \mathbf{A}^{(\mathbf{l})}) \mathrm{d}\mathbf{C}^{(\mathbf{l})} \\
&\approx \text{argmax}_{\mathbf{g}^{(\mathbf{l})}} \int q(\mathbf{g}^{(\mathbf{l})}, \mathbf{C}^{(\mathbf{l})}) \mathrm{d}\mathbf{C}^{(\mathbf{l})} \\
&= \text{argmax}_{\mathbf{g}^{(\mathbf{l})}} \int \prod_l \prod_i q(g_i^{(l)}) q(\mathbf{C}^{(\mathbf{l})}) \mathrm{d}\boldsymbol{C}^{(l)} \\
&= \text{argmax}_{\mathbf{g}^{(\mathbf{l})}} \prod_l \prod_i q(g_i^{(l)}).
\end{aligned}$$

Since $q^{(l)}$ is a categorical distribution with parameter $\boldsymbol{\mu}^{(l)}$, the original problem (5.3) is equivalent to

$$(g_i^{(l)})^\star = \text{argmax}_k \, \boldsymbol{\mu}_{ik}^{(l)} \tag{5.5}$$

for each node i and layer l, and thus, the multilayer community detection boils down to a Maximum A Posteriori (MAP) estimator on each individual nodal variational parameter $\boldsymbol{\mu}_i^{(l)}$ for each layer l.

5.3.2 Learning

As per [Aicher et al., 2014], the constant model likelihood can be written as $\log \mathbb{P}(\mathbf{A^{(l)}}) = \mathcal{G}(q) + D_{KL}(q||p)$ with

$$\mathcal{G}(q) = \mathbb{E}_q \log \mathbb{P}(\mathbf{A^{(l)}}|\mathbf{g}^{(l)}, \mathbf{C}^{(l)}, l = 1, 2) + \mathbb{E}_q \frac{p^\star}{q} \tag{5.6}$$

where $p^\star$ is the prior distribution assigned to the parameters $\mathbf{g}^{(l)}, \mathbf{C}^{(l)}$. Since the likelihood is constant, minimizing $D_{KL}(q||p)$ (and thus making the approximation q to be the closest to the sought posterior p) is equivalent to maximizing $\mathcal{G}(q)$ over the variational parameters. In the sequel, we devise a procedure to learn the parameters for which $\mathcal{G}(q)$ is maximized.

We next address how to find the variational parameters $\tau^{(l)}$, $\boldsymbol{\mu}^{(l)}$ for which $\mathcal{G}(q)$ is maximized. To this end, let us first compute $\mathcal{G}(q)$ with the forms of the prior $p^\star$ and the approximation q defined in the previous section. For illustration, we specialize to $L = 2$ but the same principle applies to any number of layers. We have

$$\begin{aligned}\mathcal{G}(q) &= \mathbb{E}_q \log \mathbb{P}(\mathbf{A^{(1)}}, \mathbf{A^{(2)}}|\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \mathbf{C}^{(1)}, \mathbf{C}^{(2)}) + \mathbb{E}_q \frac{p^\star}{q} \\ &= \mathbb{E}_q \log \mathbb{P}(\mathbf{A^{(1)}}|\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \mathbf{C}^{(1)}) + \mathbb{E}_q \log \mathbb{P}(\mathbf{A^{(2)}}|\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \mathbf{C}^{(2)}) + \mathbb{E}_q \frac{p^\star}{q}\end{aligned} \tag{5.7}$$

where in the last line, we use the chain rule along with condition of conditional independence between $\mathbf{A^{(1)}}$ and $\mathbf{A^{(2)}}$ given $\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \mathbf{C}^{(1)}, \mathbf{C}^{(2)}$.The structure of the heterogeneous Joint Stochastic Block Model (Section 5.2) couples the random variables $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ in a simple manner that can be decomposed into the following cases, which we call the *dependency cases*:

- For a vertex pair (i, j) belonging to a block with $\mathbf{C}_1^{(l)}$, $g_i^{(1)} = g_i^{(2)}$ and $g_j^{(1)} = g_j^{(2)}$.
- For a vertex pair (i, j) belonging to a block with $\mathbf{C}_2^{(l)}$, $g_i^{(1)} = g_i^{(2)}$ and $g_j^{(1)} \neq g_j^{(2)}$.
- For a vertex pair (i, j) belonging to a block with $\mathbf{C}_3^{(l)}$, $g_i^{(1)} \neq g_i^{(2)}$ and $g_j^{(1)} = g_j^{(2)}$.
- For a vertex pair (i, j) belonging to a block with $\mathbf{C}_4^{(l)}$, $g_i^{(1)} \neq g_i^{(2)}$ and $g_j^{(1)} \neq g_j^{(2)}$.

Using these *dependency cases* with (5.1), we obtain an expression for $\mathcal{G}(q)$ as per (5.7). After differentiation with respect to the sought variational parameters, we obtain updates for $\tau^{(l)}$, $\boldsymbol{\mu}^{(l)}$, which are stationary points of $\mathcal{G}(q)$ and correspond to local maxima. The precision of the local maxima depends on the initial values for $\boldsymbol{\mu}^{(l)}$. A single run of a

single-layer clustering algorithm shall lead to the optimal solution basin of attraction. Due to the *dependency cases*, the community memberships variational parameters $\mu_{ik}^{(l)}$ depends on $g_i^{(l)}$ either belonging to the set of shared communities $\{1, \ldots, k\}$, or to the set of unshared communities ($\{k+1, \ldots, k^{(l)}\}$). The derivation details are provided in Appendix C.

Algorithm 4 provides the necessary equations for the updates of the variational parameters $\tau^{(l)}$ and $\boldsymbol{\mu}^{(l)}$. Due to Equation (5.5), a max decision rule can then be used on $\boldsymbol{\mu}^{(l)}$ to assign labels to each node, namely $\text{argmax}_k\, \mu_{ik}^{(l)}$ gives the label assigned to node i in graph $\mathcal{G}^{(l)}$. The label of node i is shared between different graphs $\{\mathcal{G}^{(l)}\}$ when $\text{argmax}_k\, \mu_{ik}^{(l)} \in \{1, \ldots, k\}$ and the label is unshared otherwise.

Algorithm 4 is an extension of the variational Bayes algorithm for inferring hidden communities from single-layer graphs [Aicher et al., 2014, Zhang and Zhou, 2017] to the inference of hidden *shared* and *unshared* communities from multilayer graphs. In Algorithm 4, the updates for the parameters $\tau^{(l)}$ are done independently for each graph as in [Aicher et al., 2014]. As for the community membership variational parameters $\boldsymbol{\mu}^{(l)} \in \mathbb{R}^{n \times k^{(l)}}$, the updates of the first k columns of $\boldsymbol{\mu}^{(l)}$ are identical and are computed by adding the contributions of each graph. The last $k^{(l)} - k$ columns of $\boldsymbol{\mu}^{(l)}$ are updated independently using only the information about each graph. This is quite intuitive since the first k columns of $\boldsymbol{\mu}^{(l)}$ correspond to the shared community evidences and thus they should be updated using the contributions of the graphs altogether, while the last columns correspond to unshared communities and thus the updates should be done independently for each graph.

5.4 Experiments

5.4.1 Synthetic graphs

We first consider two Bernoulli SBM graphs $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ with the same intra-community probabilities and different inter-community probabilities in such a way that one graph is noisier than the other. Blindly identifying the community labels from each of the graphs would yield poor performances since we do not know in advance which graph has a clearer community structure than the other. $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are constructed with $n = 500$ vertices, each partitioned into $k^{(l)} = 4$ ($l = 1, 2$) communities respectively among which $k = 2$ are shared between the two graphs. The community labels $g_i^{(l)}$ are assigned uniformly overly the intervals $\{1, \ldots, k^{(l)}\}$ in such a way that $g_i^{(1)} = g_i^{(2)}$ when $g_i^{(1)} \in \{1, \ldots, k\}$. Given the community labels $g_i^{(l)}$, the entries of the adjacency matrices are generated as $A_{ij}^{(l)} \sim \mathit{Bernoulli}(\mathbf{C}^{(l)})$ with $\mathbf{C}^{(1)} = (p-q)\boldsymbol{I}_4 + q\mathbf{1}_4\mathbf{1}_4^\intercal$ and $\mathbf{C}^{(2)} = (p-q')\boldsymbol{I}_4 + q'\mathbf{1}_4\mathbf{1}_4^\intercal$ where we fix $p = 0.6$, $q = 0.2$ and we vary q' between 0.2 and 0.5. The larger q' is, the noisier the graph $\mathcal{G}^{(2)}$ is in comparison with $\mathcal{G}^{(1)}$ and the more difficult community recovery is when applying a community detection algorithm on $\mathcal{G}^{(2)}$ solely. The left figure in Figure 5.3 shows that our Joint mean field algorithm outperforms the competing method

Multilayer Extraction[Wilson et al., 2017] (M-E) in extracting both shared and unshared communities from the two correlated graphs. Both methods are designed to exploit the graph $\mathcal{G}^{(1)}$ to identify the communities of $\mathcal{G}^{(2)}$ (among which some are shared with $\mathcal{G}^{(1)}$). Both joint methods significantly outperform a spectral clustering algorithm and a mean field variational algorithm applied on $\mathcal{G}^{(2)}$ alone.

We next consider the same graph settings as before but with disparate distributions $A_{ij}^{(1)} \sim Bernoulli(\mathbf{C}^{(1)})$ and $A_{ij}^{(2)} \sim Poisson(\mathbf{C}^{(2)})$. Here the M-E algorithm is not exploitable since the latter is designed only for binary graphs (Bernoulli). Our joint variational algorithm is thus compared with single-graph clustering algorithms. The results are reported in the right figure of Figure 5.3 where our joint algorithm outperforms single-layer clustering algorithms (spectral clustering and mean field variational Bayes).

Although our algorithm is designed for detecting shared and unshared communities from multilayer graphs, here we compare our method with the state-of-the-art algorithms designed to find only shared communities between multilayer networks which is a particular case of our model with $k^{(1)} = k^{(2)} = k$. We use the synthetic dataset $\mathbf{mL{-}128}$ designed by [Brodka and Grecki, 2012] which is an extension of the LFR benchmark [Lancichinetti et al., 2008] to multilayer networks. The parameter μ characterizes the variation in the vertex degrees among layers (the higher μ the more variations in the layers vertices' degrees). In Figure 5.4, by varying the number of layers ℓ, we compare in terms of Normalized Mutual Information (NMI) for increasing μ, the output of our joint mean algorithm with some state-of-the-art methods for the identification of shared communities in multilayer networks. Although our method is designed to seek for shared and unshared communities at the same time, it competes well with the PMM [Tang et al., 2009] and the MLMAOP [Pizzuti and Socievole, 2017] methods, only optimized for shared community detection in multi-layer networks. In addition, as shown above, our method is able to also recover unshared communities between different layers of the network.

5.4.2 Real world graphs

In this section, we make use of our novel approach to understand the interplay between genome structure (form) and transcription (function) based on a human fibroblast proliferation dataset [Chen et al., 2015a]. This dataset consists of Hi-C contact maps [Lieberman-Aiden et al., 2009] that capture chromatin architectures and RNA-seq data that provide gene expression levels over 8 time points. We first build a correlation matrix between the RNA-seq values, where thresholding is applied to obtain a binary adjacency matrix $\mathbf{A}^{(1)}$ representing functional correspondence between different genes. The threshold was determined using the asymptotic expression in [Hero and Rajaratnam, 2011] for the minimal RNA-seq correlation necessary to maintain functional interaction between genes for a given number of samples n (here the number of time points), the number of variables p (here the number of genes in one chromosome) and a given level of significance. We then construct an average (over the 8 time points) Hi-C matrix $\mathbf{A}^{(2)}$ and round each entry of the average matrix to the closest integer value. For the application of the variational

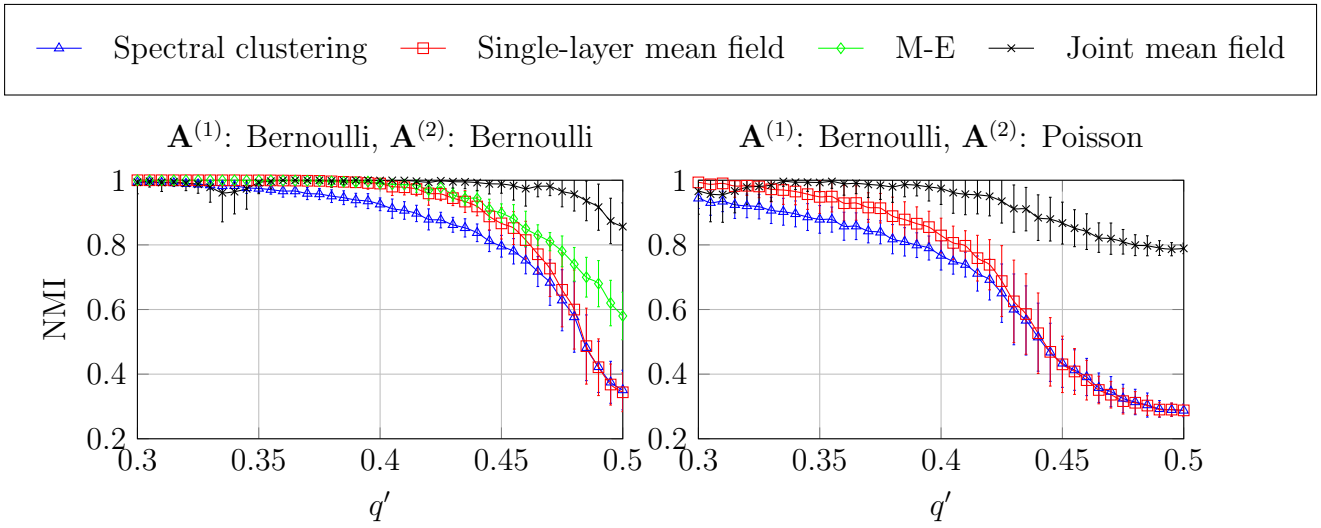


Figure 5.3: Normalized Mutual Information (NMI) between communities (of noisier graph $\mathcal{G}^{(2)}$) identified by different community detection algorithms and ground truths, $n = 500$, $k = 2$ shared communities between the two graphs, $k^{(l)} = 4$. Averages over 100 randomly generated graphs.

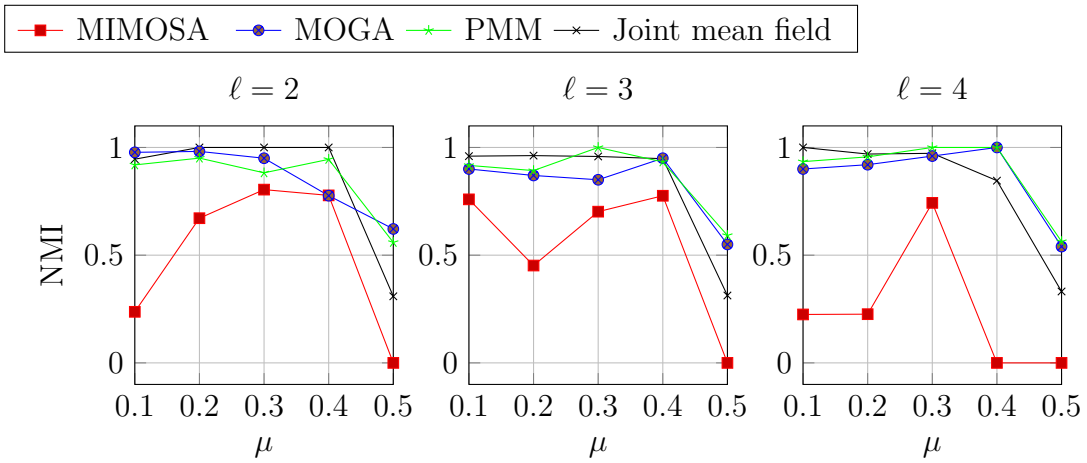
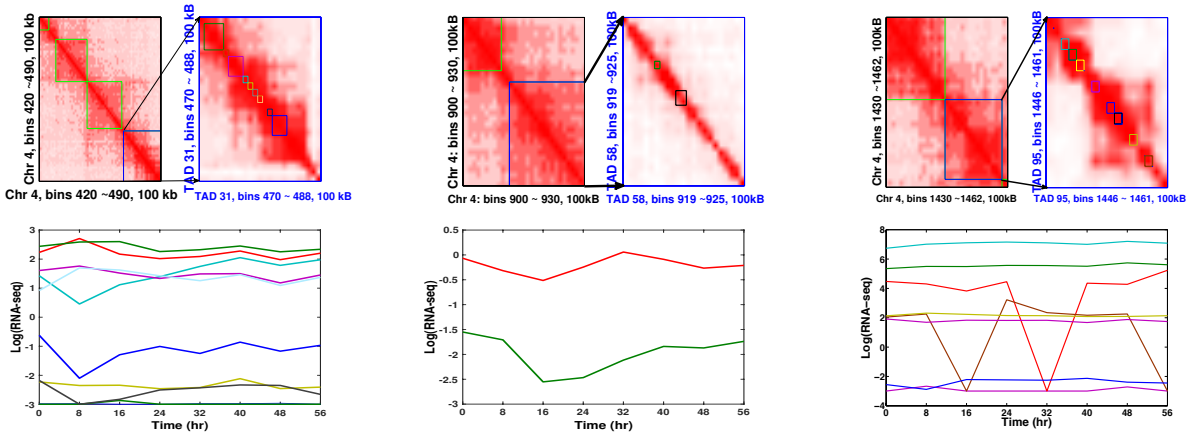


Figure 5.4: mLFR-128 networks with increasing μ and number of layers ℓ. Normalized Mutual Information (NMI) between extracted communities and ground truths.



(a) Shared cluster (b) RNA-seq private cluster (c) Hi-C private cluster

Figure 5.5: **Top**: *Left*) A portion of chromosome 4 with bins coordinates at the unit of 100 kilo-base (Kb) pair [Chen et al., 2015a]. TADs are represented by boxes, where the blue box indicates the TAD of our interest. *Right*) The deployment of genes (identified by our algorithm) in the TAD highlighted at the left plots. **Bottom**: Expression levels of genes that our algorithm identified belonging to the marked TADs in second columns of top sub-figures.

Bayes algorithm, the entries of $\mathbf{A}^{(1)}$ are considered to be Bernoulli distributed while those of $\mathbf{A}^{(2)}$ are considered to be Poisson distributed. More sophisticated models for the sample correlation graph, e.g., Wishart distributions, could also be considered but this is left for future work.

It is shown in [Dixon et al., 2012, Chen et al., 2015a, Chen et al., 2016] that the genome structure per chromosome can be divided into different topologically associating domains (TADs), each of which may contain differently expressed genes. Although [Chen et al., 2015a] found that some genes in a single TAD can maintain similar expression levels, it is unclear how to effectively find such a mapping between TADs and gene expression considering the fact that there are more than 22000 genes in the human genome. In Figure 5.5, we focus on chromosome 4 as an example in order to show how our proposed method provides an elegant way to gain insights on the genomic form-function relationship. Figure 5.5-(a) shows Hi-C contacts and gene expressions corresponding to a subset of genes in one of the shared clusters that we found. We observe from Figures 5.5-(a) that the genes in a shared cluster aggregate in one TAD, which indicates their frequent interactions. In this shared cluster, the same group of genes has very similar expressions. Our results confirm the biological findings in [Chen et al., 2015a] that co-expressed genes exist in a single TAD. Moreover, as shown in Figure 5.5-(b), our analysis establishes the form-function relationship for genes in a RNA-seq private cluster. We observe that, as compared to Figure 5.5-(a), fewer genes belong to the same TAD even though they are more strongly co-expressed. Finally, Figure 5.5-(c) shows that for a Hi-C private cluster, the genes are possibly aggregated in a small region of the chromosome but they have significantly different expression profiles. To sum up, in contrast to a single-layer com-

munity detection algorithm, our method allows to differentiate groups of genes **i)** loosely co-expressed but highly interconnected, **ii)**loosely interconnected but highly co-expressed **iii)** highly co-expressed and highly interconnected.

5.5 Conclusion

Our proposed joint mean field variational algorithm is capable of extracting shared communities across all graph layers as well as identifying communities unique to each layer. The method is applicable to any multilayer network (with or without edge weights) and can provide important insights in the analysis of real-world systems as demonstrated for the human fibroblast dataset. Mean field variational methods are faster to implement but less accurate than Belief propagation (BP) variational approaches. It might be interesting to derive a BP algorithm to solve this multilayer community detection problem and compare the complexity/performance of the two approaches. Algorithm 4 requires as inputs the number of shared clusters and the total number of clusters of each layer. In practice, those are also unknown variables which need to be selected by fine-tuning using some criteria provided by the practitioner. In this work, those numbers of clusters are assumed to be given to the algorithm. Future work might consider optimizing the number of clusters based on reasonable criteria. Finally, the present work assumes that the shared clusters are common to all layers while in practice, only a subset of layers might share communities [Wilson et al., 2017]. An interesting direction of future investigation would be to consider extensions to that more general case.

Algorithm 4: Mean field inference of heterogeneous communities in multilayer graphs.

Inputs: For $l = 1, \dots, L$, layers adjacencies $\mathbf{A}^{(l)}$, layer distributions $\mathcal{D}^{(l)} = [T^{(l)}, \eta^{(l)}, Z^{(l)}]$, number of shared communities k, total number $k^{(l)}$ of communities.
Output: $\boldsymbol{\mu}^{(1)}, \dots, \boldsymbol{\mu}^{(L)}$.
For $l = 1, \dots, L$, initialize $\boldsymbol{\mu}^{(l)}$ and choose hyperparameters $\tau_0^{(l)}$.
repeat
  for $l = 1$ **to** L **do**
    for $r^{(l)} = 1$ **to** $(k^{(l)})^2$ **do**
      $\tau^{(l)}_{r^{(l)}} = \tau_0^{(l)} + \sum_{ij} \sum_{(g_i^{(l)}, g_j^{(l)}) = r^{(l)}} T^{(l)}(A_{ij}^{(l)}) \mu^{(l)}_{i,g_i^{(l)}} \mu^{(l)}_{j,g_j^{(l)}}$.
    end for
  end for
  repeat
    for $i = 1$ **to** n **do**
      for $a = 1$ **to** k **do**

$$\mu_{ik}^{(1)} = \exp\Big\{ \frac{1}{L} \sum_{l=1}^{L} \Big[\sum_{\substack{r_1^{(l)}, j \neq i \\ (a, g_j^{(l)}) = r_1^{(l)}}} T^{(l)}(A_{ij}^{(l)}) \mu^{(l)}_{j,g_j^{(l)}} \bar{\eta}^{(l)}_{r_1^{(l)}} + \sum_{\substack{r_2^{(l)}, j \neq i \\ (a, g_j^{(l)}) = r_2^{(l)}}} T^{(l)}(A_{ij}^{(l)}) \mu^{(l)}_{j,g_j^{(l)}} \bar{\eta}^{(l)}_{r_2^{(l)}} \Big] \Big\}$$

$$\bar{\eta}^{(l)}_{r^{(l)}} = \frac{\partial \log Z^{(l)}}{\partial \tau^{(l)}} \Big|_{\tau^{(l)} = r^{(l)}}$$

      end for
      for $l = 1$ **to** L **do**
        $\mu^{(l)}_{i,1:k} = \mu^{(1)}_{i,1:k}$.
        for $a = k + 1$ **to** $k^{(l)}$ **do**

$$\mu_{ik}^{(l)} = \exp\Big\{ \sum_{\substack{r_3^{(l)}, j \neq i \\ (a, g_j^{(l)}) = r_3^{(l)}}} T^{(l)}(A_{ij}^{(l)}) \mu^{(l)}_{j,g_j^{(l)}} \bar{\eta}^{(l)}_{r_3^{(l)}} + \sum_{\substack{r_4^{(l)}, j \neq i \\ (a, g_j^{(l)}) = r_4^{(l)}}} T^{(l)}(A_{ij}^{(l)}) \mu^{(l)}_{j,g_j^{(l)}} \bar{\eta}^{(l)}_{r_4^{(l)}} \Big\}$$

        end for
        For $a = 1, \dots, k^{(l)}$, $\mu^{(l)}_{i,a} = \mu^{(l)}_{i,a} / \sum_{b=1}^{k^{(l)}} \mu^{(l)}_{i,b}$.
      end for
    end for
  until convergence
until convergence

Chapter 6

Inner product kernel spectral clustering

6.1 Introduction

The objective of this chapter is to conduct a comprehensive study of *kernel spectral clustering* of large and numerous data. As the study of the eigenvalues and eigenvectors of data-driven kernel matrices is not accessible, we shall consider that the data is drawn from a Gaussian mixture model. It is shown in [Couillet and Benaych-Georges, 2016] that this is not an undesirable model since extremely close fit in performances are obtained between real-world datasets (with the MNIST database in particular) and Gaussian mixture data generated with the same empirical means and covariances as the real data. We focus in this thesis on kernel affinities of the type $K_{ij} = f\left(\frac{1}{p}\mathbf{x}_i^\mathsf{T}\mathbf{x}_j\right)$, which are much simpler and more interpretable than the previously studied kernel $K_{ij} = f\left(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ in [Couillet and Benaych-Georges, 2016]. As explained in the introduction, the notion of "closeness" between data points of the same class is not trivial in the large dimensional regime, since distances between pairs of data tend to be concentrated into one value. But this concentration phenomenon allows for a Taylor expansion of the kernel matrix $\mathbf{K}$ leading then to a linear tractable random matrix which takes the form of an additive spiked random matrix. This structure allows for a precise analysis of the eigenvalues and eigenvectors of the kernel matrix $\mathbf{K}$. Due to the aforementioned concentration effect, the performances of high dimensional kernel spectral clustering depend only on a local behavior of the kernel function f and thus proper kernel choices need to be done. This work proposes a new family of kernel functions enabling to discriminate the data based upon the statistical difference between the class means and between class covariances to better *discriminative rates*.

6.2 Large dimensional Gausian Mixture Model

We consider $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$ a set of vectors to be classified into k similarity classes $\mathcal{C}_1, \cdots, \mathcal{C}_k$ such that for $\mathbf{x}_i \in \mathcal{C}_a$, $\mathbf{x}_i = \boldsymbol{\mu}_a + \mathbf{w}_i$, with some vector $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and $\mathbf{w}_i \sim \mathcal{N}(0, \mathbf{C}_a)$ where $\mathbf{C}_a$ satisfy the following growth rate control

Assumption 2 (Growth rates control)**.** *The matrices $\mathbf{C}_a$ are nonnegative definite and invertible and, as $p, n \to \infty$, for $\|\cdot\|$ the operator norm,*

$$\liminf_p \max\{\|C_a\|, \|C_a^{-1}\|\} < \infty.$$

We assume without loss of generality that the vectors are sorted by classes as

$$\mathbf{x}_{n_1+\cdots+n_{a-1}+1}, \cdots, \mathbf{x}_{n_1+\cdots+n_a} \in \mathcal{C}_a, a = 1, \cdots, k$$

where n_a denotes the number of vectors in $\mathcal{C}_a$. We assume the large dimensional regime where both p and n grow large at the same rate. As a basis of comparison, we derive below in an *oracle* setting (a supervised setting where the statistical means and covariances of each vector $\mathbf{x}_i$ are known and the task consists in retrieving the classes) , the minimum distance rates on the class means and class covariances necessary to achieve non trivial classification error (i.e., with probability of error neither 0 nor $1/k$) asymptotically. Through a complete study of the kernel matrix $\mathbf{K}$ (for generic kernel functions f), we will derive the minimum distance rates for which non trivial classification error can be achieved when using spectral clustering with the kernel matrix $\mathbf{K}$. We will then show that there exists a family of kernel functions for which one can achieve non trivial classification performance with distance rates (on the class means and class covariances) *closed* to the aforementioned *optimal oracle distances*.

6.3 Minimal distance rates in oracle (supervised) setting

As a reference, we derive in this section the minimum data rates necessary to achieve non trivial classification error in the best scenario where the class means and covariances are perfectly known. We take the same data setting as in Section 6.2 but in the simpler case $k = 2$ with uniform prior probability to belong to each class i.e., $\frac{1}{2}$. For $\mathbf{x} \in \mathbb{R}^p$, a Neyman-Pearson test (assuming known means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and known covariances $\mathbf{C}_1, \mathbf{C}_2$) for $\mathbf{x}$ belonging to class $\mathcal{C}_1$ consists in the following comparison

$$(\mathbf{x} - \boldsymbol{\mu}_2)^\mathsf{T} \mathbf{C}_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) - (\mathbf{x} - \boldsymbol{\mu}_1)^\mathsf{T} \mathbf{C}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) > \log \frac{|\mathbf{C}_1|}{|\mathbf{C}_2|} \tag{6.1}$$

When $\mathbf{x}$ is a vector genuinely belonging to class $\mathcal{C}_1$ i.e., $\mathbf{x} = \boldsymbol{\mu}_1 + \mathbf{C}_1^{\frac{1}{2}}\mathbf{w}$ with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, the test (6.1) boils down to verifying whether $S(\mathbf{x}) > 0$ where for $\Delta\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

$$S(\mathbf{x}) = \frac{1}{p}\mathbf{w}^\mathsf{T}\left(\mathbf{C}_1^{\frac{1}{2}}\mathbf{C}_2^{-1}\mathbf{C}_1^{\frac{1}{2}} - \mathbf{I}_p\right)\mathbf{w} + \frac{2}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_2^{-1}\mathbf{C}_1^{\frac{1}{2}}\mathbf{w} + \frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_2^{-1}\Delta\boldsymbol{\mu} - \frac{1}{p}\log\frac{|\mathbf{C}_1|}{|\mathbf{C}_2|}. \tag{6.2}$$

The random variable $S(x)$ can be written as the sum of p independent random variables and thus by carefully applying Lyapunov's central limit theorem [Billingsley, 1995] along with Assumption 2, we have as $p \to \infty$, the following central limit result on $S(x)$

$$V_S^{-\frac{1}{2}}\left(S(\mathbf{x}) - E_S\right) \to \mathcal{N}(0,1) \tag{6.3}$$

in distribution where

$$E_S = \frac{1}{p}\operatorname{tr}\mathbf{C}_1\mathbf{C}_2^{-1} - 1 + \frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_2^{-1}\Delta\boldsymbol{\mu} - \frac{1}{p}\log\frac{|\mathbf{C}_1|}{|\mathbf{C}_2|} \tag{6.4}$$

$$V_S = \frac{2}{p^2}\operatorname{tr}\left(\mathbf{C}_1^{\frac{1}{2}}\mathbf{C}_2^{-1}\mathbf{C}_1^{\frac{1}{2}} - \mathbf{I}_p\right)^2 + \frac{4}{p^2}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_2^{-1}\mathbf{C}_1\mathbf{C}_2^{-1}\Delta\boldsymbol{\mu}. \tag{6.5}$$

In order to avoid perfect classification or impossible classification for the vector $\mathbf{x}$ into its genuine class, one should have the mean and standard deviation of $S(x)$ to be of the same order of magnitude (with respect to p). We will consider some particular cases on $\boldsymbol{\mu}_a$ and $\mathbf{C}_a$ to determine minimal requirements on the distance rates between the $\boldsymbol{\mu}_a$'s and between the $\mathbf{C}_a$'s such that the aforementioned non-trivial classification can be achieved.

- We start with the case where class covariances are equal i.e $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$. In that case

$$E_S = \frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{\mu} = \mathcal{O}\left(\frac{\|\boldsymbol{\mu}\|^2}{p}\right) \tag{6.6}$$

$$V_S = \frac{4}{p^2}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{\mu} = \mathcal{O}\left(\frac{\|\boldsymbol{\mu}\|^2}{p^2}\right) \tag{6.7}$$

  which implies that in order to have $E_S \sim_p \sqrt{V_S}$, $\|\Delta\boldsymbol{\mu}\| = \mathcal{O}(1)$.

- As a second case, we take $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ such that $\|\Delta\boldsymbol{\mu}\| = \mathcal{O}(1)$ (thus meeting the minimal requirement on the means distance rate) and $\mathbf{C}_1$, $\mathbf{C}_2$ such that $\|\mathbf{C}_1 - \mathbf{C}_2\| = o(1)$. We thus set $\mathbf{C}_1 = \mathbf{C}$ and $\mathbf{C}_2 = \mathbf{C} + \mathbf{E}$ with $\|\mathbf{E}\| = o(1)$. We can then write

$$\begin{aligned} E_S &= \frac{1}{p}\operatorname{tr}\mathbf{C}(\mathbf{C}+\mathbf{E})^{-1} - 1 + \frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_2^{-1}\Delta\boldsymbol{\mu} - \frac{1}{p}\log\frac{|\mathbf{C}|}{|(\mathbf{C}+\mathbf{E})|} \\ &= \frac{1}{p}\operatorname{tr}\left(\mathbf{I}_p + \mathbf{C}^{-\frac{1}{2}}\mathbf{E}\mathbf{C}^{-\frac{1}{2}}\right)^{-1} - 1 + \frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-\frac{1}{2}}\left(\mathbf{I}_p + \mathbf{C}^{-\frac{1}{2}}\mathbf{E}\mathbf{C}^{-\frac{1}{2}}\right)^{-1}\mathbf{C}^{-\frac{1}{2}}\Delta\boldsymbol{\mu} \\ &\quad + \frac{1}{p}\log\det\left(\mathbf{I}_p + \mathbf{C}^{-\frac{1}{2}}\mathbf{E}\mathbf{C}^{-\frac{1}{2}}\right). \end{aligned}$$

By using the Taylor expansion $(\mathbf{I}_p+\mathbf{E})^{-1} = \mathbf{I}_p-\mathbf{E}+\mathbf{E}^2+\mathcal{O}(\|\mathbf{E}\|^3)$ (with $\|\mathbf{E}\| = o(1)$), we get

$$E_S = \frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{\mu} + \frac{2}{p}\operatorname{tr}(\mathbf{C}^{-1}\mathbf{E})^2 + o(p^{-1}).$$

so that the choice $\|\mathbf{E}\| = \mathcal{O}(p^{-\frac{1}{2}})$ is necessary to have $E_S \sim_p \sqrt{V_S} = \mathcal{O}(p^{-1})$ with

$$V_S = \frac{2}{p^2}\operatorname{tr}(\mathbf{C}^{-1}\mathbf{E})^2 + \frac{4}{p^2}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{\mu} + o(p^{-2}). \tag{6.8}$$

To sum up, when the class means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and class covariances $\mathbf{C}_1, \mathbf{C}_2$ are perfectly known, one can achieve non-trivial classification when

$$\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = \mathcal{O}(1) \tag{6.9}$$

$$\|\mathbf{C}_1 - \mathbf{C}_2\| = \mathcal{O}\left(\frac{1}{\sqrt{p}}\right). \tag{6.10}$$

We note in particular that Condition (6.10) implies

$$\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2) = \mathcal{O}(\sqrt{p}) \tag{6.11}$$

$$\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = \mathcal{O}(1) \tag{6.12}$$

with Condition 6.12 following from Cauchy-Schwarz inequality.

Conditions (6.9)–(6.12) constitute an optimal baseline for Gaussian Mixture Models classification that no statistical learning method can achieve. Those will be our reference in the sequel for the performance evaluation of (unsupervised) kernel spectral clustering.

6.4 Random-matrix asymptotics of inner product kernel spectral clustering

We consider $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$ n independent vectors belonging to k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ such that

$$\mathbf{x}_{n_1+\cdots+n_{a-1}+1}, \ldots, \mathbf{x}_{n_1+\cdots+n_a} \in \mathcal{C}_a$$

with $n_0 = 0$ and $\sum_i n_i = n$. We assume that when $\mathbf{x}_i \in \mathcal{C}_a$, $\mathbf{x}_i = \boldsymbol{\mu}_a + \sqrt{p}\mathbf{w}_i$ where $\mathbf{w}_i \sim \mathcal{N}(0, \mathbf{C}_a/p)$ with $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and $\mathbf{C}_a$ a non negative definite matrix. We will consider the following growth rate assumptions for $n, p \to \infty$ leading to non-trivial classification errors

Assumption 3 (Growth rate)**.** *As $n \to \infty$, we assume the following conditions hold.*

1. *For $c_n = \frac{p}{n}$, $0 < \liminf_n c_n \leq \limsup_n c_n < \infty$.*

2. *For each* $a \in \{1, \ldots, k\}$ *and for* $c_a = \frac{n_a}{n}$, *we have* $0 < \liminf_n c_a \leq \limsup_n c_a < \infty$. *We shall denote in the sequel* $c = \{c_a\}_{a=1}^k$.

3. *Let* $\boldsymbol{\mu}^\circ = \sum_{a=1}^k c_a \boldsymbol{\mu}_a$ *and for each* $a \in \{1, \ldots, k\}$, $\boldsymbol{\mu}_a^\circ = \boldsymbol{\mu}_a - \boldsymbol{\mu}^\circ$. *We have* $\limsup_n \max_{1 \leq a \leq k} \|\boldsymbol{\mu}_a^\circ\| < \infty$.

4. *Let* $\mathbf{C}^\circ = \sum_{a=1}^k c_a \mathbf{C}_a$ *and for each* $a \in \{1, \ldots, k\}$, $\mathbf{C}_a^\circ = \mathbf{C}_a - \mathbf{C}^\circ$. *We have* $\limsup_n \max_{1 \leq a \leq k} \|\mathbf{C}_a\| < \infty$ *and* $\limsup_n \max_{1 \leq a \leq k} \frac{1}{\sqrt{n}} \operatorname{tr} \mathbf{C}_a^\circ = \mathcal{O}(1)$.

For subsequent use, we introduce the following notations

$$\begin{aligned}
\mathbf{M} &\triangleq [\boldsymbol{\mu}_1^\circ, \cdots, \boldsymbol{\mu}_k^\circ] \in \mathbb{R}^{p \times k} \\
\mathbf{T} &\triangleq \left\{ \frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ \right\}_{a,b=1}^k \\
\mathbf{W} &\triangleq [\mathbf{w}_1, \ldots, \mathbf{w}_n] \in \mathbb{R}^{p \times n} \\
\mathbf{J} &\triangleq [\mathbf{j}_1, \cdots, \mathbf{j}_k] \in \mathbb{R}^{n \times k} \\
\mathbf{P} &\triangleq \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \in \mathbb{R}^{n \times n}
\end{aligned}$$

with $\mathbf{j}_a \in \mathbb{R}^n$ the canonical vector of cluster $\mathcal{C}_a$ defined by $(\mathbf{j}_a)_i = \boldsymbol{\delta}_{\mathbf{x}_i \in \mathcal{C}_a}$.

Items $1-2$ of Assumption 3 set the big data regime under consideration usually called the *random matrix regime*. Item 3 corresponds to the optimal distance rate for the means (Equation (6.9)) while Item 4 correspond to the optimal distance rate for the covariances (Equation (6.11)).

We take the Kernel matrix $\mathbf{K}$ to be: $\mathbf{K} \triangleq \left[f\left(\frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} \right) \right]_{i,j=1}^n$ where $\mathbf{x}_i^\circ = \mathbf{x}_i - \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j$ is the recentered data $\mathbf{x}_i$ and f satisfies the following conditions

Assumption 4 (On the kernel function)**.** *The kernel function* f *is three-times continuously differentiable in a neighborhood of* 0 *with* $f(0), f'(0), f''(0), f'''(0)$ *constant with* p.

It is common usage in the literature to require that f is such that the kernel matrix $\mathbf{K}$ is non-negative definite. Examples of kernels ensuring this property are:

- Linear kernel:
$$f(x) = x$$
- Exponential kernel:
$$f(x) = \exp(\beta x) \quad \beta \geq 0$$
- Polynomial kernel:
$$f(x) = (x + c)^d \quad c \geq 0, d \in \mathbb{N}$$

where the conditions $\beta \geq 0$ and $c \geq 0$ with $d \in \mathbb{N}$ are here imposed to ensure the positivity of the resulting kernel matrix $\mathbf{K}$. The rationale behind this choice stems from the fact that $\mathbf{K}$ is thought of as being the Gram matrix of the observations embedded in a higher dimensional feature space. As will be shown in this work and also through a series of numerical illustrations, the positivity of the kernel function is not mandatory, clustering being also possible when the kernel matrix is not non-negative: *what matters in high dimensions is the behavior of f and its first derivatives around 0 (for inner products kernels).*

As $p \to \infty$, we have

- For $i = j$, $\frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} = \frac{1}{p} \operatorname{tr} \mathbf{C}_a + \mathcal{O}(p^{-\frac{1}{2}}) = \frac{1}{p} \operatorname{tr} \mathbf{C}^\circ + \frac{1}{p} \operatorname{tr} \mathbf{C}_a^\circ + \mathcal{O}(p^{-\frac{1}{2}}) = \underbrace{\frac{1}{p} \operatorname{tr} \mathbf{C}^\circ}_{\tau} + \mathcal{O}(p^{-\frac{1}{2}})$.

- For $i \neq j$, $\frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} = \mathcal{O}(p^{-\frac{1}{2}})$.

Since for $i \neq j$, $\frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} \to 0$ as $p \to \infty$ and for $i = j$, $\frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} \to \tau$, we shall write a Taylor expansion of f around 0 to give an estimate of the entries K_{ij} $(i \neq j)$ and a Taylor expansion around τ to get approximations for entries K_{ii}. We thus have for $i \neq j$,

$$K_{ij} = f(0) + f'(0) \frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} + f''(0) \left(\frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} \right)^2 + k_p^{ij}$$

while

$$K_{ii} = f(\tau) + f'(\tau) \frac{\|\mathbf{x}_i^\circ\|^2}{p} + f''(\tau) \left(\frac{\|\mathbf{x}_i^\circ\|^2}{p} \right)^2 + k_p^{ii}$$

where k_p^{ij} and k_p^{ii} represent terms vanishing asymptotically. We have for $\mathbf{x}_i \in \mathcal{C}_a$ and $\mathbf{x}_j \in \mathcal{C}_b$

$$\begin{aligned} \frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} &= \frac{\left((\boldsymbol{\mu}_a^\circ)^\mathsf{T} + \sqrt{p} (\mathbf{w}_i^\circ)^\mathsf{T} \right) \left(\boldsymbol{\mu}_b^\circ + \sqrt{p} \mathbf{w}_j^\circ \right)}{p} \\ &= \underbrace{\frac{(\boldsymbol{\mu}_a^\circ)^\mathsf{T} \boldsymbol{\mu}_b^\circ}{p}}_{\mathcal{O}(p^{-1})} + \underbrace{\frac{(\boldsymbol{\mu}_a^\circ)^\mathsf{T} \mathbf{w}_j^\circ}{\sqrt{p}}}_{\mathcal{O}(p^{-1})} + \underbrace{\frac{(\mathbf{w}_i^\circ)^\mathsf{T} \boldsymbol{\mu}_b^\circ}{\sqrt{p}}}_{\mathcal{O}(p^{-1})} + \underbrace{(\mathbf{w}_i^\circ)^\mathsf{T} \mathbf{w}_j^\circ}_{\mathcal{O}(p^{-\frac{1}{2}})} \end{aligned}$$

and

$$\left(\frac{(\mathbf{x}_i^\circ)^T\mathbf{x}_j^\circ}{p}\right)^2 = \underbrace{\left(\frac{(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ}{p}\right)^2}_{\mathcal{O}(p^{-2})} + \underbrace{\left(\frac{(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\mathbf{w}_j^\circ}{\sqrt{p}}\right)^2}_{\mathcal{O}(p^{-2})} + \underbrace{\left(\frac{(\mathbf{w}_i^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ}{\sqrt{p}}\right)^2}_{\mathcal{O}(p^{-2})} + \underbrace{\left((\mathbf{w}_i^\circ)^\mathsf{T}\mathbf{w}_j^\circ\right)^2}_{\mathcal{O}(p^{-1})} + \underbrace{\frac{2}{p\sqrt{p}}(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ(\boldsymbol{\mu}u_a^\circ)^\mathsf{T}\mathbf{w}_j^\circ}_{\mathcal{O}(p^{-2})}$$
$$+ \underbrace{\frac{2}{p\sqrt{p}}(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ(w_i^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ}_{\mathcal{O}(p^{-2})} + \underbrace{\frac{2}{p}(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ(w_i^\circ)^\mathsf{T}\mathbf{w}_j^\circ}_{\mathcal{O}(p^{-2})} + \underbrace{\frac{2}{p}(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\mathbf{w}_j^\circ(\mathbf{w}_i^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ}_{\mathcal{O}(p^{-2})}$$
$$+ \underbrace{\frac{2}{\sqrt{p}}(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\mathbf{w}_j^\circ(\mathbf{w}_i^\circ)^\mathsf{T}\mathbf{w}_j^\circ}_{\mathcal{O}(p^{-2})} + \underbrace{\frac{2}{\sqrt{p}}(\mathbf{w}_i^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ(\mathbf{w}_i^\circ)^\mathsf{T}\mathbf{w}_j^\circ}_{\mathcal{O}(p^{-2})}$$

We shall now write the matrix $\mathbf{K}$ in such a way that its corresponding terms have non-vanishing operator norm. The matrices corresponding to the terms of $\left(\frac{(\mathbf{x}_i^\circ)^T\mathbf{x}_j^\circ}{p}\right)^2$ except $\left((\mathbf{w}_i^\circ)^\mathsf{T}\mathbf{w}_j^\circ\right)^2$ will give rise to matrices with vanishing operator norms. Based on that, we can write

$$\mathbf{K} = f(0)\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + f'(0)\left(\left[(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p}\right]_{a,b=1}^k + \left[\mathbf{1}_{n_a}\frac{(\boldsymbol{\mu}_a^\circ)^\mathsf{T}}{\sqrt{p}}(\mathbf{WP})_b\right]_{a,b=1}^k + \left[(\mathbf{WP})_a^\mathsf{T}\frac{(\boldsymbol{\mu}_b^\circ)}{\sqrt{p}}\mathbf{1}_{n_b}^\mathsf{T}\right]_{a,b=1}^k\right.$$
$$\left.+\mathbf{PW}^\mathsf{T}\mathbf{WP}\right) + f''(0)\left(\left[\frac{1}{p}\operatorname{tr}\mathbf{C}_a\mathbf{C}_b\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p}\right]_{a,b=1}^k\right) + \left(f(\tau) - f(0) - f'(0)(\tau)\right)\mathbf{I}_n + \mathcal{O}(p^{-\frac{1}{2}}).$$

By rearranging the terms of $\mathbf{K}$, we get the following approximation

Theorem 29. *Under Assumption 3 and 4, let $\hat{\mathbf{K}}$ be given by:*

$$\hat{\mathbf{K}} = f'(0)\mathbf{PW}^\mathsf{T}\mathbf{WP} + \mathbf{V}\boldsymbol{\Omega}\mathbf{V}^\mathsf{T} \tag{6.13}$$

with

$$\boldsymbol{\Omega} = \begin{pmatrix} f'(0)\mathbf{M}^\mathsf{T}\mathbf{M} + \frac{f''(0)}{2}\left\{\frac{1}{p}\operatorname{tr}\mathbf{C}_a^\circ\mathbf{C}_b^\circ\right\}_{a,b=1}^k & f'(0)\mathbf{I}_k \\ f'(0)\mathbf{I}_k & \mathbf{0}_k \end{pmatrix},$$
$$\mathbf{V} = \left[\frac{\mathbf{J}}{\sqrt{p}}, \mathbf{PW}^T\mathbf{M}\right]$$

Then,

$$\left\|[\mathbf{PKP} - (f(\tau) - f(0) - \tau f'(0))\mathbf{P}] - \hat{\mathbf{K}}\right\| \xrightarrow{\text{a.s.}} 0.$$

Up to recentering and scaling (by the matrix $\mathbf{P}$), the kernel matrix $\mathbf{K}$ takes the form of an additive spiked random matrix $\hat{\mathbf{K}}$ as long as

$$\operatorname{tr}\mathbf{C}_a^\circ\mathbf{C}_b^\circ = \mathcal{O}(p). \tag{6.14}$$

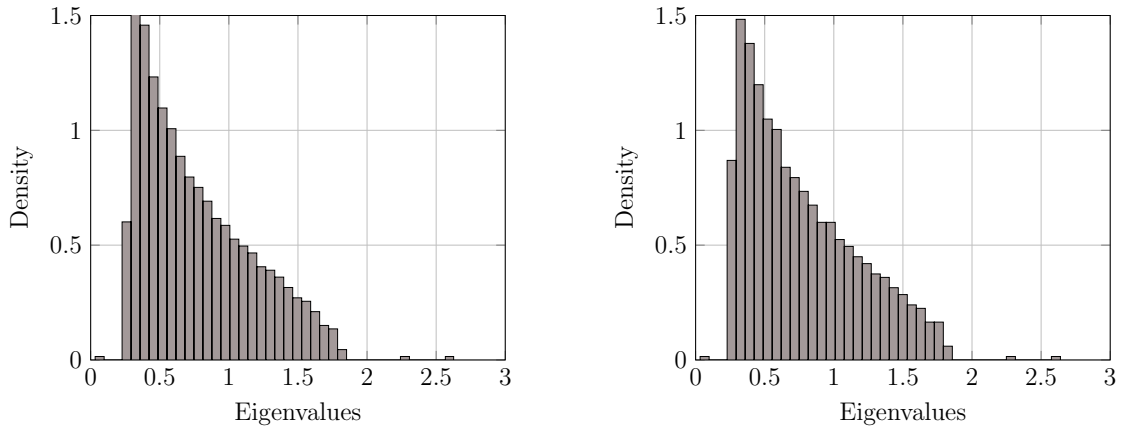


Figure 6.1: Eigenvalues of $\mathbf{K}$ (left) and $\hat{\mathbf{K}}$ (right) for $p = 2048$, $n = 1024$, $c_1 = 1/2, c_2 = c_3 = 1/4$, $[\mu_i]_j = 4\boldsymbol{\delta}_{ij}$, $\mathbf{C}_i = (1 + 6(i-1)/\sqrt{(}p))\mathbf{I}_p$, $f(x) = \exp(x/2)$

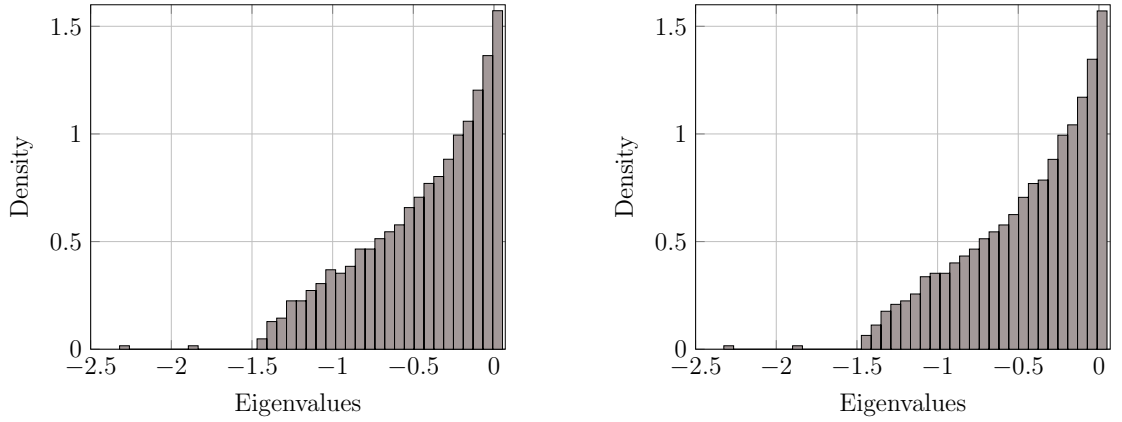


Figure 6.2: Eigenvalues of $\mathbf{K}$ (left) and $\hat{\mathbf{K}}$ (right) for $p = 2048$, $n = 1024$, $c_1 = 1/2, c_2 = c_3 = 1/4$, $[\mu_i]_j = 4\boldsymbol{\delta}_{ij}$, $\mathbf{C}_i = (1 + 6(i-1)/\sqrt{(}p))\mathbf{I}_p$, $f(x) = \exp(-x/2)$

The noise matrix $\mathbf{P}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{P}$ is a *deformed* random Wishart matrix (Marčenko Pastur) and the low rank matrix $\mathbf{V}\boldsymbol{\Omega}\mathbf{V}^\mathsf{T}$ contains linear combinations of the canonical vectors $\mathbf{j}_a$ scaled with inner products between the class means ($\mathbf{M}^\mathsf{T}\mathbf{M}$) and class covariances products ($\operatorname{tr}\mathbf{C}_a^\circ\mathbf{C}_b^\circ$). There is thus a phase transition beyond which eigenvalues of $\mathbf{K}$ isolate from the bulk. Spectral clustering using the eigenvectors of $\mathbf{K}$ associated with the isolated eigenvalues will induce non trivial classification performance since those eigenvectors ought to be correlated to the class canonical vectors. Figure 6.1 shows the spectrum of $\mathbf{K}$ and that of $\hat{\mathbf{K}}$ which can be seen to be asymptotically equivalent. Both present a bulk of the same shape (a deformed Marčenko Pastur) followed by two isolated eigenvalues (spikes).

The expression of $\hat{\mathbf{K}}$ reveals that the clustering performance does not depend on the positivity of the kernel matrix. The only condition that needs to be required is the appearance of isolated eigenvalues. Figure 6.2 reproduces the same experiment as Figure 6.1 with a kernel function $f(x) = \exp(-x/2)$. As can be seen, in such a scenario, isolated

eigenvalues arise at the left side of the spectrum, suggesting that the smallest eigenvalues carry the information about clustering.

Compared to the optimal oracle condition for the covariances rates (condition (6.12)), the condition (6.14) implies that the kernel functions f satisfying the growth rates studied so far, are not optimal. For kernel functions such that $f'(0) = 0$, Equation (6.13) reads

$$\mathbf{PKP} = (f(\tau) - f(0))\mathbf{P} + \frac{f''(0)}{2p}\mathbf{J}\left\{\frac{1}{p}\operatorname{tr}\mathbf{C}_a^\circ\mathbf{C}_b^\circ\right\}_{a,b=1}^{k}\mathbf{J}^\mathsf{T} + \mathcal{O}(p^{-\frac{1}{2}}), \tag{6.15}$$

then leading to a completely deterministic kernel matrix (absence of noise matrix) and thus asymptotic classification can be achieved when using such kernels under the distance rates on the class means (Assumption 3) and class covariances (Condition (6.14)) above. This means that with such kernels, one can reduce the rate of $\operatorname{tr}\mathbf{C}_a^\circ\mathbf{C}_b^\circ$ from $\mathcal{O}(p)$ to $\mathcal{O}(p^{\frac{1}{2}})$ to achieve non trivial classification error since in that case a new noise order term $\left((\mathbf{w}_i^\circ)^\mathsf{T}\mathbf{w}_j^\circ\right)^2$ (which was negligeable in the previous suboptimal regime) is comparable to $\operatorname{tr}\mathbf{C}_a^\circ\mathbf{C}_b^\circ$ in order of magnitude. This case was investigated in [Kammoun and Couillet, 2017] where f is chosen such that $f'(0) = 0$. It was shown in [Kammoun and Couillet, 2017] that with such kernel functions ($f'(0) = 0$), the limiting eigenvalue distribution of $\mathbf{K}$ is a semi-circle distribution instead of a Marčenko-Pastur distribution (in the case $f'(0) \neq 0$). However, while this kernel choice $f'(0) = 0$ leads to an improvement in the distance rate between products of covariances $\operatorname{tr}\mathbf{C}_a^\circ\mathbf{C}_b^\circ$, it has the undesirable effect of completely annihiling the class means as can be seen in Equation (6.15).

In the next section, we propose a new family of kernel functions designed to discriminate both statistical means and covariances to theoretically minimal distance rates (with respect to data size p).

6.5 Random matrix-improved kernels for large dimensional spectral clustering

We consider the same data setting as in Section 6.4 but for the following growth rates

Assumption 5 (Growth rate). *As $n \to \infty$, $p/n \to c_0 > 0$, $\frac{n_a}{n} \to c_a > 0$. Furthermore,*

1. *For $\boldsymbol{\mu}^\circ = \sum_{a=1}^k c_a\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_a^\circ = \boldsymbol{\mu}_a - \boldsymbol{\mu}^\circ$, $\|\boldsymbol{\mu}_a^\circ\| = \mathcal{O}(1)$.*
2. *For $\mathbf{C}^\circ = \sum_{a=1}^k c_a\mathbf{C}_a$ and $\mathbf{C}_a^\circ = \mathbf{C}_a - \mathbf{C}^\circ$, $\|\mathbf{C}_a\| = \mathcal{O}(1)$ and $\operatorname{tr}\mathbf{C}_a^\circ = \mathcal{O}(\sqrt{p})$.*
3. *$\frac{1}{\sqrt{p}}\operatorname{tr}\mathbf{C}_a^\circ\mathbf{C}_b^\circ$ converges in $[0, \infty)$.*
4. *$\frac{1}{p}\operatorname{tr}\mathbf{C}^\circ$ converges to $\tau > 0$.*

Under those conditions, it was shown in [Couillet and Benaych-Georges, 2016, Remark 12] and in [Tiomoko Ali et al., 2018a] that estimating the class labels by spectral clustering

on $\mathbf{K}$ does not perform better than random guess for generic f, unless the condition $\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ = \mathcal{O}(1)$ is relaxed to $\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ = \mathcal{O}(\sqrt{p})$. However, when f is chosen so that $f'(\tau) = 0$ for translation-invariant kernels or $f'(0) = 0$ for inner-product kernels, spectral clustering on $\mathbf{K}$ induces non trivial classification on datasets with differing class covariances. But the latter choice comes along with a complete annihilation of the class means in the spectral clustering inner workings (a setting carefully studied in [Kammoun and Couillet, 2017]).

As made clear by a careful random matrix analysis, setting instead $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$ (or $f'(\tau) = \mathcal{O}(p^{-\frac{1}{2}})$ for translation-ivariant kernels) allows for a fair treatment of both class means and covariances in the classification procedure. We focus here on the case of inner-product kernels with $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$. We thus have the following key assumption on the kernel function design.

Assumption 6 (On the kernel function)**.** *The kernel function f is three-times continuously differentiable in a neighborhood of 0 with $f(0), f''(0), f'''(0)$ constant with p while $f'(0) = \frac{\alpha}{\sqrt{p}}$ for some $\alpha \in \mathbb{R}$. We shall also denote $\beta = \frac{1}{2} f''(0)$.*

For instance, the kernels $f(x) = \beta(x + p^{-\frac{1}{2}}\beta^{-1}\alpha)^2$ or $f(x) = e^{-\beta(x+p^{-\frac{1}{2}}\beta^{-1}\alpha)^2}$ satisfy the conditions of Assumption 6.

The kernel studied in [Kammoun and Couillet, 2017] thus corresponds to the particular case of Assumption 6 with $\alpha = 0$. As for [Couillet and Benaych-Georges, 2016,Tiomoko Ali et al., 2018a], we shall see that it might be considered as a limiting setting where α is arbitrarily large.

Having specified the conditions on f, let us now define $\mathbf{K}$ as the inner-product random matrix

$$\mathbf{K} \triangleq \begin{cases} f\left(\frac{1}{p}(\mathbf{x}_i^\circ)^\mathsf{T} \mathbf{x}_j^\circ\right) & , i \neq j \\ 0 & , i = j \end{cases}$$

with $\mathbf{x}_i^\circ = \mathbf{x}_i - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ and f satisfying Assumption 6. Setting the diagonal elements of $\mathbf{K}$ is done for mathematical convenience and has no impact on the spectral clustering performance. Under this parametrization, we shall successively show that the matrix $\mathbf{K}$ composed of non-linear and intricately dependent entries asymptotically behaves in a simpler "almost linear" manner. From this simplified form, the asymptotic spectral characterization of $\mathbf{K}$ will be understood, in particular its dominant eigenvector contents.

As in [El Karoui et al., 2010] and following-up works, the non-linearity in $\mathbf{K}$ is treated by noticing that, as $p \to \infty$, $K_{ij} \to 0$ for all $i \neq j$, thereby allowing for an entry-wise Taylor expansion of $\mathbf{K}$. The theoretical difficulty next lies in the random matrix analysis of

all matrix terms arising from the Taylor expansion. The key particularity that makes the setting $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$ so fundamental is that, in [Couillet and Benaych-Georges, 2016], the terms affected by the differences $\operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ$ used to vanish (as a result of being absorbed by background noise) when $\operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ = \mathcal{O}(\sqrt{p})$; by letting $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$, the dominant background noise (but also the differences in means) are reduced and now comparable to the terms involving $\operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ$ (as long as $\beta = \frac{1}{2} f''(0) \neq 0$). An interesting side effect is that a second noise term then arises, and leads to a peculiar phenomenon where a mixture between a Marcenko–Pastur [Marchenko and Pastur, 1967] type and a semi-circle type [Wigner, 1993] noise eigenvalue distribution is observed in the limiting spectrum of $\mathbf{K}$. Still, this complication in the "noise spectrum" paradoxically comes along with a much simplified "signal spectrum", as shown in the subsequent results.

Theorem 30. *Under Assumption 1 and 6, let* $\hat{\mathbf{K}}$ *be given by:*

$$
\begin{aligned}
\hat{\mathbf{K}} &= \alpha \mathbf{P} \mathbf{W}^T \mathbf{W} \mathbf{P} + \beta \mathbf{P} \mathbf{\Phi} \mathbf{P} + \mathbf{V} \mathbf{\Omega} \mathbf{V}^T \\
\mathbf{\Omega} &= \begin{bmatrix} \alpha \mathbf{M}^T \mathbf{M} + \beta \mathbf{T} & \alpha \mathbf{I}_k \\ \alpha \mathbf{I}_k & \mathbf{0}_k \end{bmatrix} \\
\mathbf{V} &= \left[\frac{\mathbf{J}}{\sqrt{p}}, \mathbf{P} \mathbf{W}^T \mathbf{M} \right] \\
\frac{\mathbf{\Phi}}{\sqrt{p}} &= \left\{ ((\boldsymbol{\omega}_i^\circ)^T \boldsymbol{\omega}_j^\circ)^2 \boldsymbol{\delta}_{i \neq j} \right\}_{i,j=1}^n - \left\{ \frac{\operatorname{tr}(\mathbf{C}_a \mathbf{C}_b)}{p^2} \mathbf{1}_{n_a} \mathbf{1}_{n_b}^T \right\}_{a,b=1}^k .
\end{aligned}
$$

Then,

$$
\left\| \sqrt{p} \left(\mathbf{P} \mathbf{K} \mathbf{P} + \left(f(0) + \tau f'(0) \right) \mathbf{P} \right) - \hat{\mathbf{K}} \right\| \xrightarrow{\text{a.s.}} 0.
$$

Theorem 30 (proved in Appendix C) states that, up to centering and scaling, $\mathbf{K}$ is asymptotically equivalent to $\hat{\mathbf{K}}$. In particular, an immediate corollary of Theorem 30 is that both matrices asymptotically share (again, up to centering and scaling) the same eigenvalues as well as *isolated* eigenvectors (i.e., eigenvectors associated to eigenvalues found at non-vanishing distance from any other eigenvalue). We may then study the asymptotic spectral properties of $\mathbf{K}$ (and as a result, the classification performance of algorithms based on $\mathbf{K}$) through $\hat{\mathbf{K}}$.

As previously hinted at, it is first interesting to note that $\hat{\mathbf{K}}$ is the sum of i) the random matrices $\alpha \mathbf{P} \mathbf{W}^T \mathbf{W} \mathbf{P}$ (of the Marcenko–Pastur type) and $\beta \mathbf{P} \mathbf{\Phi} \mathbf{P}$ (of the Wigner type, as shown in [Couillet et al., 2016c, Kammoun and Couillet, 2017]) having entries of order $\mathcal{O}(p^{-1})$ and of ii) a maximum rank $k-1$ matrix containing linear combinations of the class-wise step vectors $\mathbf{j}_a$ intricately scaled through the inner-products between class means ($\mathbf{M}^T \mathbf{M}$) and class covariance-products ($\mathbf{T}$). This may be identified as part of the large family of *spiked random matrix models* [Benaych-Georges and Nadakuditi, 2012], with the particularity that the low-rank addition is not independent of the noise part and

that the noise part itself is a mixture between random Wishart and random symmetric matrices.

Random matrix theory today possesses all necessary tools to assess the eigenspectrum of such spiked random matrix models. As a common denominator, their eigenvalues are usually composed of a tightly connected "bulk" of eigenvalues along with up to $k-1$ isolated eigenvalues, the eigenvectors associated with which align to some extent to the eigenvectors in $\mathbf{V}$ (and thus, importantly here, to linear combinations of the vectors $\mathbf{j}_1, \ldots, \mathbf{j}_k$).

In particular, understanding the asymptotic performance of spectral clustering demands to characterize the isolated eigenvectors of $\mathbf{K}$. For these to be asymptotically informative, their associated eigenvalues must be found away from the main eigenvalue "bulk". In the following results, we evaluate the conditions upon which this transition phenomenon (i.e., the appearance of spiked eigenvalues) between asymptotically uninformative and informative eigenvectors occurs. We start by identifying the defining equations for the eigenvalue distribution of $\mathbf{K}$.

Theorem 31 (Bulk of Eigenvalues)**.** *Let Assumptions 5 hold. Then, as $p \to \infty$, the spectral distribution $\nu_n \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{\lambda_i(\hat{\mathbf{K}})}$ (with $\lambda_i(\mathbf{X})$ the eigenvalues of $\mathbf{X}$) almost surely converges (in the weak sense of probability measures) to the probability measure ν defined on a compact support $\mathcal{S}$ and having Stieltjes transform $m(z) = \int \frac{\nu(dt)}{t-z}$ defined for $z \in \mathbb{C}^+$, as the unique solution in $\mathbb{C}^+$ of*

$$\frac{1}{m(z)} = -z + \frac{\alpha}{p}\operatorname{tr}\mathbf{C}^\circ\left(\mathbf{I}_p + \frac{\alpha m(z)}{c_0}\mathbf{C}^\circ\right)^{-1} - \frac{2\beta^2}{c_0}\omega^2 m(z)$$

where $\omega = \lim_{p\to\infty}\frac{1}{p}\operatorname{tr}(\mathbf{C}^\circ)^2$.

Figure 6.3 shows for different values of the parameters α and β the histogram of the eigenvalues of $\mathbf{K}$ versus the theoretical bulk ν from Theorem 31.[1]. Note that ν is indeed a mixture of the Marcenko–Pastur law (more visible when $\alpha \gg \beta$) and a Wigner semi-circle law (especially appearant as $\beta \gg \alpha$). The regime under study thus exhibits a tradeoff between the regime considered in [Couillet and Benaych-Georges, 2016] (where α is theoretically infinite and only a Marcenko–Pastur law appears in the theoretical formulas) and the regime considered in [Couillet et al., 2016c] (where $\alpha = 0$ and a semi-circle law is obtained).

With Theorem 31 in place, it now remains to determine the conditions under which isolated eigenvalues can be found in the spectrum of $\mathbf{K}$, i.e., eigenvalues falling outside the support $\mathcal{S}$ of the limiting measure ν. This is obtained by means of now standard random matrix techniques (see e.g., [Benaych-Georges and Nadakuditi, 2012]) dedicated to spiked models. The main result is provided in Theorem 32 below, explicited here for simplicity under the following assumption:

[1] Obtained from the inverse formula $\nu(dt) = \frac{1}{\pi}\lim_{\epsilon\to 0}\Im[m(t+i\epsilon)]dt$

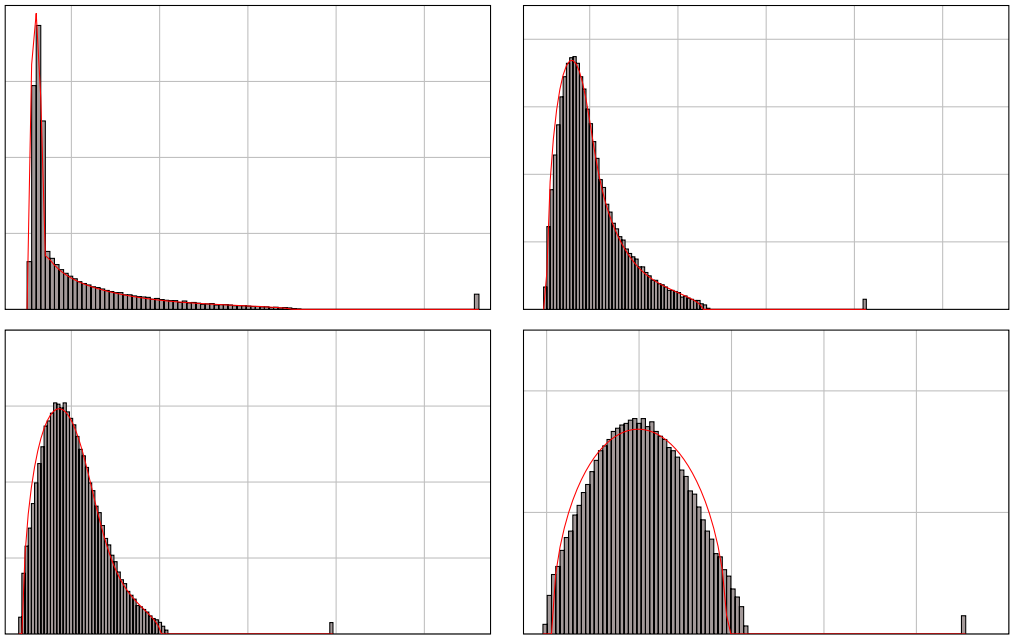


Figure 6.3: Eigenvalues of $\mathbf{K}$ (up to recentering) versus limiting law, $p = 2048$, $n = 4096$, $k = 2$, $n_1 = n_2$, $\boldsymbol{\mu}_i = 3\boldsymbol{\delta}_i$, $f(x) = \frac{1}{2}\beta\left(x + \frac{1}{\sqrt{p}}\frac{\alpha}{\beta}\right)^2$. (**Top left**): $\alpha = 8, \beta = 1$, (**Top right**): $\alpha = 4, \beta = 3$, (**Bottom left**): $\alpha = 3, \beta = 4$, (**Bottom right**): $\alpha = 1, \beta = 8$.

Assumption 7 (Symmetrical scenario)**.** *$k = 2$ with $n_1 = n_2 = \frac{n}{2}$.*

Theorem 32. *Let Assumption 5–6–7 hold and let $\rho \in \mathbb{R} \setminus \mathcal{S}$ be such that*

$$\frac{m(\rho)}{4c_0}(\alpha g(\rho)\delta + \beta\theta) + 1 = 0 \tag{6.16}$$

with $g(\rho) = \frac{1}{p}\operatorname{tr}(\mathbf{I}_p + \frac{\alpha m(\rho)}{c_0}\mathbf{C}^\circ)^{-1}$, $\delta = \|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\|^2$, and $\theta = \frac{1}{\sqrt{p}}\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$.

Then, there exists λ_j eigenvalue of $\hat{\mathbf{K}}$ such that

$$|\lambda_j - \rho| \xrightarrow{\text{a.s.}} 0.$$

Any real number ρ satisfying equation (6.16) therefore corresponds to the (almost sure) limit of some eigenvalue of $\mathbf{K}$ (again, up to a shift and scaling). This equation in general has a solution for sufficiently large differences in class means, through the Euclidean norm distance δ, or in class covariances, through the Frobenius norm distance θ. This induces a detectability phase transition depending on the values of the pair (δ, θ). Thus, for sufficiently large δ or θ and appropriately set α, β, Equation 6.16 has a solution, which implies the presence of an isolated eigenvalue outside $\mathcal{S}$ and to a corresponding eigenvector "aligned to some extent" to the canonical class vectors $\mathbf{j}_a$'s. Specifically, as hinted in Chapter 3, for every isolated eigenvalue λ of $\mathbf{K}$, the associated eigenvector $\mathbf{u}_\lambda$

can be written as a linear combination of the class canonical vectors added to residual noise. Since the data are statistically interchangeable within the classes, we can write

$$\mathbf{u}_\lambda = \eta_1 \frac{\mathbf{j}_1}{\sqrt{n_1}} + \eta_2 \frac{\mathbf{j}_2}{\sqrt{n_2}} + \sigma_1 \boldsymbol{\omega}_1 + \sigma_2 \boldsymbol{\omega}_2 \tag{6.17}$$

where $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are unit norm vectors supported respectively on the indices of class $\mathcal{C}_1$ and $\mathcal{C}_2$, and orthogonal to respectively $\mathbf{j}_1$ and $\mathbf{j}_2$. The scalars η_1, η_2 can be seen as the empirical averages of the eigenvector entries in class $\mathcal{C}_1$ and $\mathcal{C}_2$ while the σ_1 and σ_2 represent the class standard deviations of the eigenvector fluctuations around $\eta_1 \frac{\mathbf{j}_1}{\sqrt{n_1}}$ and $\eta_2 \frac{\mathbf{j}_2}{\sqrt{n_2}}$. Intuitively, the larger $|\eta_1 - \eta_2|$ the more the separation between eigenvector entries mapped to $\mathcal{C}_1$ and those mapped to $\mathcal{C}_2$ and thus the better the clustering performance. A precise analysis of the limiting values of those parameters (similar to the approach in [Couillet and Benaych-Georges, 2016]) leads to the following result.

Theorem 33 (Isolated eigenvector)**.** *Under the assumptions of Theorem 32, let λ be an isolated eigenvalue of $\hat{\mathbf{K}}$ with almost sure limit ρ, and $\mathbf{u}_\lambda$ its associated eigenvector decomposed as* (6.17)*. Then, for both $a = 1$ and $a = 2$*

$$(\eta_a)^2 = \frac{m(\rho)^2}{2m'(\rho)} \frac{1}{1 - \frac{m(\rho)^2}{4m'(\rho)} \frac{\alpha g'(\rho)}{c_0} \delta} + o(1)$$

where $m(\rho)$ and $g(\rho)$ are defined in Theorem 31 and $m'(\rho)$, $g'(\rho)$ are their respective first derivatives.

Under this model (i.e., for $k = 2$ with $n_1 = n_2$), the limiting structure of eigenvector u_λ is quite symmetric, as seen through the fact that $\eta_1 = -\eta_2 + o(1)$. This in particular immediately implies that $\sigma_1^2 = \sigma_2^2 + o(1) = \frac{1}{2} - \eta_1^2 + o(1)$.

Such a symmetric model can be for example obtained by letting $\mathbf{C}_a = \mathbf{I}_p + \sqrt{\frac{\theta}{2\kappa}} p^{-5/4} \mathbf{W}_a \mathbf{W}_a^\mathsf{T}$ for some $\kappa > 0$, and with $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{p \times \kappa p}$ two independent random matrices having i.i.d. $\mathcal{N}(0,1)$ entries so that $\frac{1}{\sqrt{p}} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 \xrightarrow{\text{a.s.}} \theta,$. In this case, the asymptotic correct classification $\mathbb{P}_c(\alpha, \beta)$ obtained by clustering eigenvector u_λ based on the signs of its entries satisfies

$$\mathbb{P}_c(\alpha, \beta) - Q\left(-\sqrt{\frac{\eta^2}{\frac{1}{2} - \eta^2}} \right) \xrightarrow{\text{a.s.}} 0 \tag{6.18}$$

where $\frac{1}{\eta^2} = \frac{2m'(\rho)}{m(\rho)^2}(1 - \frac{m(\rho)^2}{4m'(\rho)} \frac{\alpha g'(\rho)}{c_0} \delta)$.

As an illustration, Figure 6.4 depicts the limiting values for $\mathbb{P}_c(\alpha, \beta)$ as per (6.18) for different values of $\frac{\alpha}{\beta}$ and as a function of δ and θ. The figure strongly sets forth the importance of a proper choice of α, β depending on the specifics of the classification task, i.e., either means-dominant or covariance-dominant. In particular, as previously anticipated, a large value for $\frac{\alpha}{\beta}$ yields better performances in means-dominant disciminative tasks

(bottom of Figure 6.4); conversely, small values of $\frac{\alpha}{\beta}$ are adapted to covariance-dominant tasks (top of Figure 6.4).

Upon anticipation of the most discriminative attribute of the data at hand, our results therefore provide an instructive direction to appropriate kernel choice. In supervised or semi-supervised learning tasks, δ and θ can be estimated through appropriate (random matrix-based) estimators, thereby helping in the choice of appropriate values for α and β. An application of this principle is performed in the subsequent section on real datasets.

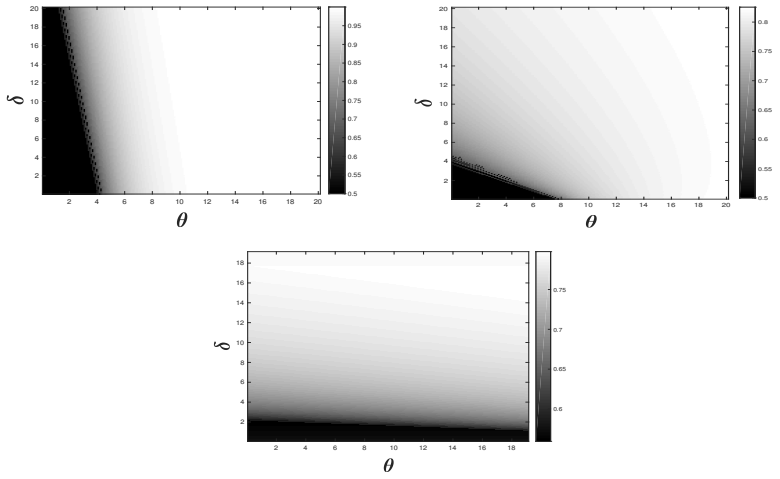


Figure 6.4: $\frac{p}{n} = \frac{1}{2}$, $k = 2$, $c_1 = c_2$, $\boldsymbol{\mu}_i = \delta\boldsymbol{\delta}_i$, $\delta \in [1:20]$, $\mathbf{C}_1, \mathbf{C}_2$ as in the symmetric setting with $\theta \in [1:20]$, $f(x) = \frac{1}{2}\beta\left(x + \frac{1}{\sqrt{p}}\frac{\alpha}{\beta}\right)^2$. Probability of correct recovery for different settings $\frac{\alpha}{\beta} = \frac{1}{8}$ (top), $\frac{\alpha}{\beta} = 1$ (Middle), $\frac{\alpha}{\beta} = 8$ (Bottom), a function of δ (x-axis) and θ (y-axis).

6.6 Applications

Our study has so far provided theoretical results for Gaussian mixture models, notably emphasizing the appropriateness of a kernel having first derivative scaling as $\mathcal{O}(p^{-\frac{1}{2}})$ with the data size p. In this section, we demonstrate that these findings are confirmed when applied to realistic datasets. The first dataset under consideration is the popular MNIST database of handwritten digits [LeCun, 1998]. In this dataset, the classes (the different digits) are evidently more discriminative in means than in covariances, as confirmed by Table 6.1. The second dataset is the epileptic EEG database from [Andrzejak et al., 2001] which consists of five sets (A to E), each containing $p = 100$ single EEG channel segments of $23.6s$ each. Sets A and B report measures on 5 healthy volunteers and sets $C - E$ on 5 epileptic patients; each set is composed of 4097 samples. This dataset demonstrates more variations in the class covariances as shown again in Table 6.1.

For both examples, kernel spectral clustering is performed on the dominant eigenvec-

Table 6.1: Class means and class covariances differences for some real datasets.

DATASETS	$\|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\|^2$	$\frac{1}{\sqrt{p}}\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$
MNIST (DIGITS $1, 7$)	613	1990
MNIST (DIGITS $3, 6$)	441	1119
MNIST (DIGITS $3, 8$)	212	652
EEG (SETS A, E)	2.4	109

tor of a subset of two classes (here $n = 1024$ samples), using k-means (rather than the eigenvector entry signs) to discriminate the classes. The results are depicted in Figure 6.5 for the MNIST data and Figure 6.6 for the EEG data. A clear observation is that extremely poor performances (proportion of correctly classified images) are obtained in the MNIST case for $\frac{\alpha}{\beta} \simeq 0$ while conversely extremely good performances are found on EEG for that setting, as was anticipated. Yet, note that the optimal value of $\frac{\alpha}{\beta}$ for the MNIST case does not demand that $\beta \to 0$; rather, an optimal value for $\frac{\alpha}{\beta}$ is found within the range $[0, 10]$, thereby suggesting that the differences in covariance are also exploited to some extent.

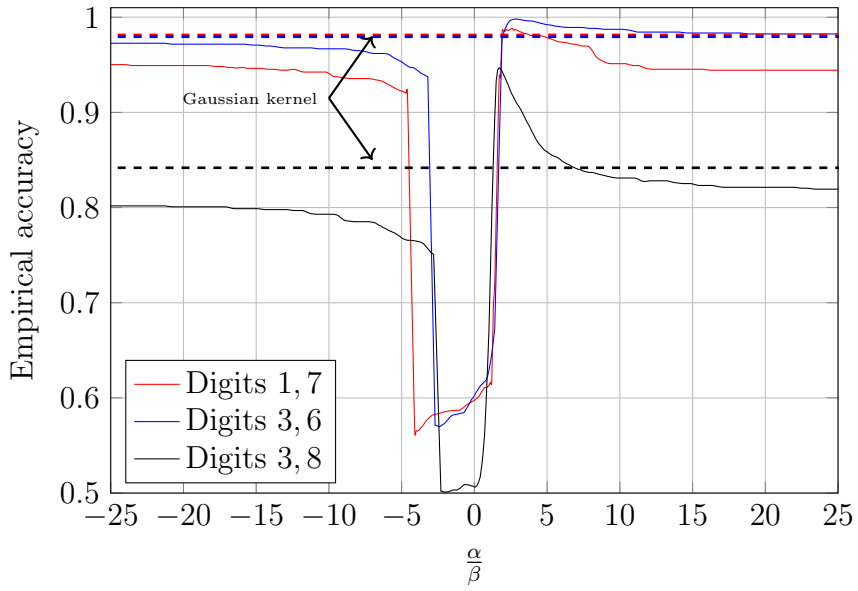


Figure 6.5: Spectral clustering of the MNIST database for varying $\frac{\alpha}{\beta}$ versus Gaussian kernel ($K_{ij} = e^{-\frac{1}{2}\|x_i - x_j\|^2}$).

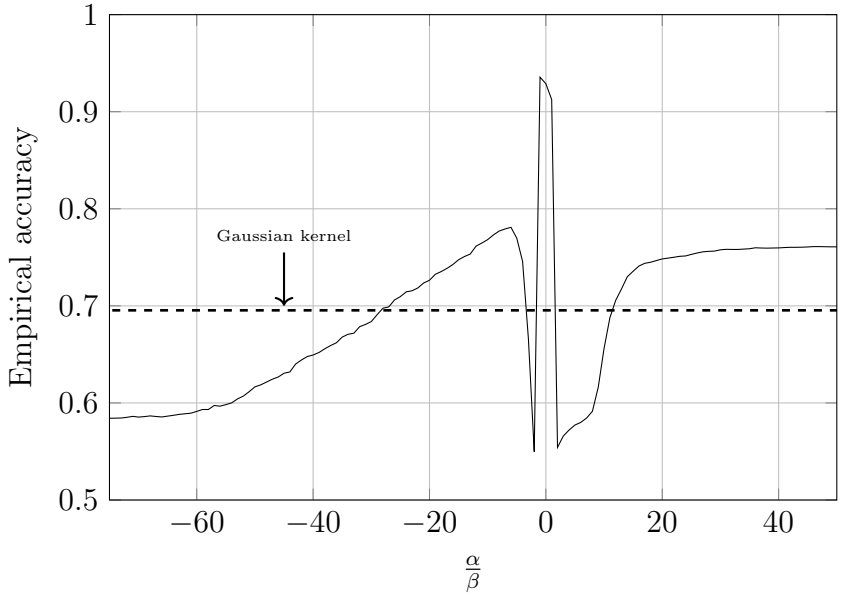


Figure 6.6: Spectral clustering of the EEG database for varying $\frac{\alpha}{\beta}$ versus Gaussian kernel ($K_{ij} = e^{-\frac{1}{2}\|x_i - x_j\|^2}$).

6.7 Conclusion

By leveraging on recent advances in random matrix theory, this chapter has proceeded to a complete study of the eigenspectrum of inner product kernel random matrices of finite mixture of large dimensional gaussian classes, in a non trivial regime of classification. This analysis shows the importance of the kernel function choice for a better exploitation of the discriminative power of kernel spectral methods. A new kernel function model is proposed and heavily relies on the need to balance statistical means and covariances in the data classes. The importance of such a kernel choice is confirmed on real datasets although the analysis is performed on Gaussian mixture models. So far, the proposed method which relies on fine-tuning the first derivative of the kernel function (through the value α) and the second derivative (through β), does not provide a clear recipe for performing such fine-tuning offline. The best ratio $\frac{\alpha}{\beta}$ might be chosen as the one maximizing the absolute difference between "the empirical average of the eigenvector entries mapped to the first class" and the "average of those mapped to the second class" since the larger this difference is, the better is the clustering performance. As can be seen in Theorem 33, there are unknown parameters (difference in means δ and difference in covariances θ) which need to be estimated prior to the optimization. Those can be easily estimated in supervised or semi-supervised settings but are more challenging in the unsupervised case. One approach might be to first run a spectral clustering on the data with a simple kernel, and estimate the class parameters under this partitioning, which can be used later to optimize on $\frac{\alpha}{\beta}$ but this comes with additional complexity and the estimated parameters might be biased since the clustering would be suboptimal with the "simple" kernel. Assuming the parameters (δ

and θ) are well estimated, one needs to find a computational efficient optimization of $\frac{\alpha}{\beta}$ better than a grid search in a certain range. Those different questions are left for future work.

Chapter 7

Conclusions and Perspectives

7.1 Data clustering

The larger part of this thesis has been focused on the theoretical understanding of large dimensional spectral clustering methods on graphs in the one hand and on data on the other hand. This involves the understanding of the eigenvalues and eigenvectors of large random matrices "not classical" in RMT as they either present some non linearities between the entries or dependency between them. We tackle this problem by benefiting from a concentration phenomena in a high dimensional regime where classification is not trivial (neither impossible nor perfect) which allows for a linearization around the limiting objects. The obtained matrices after linearization are shown to be asymptotically equivalent to spiked random matrices for which the treatment of the eigenvalues and eigenvectors is classical in RMT. The eigenspectrum analysis of those matrices allows us to provide some improvements over existing methods in dense heterogeneous community detection and in kernel spectral clustering. It might be objected that our limiting "non-trivial' assumption does not cover for all pratical classification problems; indeed, most competing articles in the field instead assume asymptotically perfect classification regimes. We believe that our setting is however most relevant to tackle classification questions were difficulties do arise, rather than studying problems where classification is deemed "simple". Besides, our practical experiments have shown multiple times that supposedly difficult classification tasks fall close to the expected outcomes under our "non-trivial" assumptions.

A smaller part of this thesis has been dedicated to the community detection problem in multilayer networks where some communities might be shared by the different layers while others might not be. Here, a spectral clustering algorithm would not work out except if the objective is to only detect shared communities in which case one can collapse the different layers into one graph on which a spectral method or other single-layer community detection approaches can be applied. We made use instead of a statistical inference method based on a multilayer weighted SBM with constraints on the heterogeneous structure of the communities.

In the following, we give possible future research directions in light of our findings, for each of the applications covered in this thesis.

7.1.1 Heterogenous single-layer community detection

The research has been focused on the challenging community detection problem on sparse graphs where classical methods based on *adjacency-family* of matrices are sub-optimal. In the sparse regime, the state-of-the art spectral methods close to the Bayes-optimal performances are the methods based on the Non-backtracking and the Bethe Hessian matrices. It was shown in [Gulikers et al., 2016] through a study of the Non-backtracking (NB) matrix in sparse DCSBM graph models, that no normalization is required even in the heterogeneous case as the eigenvectors of the NB matrix are shown to carry relevant information. However, as discussed in this thesis and in previous works [Gulikers et al., 2015,Newman, 2013], the eigenvectors of the similarity matrices of dense graphs are biased and proper normalizations are required. In addition to that, there has always been this recurring debate on the "best" similarity matrix to use for efficient spectral clustering in dense graphs. In this thesis, we try to provide an answer to that question by finding the similarity matrix for which the informative eigenvalues are well separated from the noise and we propose the proper normalization to perform on their corresponding eigenvectors prior to classification. But our study is so far limiting to a restricted range of similarity matrices, that is the class of $\mathbf{L}_\alpha$ matrices, and it is clear that there must exist more involved similarity matrix structures that better capture the information within the DCSBM model.

The statistical physics literature has provided new matrices (e.g., NB and Bethe Hessian matrices for sparse graphs) to use for spectral clustering in community detection, which are obtained from the linearization of the Belief Propagation (BP) algorithm and, as a consequence, likely (but not provably) have performances close to the optimal Bayes ones. In the dense SBM case, the linearization of BP leads to a spectral method on the so-called Fisher-score matrix [Lesieur et al., 2015, Lesieur et al., 2017] and is shown to exhibit a better phase transition than the adjacency family of matrices. As a future investigation, one might derive the expression for the Fisher score matrix in the dense DCSBM (by linearizing the BP equations for this specific model) and then conduct a similar analysis as in this thesis i.e., study the eigenvalues and eigenvectors of this matrix, derive the phase transition and compare with the one for the normalized modularity matrix (considered in this thesis), and find the processing to perform on the eigenvectors for better performances.

Pushing further the applicability reach of the present study, note that one of the remaining issues of spectral clustering methods especially for large graphs is the expensive computational complexity of the eigenvectors of the large random matrices representing those graphs. This computation burden is reduced when using power methods and is thus less expensive for sparse graphs (with many zeros in the corresponding matrix). This suggests potential computational gains incurred by smartly removing some edges in the graph to make it sparse prior to eigenvectors computation, which will of course reduce

the performance to the benefit of the computational cost. Our mathematical framework may allow us to study the tradeoff between complexity/cost and performances of spectral methods for community detection on such subsampled large dense graphs.

7.1.2 Kernel spectral clustering

Our random matrix framework enables us to analyze the eigenspectrum of inner-products kernel random matrices to understand the different mechanisms of kernel spectral clustering. This study reveals that the choice of the kernel functions is important depending on the specificities of the classification task (means dominant datasets, covariances dominant). We propose a family of kernel functions which are better tuned for taking into account the balance between the class means and class covariances. A similar study can be conducted with translation-invariant kernels (although much more involved) $f(\|\mathbf{x}_i-\mathbf{x}_j\|^2/p)$ with f such that $f'(\tau) = \mathcal{O}(p^{-\frac{1}{2}})$ (with τ the limiting value of $\|\mathbf{x}_i-\mathbf{x}_j\|^2/p$), which might provide further performance improvements.

As stated in Section 6.7, a next step following our analysis would be to optimize offline the proposed kernels in order to better fit the discriminative power of the datasets in hand. This can be performed using our obtained expressions for the empirical averages of the dominant eigenvectors entries by finding consistent estimates of the class statistics parameters and a good optimization framework. We obtained those expressions for the eigenvectors in the simpler case of 2 classes with the same proportions for which the optimization of the kernels is more accessible. A more general setting (more than 2 classes with unequal number of elements) needs to be investigated in the future.

The study of kernel random matrices was performed so far (in this work and in previous ones) on a Gaussian mixture model assumption. Although extremely close fit are obtained when considering MNIST and EEG datasets with their empirical means and covariances, other real-world datasets exhibit other moments than the means and covariances and are mostly treated through a first feature extraction procedure (such as through the popular HOG or VGG feature extraction in image processing). A study of spectral clustering on other more involved distributions such as heavy tailed ones, is left for a future investigation.

Beyond spectral clustering, we believe that this theoretical work might help designing more interesting kernel functions than classically used ones in many high-dimensional statistical learning problems such as kernel-based semi-supervised classification [Chapelle et al., 2003], kernel Support Vector Machine [Scholkopf and Smola, 2001], Generative Adversarial Networks (GANs) training using Kernel Inception Distance (KID) [Bińkowski et al., 2018]. For the latter, recent work [Bińkowski et al., 2018] has been considering the so-called Maximum Mean Discrepancy (MMD) as the loss metric for training GANs. The MMD is an integral probability metric to measure the distance between $\mathbb{P}$ and $\mathbb{Q}$, two probability distributions by using independent samples drawn from each distribution. Given samples $X = \{x_i\}_{i=1}^m$ and $Y = \{y_j\}_{j=1}^n$ drawn from $\mathbb{P}$ and $\mathbb{Q}$ respectively, it is shown

in [Gretton et al., 2012] that an unbiased estimator of the squared MMD is given by

$$MMD^2(X,Y) = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j), \tag{7.1}$$

where k is a "well-chosen" kernel function. It is discussed in [Bińkowski et al., 2018] that different kernel choices lead to the comparison of different statistics of the data distribution. For e.g., the simple linear kernel $k(x_i, x_j) = x_i^\mathsf{T} x_j$ only compares the means of the data distributions while more sophisticated kernels might compare means and/or covariances and/or skewness. Simple kernels of the type $k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ and $\left(1 + \frac{\|x_i - x_j\|^2}{\alpha}\right)^{-\alpha}$ with σ, α hand tuned, are used in [Bińkowski et al., 2018] to train GANs using the MMD as the training loss function and are shown to outperform other metrics for GANs learning with high dimensional features. We believe that similar studies as the ones conducted in this thesis, might help in designing better kernels for improving MMD-based GANs learning. A study of kernel matrices with entries drawn from heavy tailed distributions for instance might give intuitions on the kernel functions to use in order to compare the distributions based upon their means, covariances, skewness, high order statistics or mixture of those.

7.1.3 Multilayer community detection

The variational Bayes algorithm devised for the automatic detection of the shared and unshared communities between the different layers requires inputting the number of shared clusters and the number of private clusters for each layer. While the question of the number of clusters selection is an open problem in community detection in general, there exists some model selection approaches optimizing some metrics such as the Bayesian Information Criteria (BIC) to determine the number of clusters. The difficulty in the multilayer community detection problem is that there are many hyperparameters (number of clusters) to optimize. One might think of a BIC-like metric taking into account the joint likelihood of the multilayer graphical model, which will then be optimized on the different number of clusters.

The proposed approach assumes that the shared communities are common to all layers while a more realistic scenario would be that only subset of layers share communities. A future investigation would be to find a method that can encompass this more general case on weighted multilayer graphical models.

7.2 Random matrices as a tool to understand new and old machine learning methods

Spurred with the recent results of RMT in machine learning (ML), the former is undoubtedly a leading candidate as a theoretical tool for the understanding and improvement of big data machine learning algorithms. As stated throughout this thesis, most of the current ML algorithms are inconsistent in the large dimensional regime e.g., due to the "closeness" notion not exactly valid in high dimensions or inconsistency of classical estimators (empirical covariance matrix). Besides spectral clustering, those inconsistencies have been revealed and improved through a random matrix analysis of large dimensional semi-supervised classification [Mai and Couillet, 2017], least-square Support Vector Machine [Liao and Couillet, 2017] and the understanding of methods based on non linear random features maps [Louart et al., 2018, Liao and Couillet, 2018].

More importantly, one of the biggest challenges in ML nowadays is to provide some theoretical understanding to the astonishing performances of the popular deep neural networks which are to date, not well understood. Recent RMT works have tried to provide some theoretical explanations to the inner-workings behind deep learning by: computing their asymptotic performances in simple settings and architectures [Pennington and Worah, 2017, Louart et al., 2018], showing that some classes of activation functions have favorable properties [Pennington and Worah, 2017, Louart et al., 2018], showing that (under some approximation) deep neural network parameters shall converge to "good" local minima [Choromanska et al., 2015, Dauphin et al., 2014, Liao et al.,]. Those initial works leveraging random matrix theory as a core tool, although relying on simple cases or on some approximation, provide powerful grounds which might open the doors of the black box that is deep learning.

Appendix A

Supplementary material Chapter 4

Preliminaries

The random matrix under study $\mathbf{L}_\alpha$ is not a classically studied matrix in random matrix theory. We will thus first find in Section A.1 an approximate tractable random matrix $\tilde{\mathbf{L}}_\alpha$ which asymptotically preserves the eigenvalue distribution and the extreme eigenvectors of $\mathbf{L}_\alpha$. In Section A.2, we study the empirical distribution of the eigenvalues of $\mathbf{L}_\alpha$ and in Section A.3, we characterize the exact localizations of those eigenvalues. Finally, a thorough study of the eigenvectors associated to the aforementioned eigenvalues is investigated in Sections A.4 and A.5.

We follow here the proof technique of [Couillet and Benaych-Georges, 2016]. In the sequel, we will make some approximations of random variables in the asymptotic regime where $n \to \infty$. For the sake of random variables comparisons, we give the following stochastic definitions. For $x \equiv x_n$ a random variable and $u_n \geq 0$, we write $x = \mathcal{O}(u_n)$ if for any $\eta > 0$ and $D > 0$, $n^D \mathbb{P}(x \geq n^\eta u_n) \to 0$ as $n \to \infty$. For $\mathbf{v}$ a vector or a diagonal matrix with random entries, $\mathbf{v} = \mathcal{O}(u_n)$ means that the maximal entry of $\mathbf{v}$ in absolute value is $\mathcal{O}(u_n)$ in the sense defined previously. When $\mathbf{M}$ is a square matrix, $\mathbf{M} = \mathcal{O}(u_n)$ means that the operator norm of $\mathbf{M}$ is $\mathcal{O}(u_n)$. For $\mathbf{x}$ a vector or a matrix with random entries, $\mathbf{x} = o(u_n)$ means that there is $\kappa > 0$ such that $\mathbf{x} = \mathcal{O}(n^{-\kappa} u_n)$.

Most of the proofs here are classical in random matrix theory (see e.g., [Baik and Silverstein, 2006]) but require certain controls inherent to our model. The goal of the article not being an exhaustive development of the proofs techniques, we will admit a number of technical results already studied in the literature. However, we will exhaustively develop the calculus to obtain our final results which are not trivial.

A.0.1 Stein Lemma and Nash Poincare inequality

Lemma 34. *Let x be a standard real Gaussian random variable and $f : \mathbb{R} \to \mathbb{R}$ be a $\mathbb{C}^1$ function with first derivative $f^{'}(x)$ having at most polynomial growth. Then,*

$$\mathbb{E}[xf(x)] = \mathbb{E}[f^{'}(x)].$$

Lemma 35. *Let x be a standard real Gaussian random variable and $f : \mathbb{R} \to \mathbb{R}$ be a $\mathbb{C}^1$ function with first derivative $f^{'}(x)$. Then, we have*

$$\mathrm{Var}[f(x)] \leq \mathbb{E}[|f^{'}(x)|^2].$$

The proofs of those lemma can be found in [Pastur and Shcherbina, 2011].

A.1 Proof of Theorem 15

The matrix $\mathbf{L}_\alpha = (\mathbf{d}^\mathsf{T}\mathbf{1}_n)^\alpha \frac{1}{\sqrt{n}}\mathbf{D}^{-\alpha}\left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{\mathbf{d}^\mathsf{T}\mathbf{1}_n}\right]\mathbf{D}^{-\alpha}$ has non independent entries and is not a classical random matrix model. The idea is thus to approximate $\mathbf{L}_\alpha$ by a more tractable random matrix model $\tilde{\mathbf{L}}_\alpha$ in such a way that they share asymptotically the same set of outlying eigenvalues/eigenvectors which are of interest in our clustering scenario. We recall that condition to the knowledge of the intrinsic weights q_i's and of the connectivity matrix $\mathbf{M}$, the entries A_{ij} of the adjacency matrix are defined from the DCSBM model as independent Bernoulli random variables with parameter $q_i q_j\left(1 + \frac{M_{g_i g_j}}{\sqrt{n}}\right)$; one may thus write

$$A_{ij} = q_i q_j + q_i q_j \frac{M_{g_i g_j}}{\sqrt{n}} + X_{ij}$$

where X_{ij}, $1 \leq i, j \leq n$, are independent (up to the symmetry) zero mean random variables of variance $q_i q_j(1 - q_i q_j) + \mathcal{O}(n^{-\frac{1}{2}})$, since A_{ij} has mean $q_i q_j + q_i q_j \frac{M_{g_i g_j}}{\sqrt{n}}$ and variance $q_i q_j(1 - q_i q_j) + \mathcal{O}(n^{-\frac{1}{2}})$. We can then write the normalized adjacency matrix as follows

$$\frac{1}{\sqrt{n}}\mathbf{A} = \frac{1}{\sqrt{n}}\mathbf{q}\mathbf{q}^\mathsf{T} + \frac{1}{n}\left\{\mathbf{q}_{(a)}\mathbf{q}_{(b)}^\mathsf{T} M_{ab}\right\}_{a,b=1}^{K} + \frac{1}{\sqrt{n}}\mathbf{X} \tag{A.1}$$

$$= \underbrace{\frac{\mathbf{q}\mathbf{q}^\mathsf{T}}{\sqrt{n}}}_{\mathbf{A}_{d,\sqrt{n}}} + \underbrace{\frac{1}{n}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q}_{\mathbf{A}_{d,1}} + \underbrace{\frac{\mathbf{X}}{\sqrt{n}}}_{\mathbf{A}_{r,1}}, \tag{A.2}$$

where[1] $\mathbf{q}_{(i)} = [q_{n_1+\ldots+n_{i-1}+1}, \ldots, q_{n_1+\ldots+n_i}]^\mathsf{T} \in \mathbb{R}^{n_i}$ $(n_0 = 0)$, $\mathbf{X} = \{X_{ij}\}_{i,j=1}^n$ and $\mathbf{D}_q = \mathcal{D}(\mathbf{q})$. The idea of the proof is to write all the terms of $\mathbf{L}_\alpha$ based on Equation (A.2), since all those terms depend on $\mathbf{A}$. To this end, we will evaluate successively $\mathbf{d} = \mathbf{A}\mathbf{1}_n$,

[1] We recall that subscript 'd, n^k' stands for deterministic term whose operator norm is of order n^k and 'r, n^k' for random term with operator norm of order n^k.

$\mathbf{D} = \mathcal{D}(\mathbf{d})$, $\mathbf{d}\mathbf{d}^\mathsf{T}$ and $2m = \mathbf{d}^\mathsf{T}\mathbf{1}_n$. It will appear that $\mathbf{D}$ and $\mathbf{d}^\mathsf{T}\mathbf{1}_n$ are composed of dominant terms (with higher operator norm) and vanishing terms (with smaller operator norm); we may then proceed to writing a Taylor expansion of $\mathbf{D}^{-\alpha}$ and $(2m)^\alpha = (\mathbf{d}^\mathsf{T}\mathbf{1}_n)^\alpha$ for any α around their dominant terms to finally retrieve a Taylor expansion of $\mathbf{L}_\alpha$.

Let us start by developing the degree vector $\mathbf{d} = \mathbf{A}\mathbf{1}_n$. We have

$$\mathbf{d} = \mathbf{q}\mathbf{q}^\mathsf{T}\mathbf{1}_n + \frac{1}{\sqrt{n}}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n + \mathbf{X}\mathbf{1}_n = \mathbf{q}^\mathsf{T}\mathbf{1}_n\Big(\underbrace{\mathbf{q}}_{\mathcal{O}(n^{\frac{1}{2}})} + \underbrace{\frac{1}{\sqrt{n}}\frac{\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}}_{\mathcal{O}(n^{-\frac{1}{2}})} + \underbrace{\frac{\mathbf{X}\mathbf{1}_n}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}}_{\mathcal{O}(n^{-\frac{1}{2}})}\Big). \tag{A.3}$$

Let us then write the expansions of $\mathbf{d}^\mathsf{T}\mathbf{1}_n$, $(\mathbf{d}^\mathsf{T}\mathbf{1}_n)^\alpha$, $\mathbf{d}\mathbf{d}^\mathsf{T}$ and $\frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{(\mathbf{d}^\mathsf{T}\mathbf{1}_n)}$ respectively. From (A.3), we obtain

$$\mathbf{d}^\mathsf{T}\mathbf{1}_n = (\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2\Big[1 + \underbrace{\frac{1}{\sqrt{n}}\frac{\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}}_{\mathcal{O}(n^{-\frac{1}{2}})} + \underbrace{\frac{\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}}_{\mathcal{O}(n^{-\frac{1}{2}})}\Big]. \tag{A.4}$$

Thus for any α, proceeding to a 1^{st} order Taylor expansion, we may write

$$(\mathbf{d}^\mathsf{T}\mathbf{1}_n)^\alpha = (\mathbf{q}^\mathsf{T}\mathbf{1}_n)^{2\alpha}\Big[1 + \frac{\alpha}{\sqrt{n}}\underbrace{\frac{\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}}_{\mathcal{O}(n^{-\frac{1}{2}})} + \alpha\underbrace{\frac{\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}}_{\mathcal{O}(n^{-\frac{1}{2}})} + o(n^{-\frac{1}{2}})\Big]. \tag{A.5}$$

Besides, from (A.3) we have

$$\begin{aligned}\mathbf{d}\mathbf{d}^\mathsf{T} = (\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2\Big[&\underbrace{\mathbf{q}\mathbf{q}^\mathsf{T}}_{\mathcal{O}(n)} + \underbrace{\frac{1}{\sqrt{n}}\frac{\mathbf{q}\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}}_{\mathcal{O}(\sqrt{n})} + \underbrace{\frac{1}{\sqrt{n}}\frac{\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n\mathbf{q}^\mathsf{T}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}}_{\mathcal{O}(\sqrt{n})} + \underbrace{\frac{\mathbf{q}\mathbf{1}_n^\mathsf{T}\mathbf{X}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}}_{\mathcal{O}(\sqrt{n})} + \underbrace{\frac{\mathbf{X}\mathbf{1}_n\mathbf{q}^\mathsf{T}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}}_{\mathcal{O}(\sqrt{n})} \\ &+ \underbrace{\frac{1}{n}\frac{\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}}_{\mathcal{O}(1)} + \underbrace{\frac{1}{\sqrt{n}}\frac{\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{X}}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}}_{\mathcal{O}(1)} + \underbrace{\frac{1}{\sqrt{n}}\frac{\mathbf{X}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}}_{\mathcal{O}(1)} \\ &+ \underbrace{\frac{\mathbf{X}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{X}}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}}_{\mathcal{O}(1)} + o(1)\Big].\end{aligned} \tag{A.6}$$

Keeping in mind that we shall only need terms with non vanishing operator norms asymptotically, we will require $\frac{1}{\sqrt{n}}\left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{\mathbf{d}^\mathsf{T}\mathbf{1}_n}\right]$ to have terms with spectral norms of order at least $\mathcal{O}(1)$. We get from multiplying (A.6) and (A.5) (with $\alpha = -1$)

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{\mathbf{d}^\mathsf{T}\mathbf{1}_n} = &\frac{\mathbf{q}\mathbf{q}^\mathsf{T}}{\sqrt{n}} + \frac{1}{n}\frac{\mathbf{q}\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q}{\mathbf{q}^\mathsf{T}\mathbf{1}_n} + \frac{1}{n}\frac{\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n\mathbf{q}^\mathsf{T}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n} + \frac{1}{\sqrt{n}}\frac{\mathbf{q}\mathbf{1}_n^\mathsf{T}\mathbf{X}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n} + \frac{1}{\sqrt{n}}\frac{\mathbf{X}\mathbf{1}_n\mathbf{q}^\mathsf{T}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n} \\ &- \frac{1}{n}\frac{\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\mathbf{q}\mathbf{q}^\mathsf{T} - \frac{1}{\sqrt{n}}\frac{\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\mathbf{q}\mathbf{q}^\mathsf{T} + \mathcal{O}(n^{-\frac{1}{2}}).\end{aligned} \tag{A.7}$$

By subtracting (A.7) from (A.2), we obtain

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(\mathbf{A}-\frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{\mathbf{d}^\mathsf{T}\mathbf{1}_n}\right) &= \frac{1}{n}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q - \frac{1}{n}\frac{\mathbf{q}\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q}{\mathbf{q}^\mathsf{T}\mathbf{1}_n} - \frac{1}{n}\frac{\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n\mathbf{q}^\mathsf{T}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\\ &\quad + \frac{1}{n}\frac{\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\mathbf{q}\mathbf{q}^\mathsf{T} + \frac{\mathbf{X}}{\sqrt{n}} - \frac{1}{\sqrt{n}}\frac{\mathbf{q}\mathbf{1}_n^\mathsf{T}\mathbf{X}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n} - \frac{1}{\sqrt{n}}\frac{\mathbf{X}\mathbf{1}_n\mathbf{q}^\mathsf{T}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\\ &\quad + \frac{1}{\sqrt{n}}\frac{\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\mathbf{q}\mathbf{q}^\mathsf{T} + \mathcal{O}(n^{-\frac{1}{2}}). \end{aligned} \tag{A.8}$$

It then remains to evaluate $\mathbf{D}^{-\alpha}$. From (A.3), we may write $\mathbf{D} = \mathcal{D}(\mathbf{d})$ as

$$\mathbf{D} = \mathbf{q}^\mathsf{T}\mathbf{1}_n\Big(\underbrace{\mathbf{D}_q}_{\mathcal{O}(1)} + \underbrace{\mathcal{D}\left(\frac{1}{\sqrt{n}}\frac{\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\right)}_{\mathcal{O}(n^{-\frac{1}{2}})} + \underbrace{\mathcal{D}\left(\frac{\mathbf{X}\mathbf{1}_n}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\right)}_{\mathcal{O}(n^{-\frac{1}{2}})}\Big).$$

The right hand side of $\mathbf{D}$ (in brackets) having a leading term in $\mathcal{O}(1)$ and residual terms in $\mathcal{O}(n^{-\frac{1}{2}})$, the Taylor expansion of the $(-\alpha)$-power of $\mathbf{D}$ is then retrieved

$$\mathbf{D}^{-\alpha} = \left(\mathbf{q}^\mathsf{T}\mathbf{1}_n\right)^{-\alpha}\Big(\underbrace{\mathbf{D}_q^{-\alpha}}_{\mathcal{O}(1)} - \alpha\mathbf{D}_q^{-(\alpha+1)}\underbrace{\mathcal{D}\left(\frac{1}{\sqrt{n}}\frac{\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\right)}_{\mathcal{O}(n^{-\frac{1}{2}})} - \alpha\mathbf{D}_q^{-(\alpha+1)}\underbrace{\mathcal{D}\left(\frac{\mathbf{X}\mathbf{1}_n}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\right)}_{\mathcal{O}(n^{-\frac{1}{2}})} + \mathcal{O}(n^{-1})\Big). \tag{A.9}$$

By combining the expressions (A.5), (A.8) and (A.9), we obtain a Taylor approximation of $\mathbf{L}_\alpha$ as follows

$$\begin{aligned}\mathbf{L}_\alpha &= \mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha} + \frac{1}{n}\mathbf{D}_q^{1-\alpha}\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha} - \frac{1}{n}\frac{\mathbf{D}_q^{1-\alpha}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\\ &\quad - \frac{1}{n}\frac{\mathbf{D}_q^{1-\alpha}\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^{1-\alpha}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n} + \frac{1}{n}\frac{\mathbf{1}_n^\mathsf{T}\mathbf{D}_q\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\mathbf{D}_q^{1-\alpha}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^{1-\alpha} - \frac{1}{\sqrt{n}}\frac{\mathbf{D}_q^{1-\alpha}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{D}_q^{-\alpha}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\\ &\quad - \frac{1}{\sqrt{n}}\frac{\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^{1-\alpha}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n} + \frac{1}{\sqrt{n}}\frac{\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{1}_n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\mathbf{D}_q^{1-\alpha}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^{1-\alpha} + \mathcal{O}(n^{-\frac{1}{2}}).\end{aligned}$$

The three following arguments allow to complete the proof

- $\mathbf{1}_n = \mathbf{J}\mathbf{1}_K$ and $\mathbf{D}_q\mathbf{1}_n = \mathbf{q}$.
- We may write $(\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{q})_i = \frac{n_i}{n}\left(\frac{1}{n_i}\sum_{a\in\mathcal{C}_i} q_a\right)$. For classes of large sizes n_i, from the law of large numbers, $\left(\frac{1}{n_i}\sum_{a\in\mathcal{C}_i} q_a\right) \xrightarrow{\text{a.s.}} m_\mu$ and so, $\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{q} \xrightarrow{\text{a.s.}} m_\mu\mathbf{c}$ where we recall that $m_\mu = \int t\mu(\mathrm{d}t)$.
- As $\mathbf{X}$ is a symmetric random matrix having independent entries of zero mean and finite variance, from the law of large numbers, we have $\frac{1}{n}\frac{\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{1}_n}{\sqrt{n}} \xrightarrow{\text{a.s.}} 0$.

Using those three arguments, $\mathbf{L}_\alpha$ may be further rewritten

$$\begin{aligned}\mathbf{L}_\alpha = \mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha} + \frac{1}{n}\mathbf{D}_q^{1-\alpha}\mathbf{J}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha} - \frac{1}{n}\mathbf{D}_q^{1-\alpha}\mathbf{J}\mathbf{1}_K\mathbf{c}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}\\ - \frac{1}{n}\mathbf{D}_q^{1-\alpha}\mathbf{J}\mathbf{M}\mathbf{c}\mathbf{1}_K^\mathsf{T}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha} + \frac{1}{n}\mathbf{D}_q^{1-\alpha}\mathbf{J}\mathbf{1}_K\mathbf{c}^\mathsf{T}\mathbf{M}\mathbf{c}\mathbf{1}_K^\mathsf{T}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}\\ - \frac{1}{\sqrt{n}\mathbf{q}^\mathsf{T}\mathbf{1}_n}\mathbf{D}_q^{1-\alpha}\mathbf{J}\mathbf{1}_K\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{D}_q^{-\alpha} - \frac{1}{\sqrt{n}\mathbf{q}^\mathsf{T}\mathbf{1}_n}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{1}_n\mathbf{1}_K^\mathsf{T}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha} + \mathcal{O}(n^{-\frac{1}{2}}). \qquad \text{(A.10)}\end{aligned}$$

By rearranging the terms of (A.10), we obtain the expected result

$$\begin{aligned}\mathbf{L}_\alpha &= \mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\\ &\quad + \begin{bmatrix}\frac{\mathbf{D}_q^{1-\alpha}\mathbf{J}}{\sqrt{n}} & \frac{\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{1}_n}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\end{bmatrix}\begin{bmatrix}\left(\mathbf{I}_K - \mathbf{1}_K\mathbf{c}^T\right)\mathbf{M}\left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right) & -\mathbf{1}_K\\ -\mathbf{1}_K^T & 0\end{bmatrix}\begin{bmatrix}\frac{\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}}{\sqrt{n}}\\ \frac{\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{D}_q^{-\alpha}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\end{bmatrix} + \mathcal{O}(n^{-\frac{1}{2}}).\end{aligned}$$

This proves Theorem 15.

A.2 Proof of Theorem 17

It follows from Theorem 15 that $\tilde{\mathbf{L}}_\alpha = \mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$ is equivalent to an additive spiked random matrix [Chapon et al., 2012] where

$$\mathbf{V} = \begin{bmatrix}\frac{\mathbf{D}_q^{1-\alpha}\mathbf{J}}{\sqrt{n}} & \frac{\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{1}_n}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\end{bmatrix},$$

$$\mathbf{\Omega} = \begin{bmatrix}\left(\mathbf{I}_K - \mathbf{1}_K\mathbf{c}^T\right)\mathbf{M}\left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right) & -\mathbf{1}_K\\ -\mathbf{1}_K^T & 0\end{bmatrix},$$

with the difference that the deterministic part $\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$ is not independent of the random part $\mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha}$ (an issue that we solve here) and $\mathbf{V}$ is not composed of orthonormal vectors. Let us then study $\bar{\mathbf{X}} = \mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha}$ (having entries $\bar{X}_{ij}$ with zero mean and variance σ_{ij}^2/n with $\sigma_{ij}^2 = q_i^{1-2\alpha}q_j^{1-2\alpha}(1 - q_iq_j) + \mathcal{O}(n^{-\frac{1}{2}})$) and show that its empirical spectral distribution (e.s.d.) $\tilde{\pi}^\alpha$ converges weakly to $\bar{\pi}^\alpha$ with Stieljes transform $e_{00}^\alpha(z) = \int (t-z)^{-1}\,\mathrm{d}\bar{\pi}^\alpha(t)$ for $z \in \mathbb{C}^+$. This will imply (By Weyl interlacing formula) that the empirical spectral measure $\pi^\alpha \equiv \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(\tilde{\mathbf{L}}_\alpha)}$ (with $\lambda_i(\tilde{\mathbf{L}}_\alpha)$ eigenvalues of $\tilde{\mathbf{L}}_\alpha$) will also converge to $\bar{\pi}^\alpha$.

The matrix $\bar{\mathbf{X}}$ is a classical random matrix model in RMT already studied in similar cases [Pastur and Shcherbina, 2011]. It is well known for those random matrix models (having entries with given means, variances and bounded first order moments) that the law of the $\bar{X}_{ij}$'s does not change the results on the limiting law of the e.s.d. $\tilde{\pi}^\alpha$: this property is

kwown as *universality* (e.g., [Silverstein and Bai, 1995]). For technical reasons, we can thus assume that the $\bar{X}_{ij}$'s are Gaussian random variables with the same means and variances in order to use standard Gaussian calculus, introduced in [Pastur and Shcherbina, 2011]. The objective of the proof is to find the deterministic limit $e_{00}^{\alpha}(z)$ for the random quantity $\frac{1}{n}\operatorname{tr}\left(\bar{\mathbf{X}}-z\mathbf{I}_n\right)^{-1}$ which is the Stieljes transform of the e.s.d. $\tilde{\pi}^{\alpha}$. Deterministic equivalents for the Stieljes transform of empirical spectral measures associated with centered and symmetric random matrix models with a variance profile have already been studied in for example [Ajanki et al., 2015,Hachem et al., 2007]. We give in Appendix A.7 an exhaustive development of the Gaussian calculus to obtain $e_{00}^{\alpha}(z)$. The final result is as follows.

Lemma 36 (A first deterministic equivalent)**.** *Let* $\mathbf{Q}=(\bar{\mathbf{X}}-z\mathbf{I}_n)^{-1}$. *Then, for all* $z\in\mathbb{C}^+$,

$$\mathbf{Q}\leftrightarrow\bar{\mathbf{Q}}=\left(-z\mathbf{I}_n-\mathcal{D}\left(e_i(z)\right)_{i=1}^n\right)^{-1} \tag{A.11}$$

where $e_i(z)$ *the unique solution of* $e_i(z)=\frac{1}{n}\operatorname{tr}\mathcal{D}\left(\sigma_{ij}^2\right)_{j=1}^n\left(-z\mathbf{I}_n-\mathcal{D}\left(e_j(z)\right)_{j=1}^n\right)^{-1}$ *and the notation* $\mathbf{A}\leftrightarrow\mathbf{B}$ *stands for* $\frac{1}{n}\operatorname{tr}\mathbf{CA}-\frac{1}{n}\operatorname{tr}\mathbf{CB}\to 0$ *and* $\mathbf{d}_1^{\mathsf{T}}(\mathbf{A}-\mathbf{B})\mathbf{d}_2\to 0$ *almost surely, for all deterministic Hermitian matrix* $\mathbf{C}$ *and deterministic vectors* $\mathbf{d}_i$ *of bounded norms (spectral norm for matrices and Euclidian norm for vectors).*

From Lemma 36 (proof provided in Section A.7), we get directly $\frac{1}{n}\operatorname{tr}\mathbf{Q}-e_{00}^{\alpha}(z)\xrightarrow{\text{a.s.}}0$ with $e_{00}^{\alpha}(z)=\frac{1}{n}\sum_{i=1}^n\frac{1}{-z-e_i(z)}$. Observe now that

$$\begin{aligned}e_i(z)&=\frac{1}{n}\sum_{j=1}^n\frac{q_i^{1-2\alpha}q_j^{1-2\alpha}-q_i^{2-2\alpha}q_j^{2-2\alpha}}{-z-e_j(z)}\\&=q_i^{1-2\alpha}e_{11}^{\alpha}(z)-q_i^{2-2\alpha}e_{21}^{\alpha}(z)\end{aligned} \tag{A.12}$$

where

$$\begin{aligned}e_{11}^{\alpha}(z)&=\frac{1}{n}\sum_{j=1}^n\frac{q_j^{1-2\alpha}}{-z-q_j^{1-2\alpha}e_{11}^{\alpha}(z)+q_j^{2-2\alpha}e_{21}^{\alpha}(z)}\\e_{21}^{\alpha}(z)&=\frac{1}{n}\sum_{j=1}^n\frac{q_j^{2-2\alpha}}{-z-q_j^{1-2\alpha}e_{11}^{\alpha}(z)+q_j^{2-2\alpha}e_{21}^{\alpha}(z)}.\end{aligned} \tag{A.13}$$

Assuming now that the $q_i's$ are generated from an i.i.d law μ of compact support in $(0,1)$, by using the iterative deterministic equivalent approach devised in [Hoydis et al., 2011], one can show that Equation (A.13) is equivalent to

$$e_{00}^{\alpha}(z)=\int\frac{1}{-z-e_{11}^{\alpha}(z)q^{1-2\alpha}+e_{21}^{\alpha}(z)q^{2-2\alpha}}\mu(dq).$$

where for $z\in\mathbb{C}^+$ and $a,b\in\mathbb{Z}$ we define

$$e_{ab}^{\alpha}(z)=\int\frac{q^{a-2b\alpha}\mu(dq)}{-z-e_{11}^{\alpha}(z)q^{1-2\alpha}+e_{21}^{\alpha}(z)q^{2-2\alpha}}. \tag{A.14}$$

with $\mu(\mathrm{d}q) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \delta_{q_i}$. From this, we have that $e_{00}^{\alpha}(z)$ does not depend on n, so that $\frac{1}{n}\operatorname{tr}\mathbf{Q} \xrightarrow{\text{a.s.}} E_0^{\alpha}(z)$, $\tilde{\pi}^{\alpha} \to \bar{\pi}^{\alpha}$, and thus $\pi^{\alpha} \to \bar{\pi}^{\alpha}$ since $\tilde{\mathbf{L}}_{\alpha}$ and $\bar{\mathbf{X}}$ only differ by a finite rank matrix. This proves Theorem 17.

In the main core of the article, we have defined $e_{00}^{\alpha}(z) \triangleq m^{\alpha}(z)$, $e_{11}^{\alpha}(z) \triangleq f^{\alpha}(z)$ and $e_{21}^{\alpha}(z) \triangleq g^{\alpha}(z)$ for readability reasons. For future use, we define for $z, \tilde{z} \in \mathbb{C} \setminus \mathcal{S}^{\alpha}$

$$e_{ab;2}^{\alpha}(z,\tilde{z}) = \int \frac{q^{a-2b\alpha}\mu(dq)}{(-z - E_1^{\alpha}(z)q^{1-2\alpha} + E_2^{\alpha}(z)q^{2-2\alpha})(-\tilde{z} - E_1^{\alpha}(\tilde{z})q^{1-2\alpha} + E_2^{\alpha}(\tilde{z})q^{2-2\alpha})} \tag{A.15}$$

and

$$e_{ab;3}^{\alpha}(z,\tilde{z}) = \int \frac{q^{a-2b\alpha}\mu(dq)}{(-z - E_1^{\alpha}(z)q^{1-2\alpha} + E_2^{\alpha}(z)q^{2-2\alpha})^2(-\tilde{z} - E_1^{\alpha}(\tilde{z})q^{1-2\alpha} + E_2^{\alpha}(\tilde{z})q^{2-2\alpha})}. \tag{A.16}$$

We note that the canonical equations defining the stieltjes transform of the l.s.d of random symmetric matrices having independent entries with non-zero mean and a variance profile have first been introduced in [Girko, 2001, Girko, 2012] without the need of universality arguments. Those general results are applied in for e.g., [Avrachenkov et al., 2015] to the spectral analysis of Stochastic Block Models.

Convergence of the e_i's

Similar results to Lemma 36 have been derived for example in [Hachem et al., 2007] and the fixed point algorithm (C.7) which consists of iterating the e_i's is shown to converge. Since the calcultation of the e_{ab}'s is an intermediary step of (C.7) from (A.12), the fixed point algorithm (A.13) also converges. From the analyticity of the Stieljes transform outside its support, Lemma 36 extends naturally to $\mathbb{C} \setminus \mathcal{S}^{\alpha}$. This proves Theorem 17.

Remark 37. *Similarly to [Hachem et al., 2007], when none of the $(\mathbf{D}_q^{-\alpha})_{ii}$'s is isolated, the random matrix $\bar{\mathbf{X}}$ does not produce isolated eigenvalues outside the support $\mathcal{S}^{\alpha}$ of $\bar{\pi}^{\alpha}$. Here, for large n, this property is verified since from Assumption 1, the q_i's are i.i.d. arising from a law with compact support (the probability that a $(\mathbf{D}_q^{-\alpha})_{ii}$ gets isolated tends to 0 asymptotically). This gives Proposition 38 which we will not prove here; similar proofs are provided for example in [Bai and Silverstein, 1998].*

Proposition 38 (No eigenvalues outside the support)**.** *Following the statement of Theorem 17, let S_-^{α} and S_+^{α} be respectively the left and right edges of $\mathcal{S}^{\alpha}$. Then, for any $\epsilon > 0$, by letting $\mathcal{S}_{\epsilon}^{\alpha} = [S_-^{\alpha} - \epsilon; S_+^{\alpha} + \epsilon]$, for all large n almost surely,*

$$\left\{\lambda_i\left(\mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\right), 1 \le i \le n\right\} \cap (\mathbb{R} \setminus \mathcal{S}_{\epsilon}^{\alpha}) = \emptyset.$$

Remark 39. *The support $\mathcal{S}^{\alpha}$ is symmetric i.e., $\bar{\pi}^{\alpha}([a,b]) = \bar{\pi}^{\alpha}([-b,-a])$. We have in particular $S_-^{\alpha} = -S_+^{\alpha} = -S^{\alpha}$ where we denote $S_+^{\alpha} \triangleq \sup \mathcal{S}^{\alpha}$ and $S_-^{\alpha} \triangleq \inf \mathcal{S}^{\alpha}$.*

A.3 Proof of Theorem 19.

In the previous section, we have shown that the e.s.d. of $\mathbf{L}_\alpha$ converges weakly to the limiting law of the eigenvalues of $\bar{\mathbf{X}}$ since they only differ by a finite rank matrix. We shall have in addition isolated eigenvalues of $\mathbf{L}_\alpha$ induced by the aforementionned low rank matrix. We are interested here in the localization of eigenvalues of $\mathbf{L}_\alpha$ isolated from the support $\mathcal{S}^\alpha$ of the limiting law of its e.s.d. According to Proposition 38, there is almost surely no eigenvalue of $\bar{\mathbf{X}}$ at non-vanishing distance from $\mathcal{S}^\alpha$ asymptotically as $n \to \infty$ and hence the plausible isolated eigenvalues of $\mathbf{L}_\alpha$ are only due to the matrix $\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$. We follow classical random matrix approaches used for the study of the spectrum of spiked random matrices [Benaych-Georges and Nadakuditi, 2012, Chapon et al., 2012]. From Theorem 15, the eigenvalues ρ of $\mathbf{L}_\alpha$ falling at non-vanishing distance from the limiting support $\mathcal{S}^\alpha$ solve for large n, $0 = \det(\mathbf{L}_\alpha - \rho\mathbf{I}_n)$ almost surely for $\rho \notin \mathcal{S}^\alpha$. Since $\|\mathbf{L}_\alpha - \tilde{\mathbf{L}}_\alpha\| \xrightarrow{\text{a.s.}} 0$, $\rho_i(\mathbf{L}_\alpha) - \rho_i(\tilde{\mathbf{L}}_\alpha) \xrightarrow{\text{a.s.}} 0$ for all eigenvalues $\rho_i(\mathbf{L}_\alpha)$. We may then just solve $0 = \det(\mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T} - \rho\mathbf{I}_n)$. Now, as from Proposition 38, the random matrix $\bar{\mathbf{X}}$ does not have eigenvalues at non-vanishing distance from $\mathcal{S}^\alpha$ asymptotically, for $\rho \notin \mathcal{S}^\alpha$, we can thus factor and cancel out $\det(\bar{\mathbf{X}} - \rho\mathbf{I}_n)$ from the previous determinant equation, so that we are left to solve

$$0 = \det(\mathbf{I}_n + \mathbf{Q}_\rho^\alpha \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}) = \det(\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha \mathbf{V}\mathbf{\Omega})$$

where $\mathbf{Q}_\rho^\alpha = (\bar{\mathbf{X}} - \rho\mathbf{I}_n)^{-1}$. As we will show next, the matrix $\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V}\mathbf{\Omega}$ converges to a deterministic matrix, almost surely for large n. By the argument principle (similar to e.g., [Chapon et al., 2012]), the roots of $\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V}\mathbf{\Omega}$ are asymptotically those of the limiting matrix, with same multiplicity and it suffices to study the latter.

We then proceed to retrieving a limit for $\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V}\mathbf{\Omega}$. From Theorem 15, we have

$$\mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V} = \begin{pmatrix} \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}\mathbf{Q}_\rho^\alpha\mathbf{D}_q^{1-\alpha}\mathbf{J} & \frac{1}{\sqrt{n}(\mathbf{q}^\mathsf{T}\mathbf{1}_n)}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}\mathbf{Q}_\rho^\alpha\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{1}_n \\ \frac{1}{\sqrt{n}(\mathbf{q}^\mathsf{T}\mathbf{1}_n)}\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{D}_q^{-\alpha}\mathbf{Q}_\rho^\alpha\mathbf{D}_q^{1-\alpha}\mathbf{J} & \frac{1}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{D}_q^{-\alpha}\mathbf{Q}_\rho^\alpha\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{1}_n \end{pmatrix}.$$

The entries $(1,2)$, $(2,1)$ and $(2,2)$ of $\mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V}$ are random as they contain the random matrix $\mathbf{X}$ but tend to be deterministic in the limit. In fact, using the resolvent identity, we have that $\mathbf{Q}_\rho^\alpha\mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha} = \mathbf{I}_n + \rho\mathbf{Q}_\rho^\alpha$, the entry $(1,2)$ becomes $\frac{1}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n + \rho\frac{1}{\sqrt{n}(\mathbf{q}^\mathsf{T}\mathbf{1}_n)}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}\mathbf{Q}_\rho^\alpha\mathbf{D}_q^\alpha\mathbf{1}_n$ and the entry $(2,2)$ is equal to $\frac{n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\Big(\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{1}_n + \rho\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^{2\alpha}\mathbf{1}_n + \rho^2\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^\alpha\mathbf{Q}_\rho^\alpha\mathbf{D}_q^\alpha\mathbf{1}_n\Big)$. Now, we can freely use Lemma 36 to evaluate the limits of the entries of $\mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V}$ since all the terms are of the form $\mathbf{a}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{b}$ with $\mathbf{a}$ and $\mathbf{b}$ deterministic vectors. From Lemma 36, the entries $(1,1)$, $(1,2)$ and $(2,2)$ converge almost surely respectively to $\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}\bar{\mathbf{Q}}_\rho^\alpha\mathbf{D}_q^{1-\alpha}\mathbf{J}$, $\frac{1}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n + \rho\frac{1}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}\bar{\mathbf{Q}}_\rho^\alpha\mathbf{D}_q^\alpha\mathbf{1}_n$ and $\frac{n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\Big(\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{1}_n + \rho\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^{2\alpha}\mathbf{1}_n + \rho^2\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^\alpha\bar{\mathbf{Q}}_\rho^\alpha\mathbf{D}_q^\alpha\mathbf{1}_n\Big)$ for large n.

Now, using the fact that for any bounded continuous function f, from the law of large

numbers,

$$\frac{1}{n}\sum_{j\in\mathcal{C}_i} f(q_j) = \frac{n_i}{n}\frac{1}{n_i}\sum_{j\in\mathcal{C}_i} f(q_j) \xrightarrow{\text{a.s.}} c_i \int f(q)\mu(dq). \tag{A.17}$$

After some algebra, we obtain $\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}\bar{\mathbf{Q}}_\rho^\alpha\mathbf{D}_q^{1-\alpha}\mathbf{J} \xrightarrow{\text{a.s.}} e_{21}^\alpha(\rho)\mathcal{D}(\mathbf{c})$ where the e_{ij}'s are given in Theorem 17. Similarly for the terms $(1,2)$ and $(2,2)$, we obtain respectively

$$\frac{1}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)}\mathbf{J}^\mathsf{T}\mathbf{D}_q\mathbf{1}_n + \rho\frac{1}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)}\mathbf{J}^\mathsf{T}\mathbf{D}_q^{1-\alpha}\bar{\mathbf{Q}}_\rho^\alpha\mathbf{D}_q^\alpha\mathbf{1}_n \xrightarrow{\text{a.s.}} \left(1+\frac{\rho}{m_\mu}e_{10}^\alpha(\rho)\right)\mathbf{c}$$

and

$$\frac{n}{(\mathbf{q}^\mathsf{T}\mathbf{1}_n)^2}\left(\mathbf{1}_n^\mathsf{T}\mathbf{X}\mathbf{1}_n + \rho\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^{2\alpha}\mathbf{1}_n + \rho^2\mathbf{1}_n^\mathsf{T}\mathbf{D}_q^\alpha\bar{\mathbf{Q}}_\rho^\alpha\mathbf{D}_q^\alpha\mathbf{1}_n\right) \xrightarrow{\text{a.s.}} \frac{1}{m_\mu^2}\left(\rho v_\mu + \rho^2 e_{0;-1}^\alpha(\rho)\right)$$

with $v_\mu = \int q^{2\alpha}\mu(\mathrm{d}q)$ and where we have also used the fact that $\frac{1}{n}\mathbf{1}_n^\mathsf{T}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{1}_n \xrightarrow{\text{a.s.}} 0$ again from the law of large numbers.

The limit of $\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V}\boldsymbol{\Omega}$ is then obtained as

$$\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V}\boldsymbol{\Omega} \xrightarrow{\text{a.s.}} \begin{pmatrix} \mathbf{I}_K + e_{21}^\alpha(\rho)(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^\mathsf{T}) - \left(1+\frac{\rho}{m_\mu}e_{10}^\alpha(\rho)\right)\mathbf{c}\mathbf{1}_K^\mathsf{T} & -e_{21}^\alpha(\rho)\mathbf{c} \\ \frac{\rho}{m_\mu^2}\left(v_\mu + \rho e_{0;-1}^\alpha(\rho)\right)\mathbf{1}_K^\mathsf{T} & -\rho\frac{e_{10}^\alpha(\rho)}{m_\mu} \end{pmatrix}.$$

Using the Schur complement formula for the determinant of block matrices, we have that the determinant of the RHS matrix is zero whenever

$$-\rho\frac{e_{10}^\alpha(\rho)}{m_\mu}\det\Big[\mathbf{I}_K + e_{21}^\alpha(\rho)(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^\mathsf{T}) - \left(1+\frac{\rho}{m_\mu}e_{10}^\alpha(\rho)\right)\mathbf{c}\mathbf{1}_K^\mathsf{T} + \frac{\left(v_\mu + \rho e_{0;-1}^\alpha(\rho)\right)e_{21}^\alpha(\rho)}{m_\mu e_{10}^\alpha(\rho)}\mathbf{c}\mathbf{1}_K^\mathsf{T}\Big] = 0$$

or equivalently $\det(\underline{\mathbf{G}}_\rho^\alpha) = 0$ where

$$\underline{\mathbf{G}}_\rho^\alpha = \mathbf{I}_K + e_{21}^\alpha(\rho)(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^\mathsf{T}) + \theta^\alpha(\rho)\mathbf{c}\mathbf{1}_K^\mathsf{T}$$
$$\theta^\alpha(\rho) = -1 - \frac{\rho}{m_\mu}e_{10}^\alpha(\rho) + \frac{\left(v_\mu + \rho e_{0;-1}^\alpha(\rho)\right)e_{21}^\alpha(\rho)}{m_\mu e_{10}^\alpha(\rho)}.$$

The isolated eigenvalues ρ of $\mathbf{L}_\alpha$, which are the ρ for which $\det(\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho^\alpha\mathbf{V}\boldsymbol{\Omega}) = 0$, are then asymptotically the ρ such that $\det(\underline{\mathbf{G}}_\rho^\alpha) = 0$.

Remark 40 (Two types of isolated eigenvalues)**.** *From the previous paragraph,* $1+\theta^\alpha(\rho)$ *is an eigenvalue of* $\underline{\mathbf{G}}_\rho^\alpha$ *with associated left eigenvector* $\mathbf{1}_K$ *and right eigenvector* $\mathbf{c}$ *since* $\mathbf{1}_K^T\underline{\mathbf{G}}_\rho^\alpha = (1+\theta^\alpha(\rho))\,\mathbf{1}_K^T$ *and* $\underline{\mathbf{G}}_\rho^\alpha\mathbf{c} = (1+\theta^\alpha(\rho))\,\mathbf{c}$.

Letting ρ *be such that* $\det(\underline{\mathbf{G}}_\rho^\alpha) = 0$*, we can discriminate two cases*

- $1+\theta^\alpha(\rho)=0$*: isolated eigenvalues are found for those* $\rho \in \mathbb{R}\backslash\mathcal{S}^\alpha$ *such that* $1+\theta^\alpha(\rho)=0$*. We shall denote by* $\tilde{\rho}$ *such eigenvalues when they exist.*

- $1+\theta^\alpha(\rho)\neq 0$*: the left and right eigenvectors associated to the zero eigenvalues of* $\underline{\mathbf{G}}_\rho^\alpha$ *are respectively orthogonal to the right and left eigenvectors associated to the non-zero eigenvalues. So, by letting* $\mathbf{V}_l$*,* $\mathbf{V}_r$ *be matrices containing in columns the respectively left and right eigenvectors of* $\underline{\mathbf{G}}_\rho^\alpha$ *associated with the zero eigenvalues, we have* $\mathbf{V}_l^\mathsf{T}\mathbf{c}=\mathbf{0}$ *and* $\mathbf{1}_K^\mathsf{T}\mathbf{V}_r=\mathbf{0}$ *since* $1+\theta^\alpha(\rho)\neq 0$*. It is thus immediate that* $(\mathbf{V}_l,\mathbf{V}_r)$ *is also a pair of eigenvectors (with multiplicity) of* $\mathbf{I}_K + e_{21}^\alpha(\rho)\left(\mathcal{D}\left(\mathbf{c}\right)-\mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}\left(\mathbf{I}_K-\mathbf{c}\mathbf{1}_K^T\right)$ *associated to the zero eigenvalues.*

As we show in Section A.5, for $1+\theta^\alpha(\tilde{\rho})=0$, the eigenvectors associated to the aforementioned isolated eigenvalues $\tilde{\rho}$ will not contain information about the classes. This case is thus of no interest for clustering. It is nevertheless important from a practical viewpoint to note that, even in the absence of communities, spurious isolated eigenvalues may be found that may deceive the experimenter in suggesting the presence of node clusters. From now on, we will only consider the isolated eigenvalues ρ for which $1+\theta^\alpha(\rho)\neq 0$.

We now have all the ingredients to determine the conditions under which we may have eigenvalues of $\mathbf{L}_\alpha$ which isolate from $\mathcal{S}^\alpha$. Let l be a non zero eigenvalue of $\mathbf{G}_\rho^\alpha = (\mathcal{D}\left(\mathbf{c}\right)-\mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K-\mathbf{c}\mathbf{1}_K^T)$. Since $\det((\mathcal{D}(\mathbf{c})-\mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K-\mathbf{c}\mathbf{1}_K^\mathsf{T})) = \det((\mathbf{I}_K-\mathbf{c}\mathbf{1}_K^\mathsf{T})(\mathcal{D}(\mathbf{c})-\mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}) = \det((\mathcal{D}(\mathbf{c})-\mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M})$, l is also a non zero eigenvalue of $\bar{\mathbf{M}} = (\mathcal{D}(\mathbf{c})-\mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$. For each isolated eigenvalue ρ of $\mathbf{L}_\alpha$ we have a one-to-one mapping with a non zero eigenvalue l of $\bar{\mathbf{M}}$ such that $l=-\frac{1}{E_2^\alpha(\rho)}$. Hence, to show the existence of isolated eigenvalues of $\mathbf{L}_\alpha$, we need to solve for $\rho\in\mathbb{R}\setminus\mathcal{S}^\alpha$, $l=-\frac{1}{E_2^\alpha(\rho)}$ for each non zero eigenvalue l of $\bar{\mathbf{M}}$. Precisely, let us write $\mathcal{S}^\alpha=\bigcup_{m=1}^M[S_{m,-}^\alpha,S_{m,+}^\alpha]$ with $S_{1,-}^\alpha\leq S_{1,+}^\alpha<S_{2,-}^\alpha\leq\ldots<S_{M,+}^\alpha$ and define $S_{0,+}=-\infty$ and $S_{M+1,-}=+\infty$. Then, recalling that the Stieltjes transform of a real supported measure is necessarily increasing on $\mathbb{R}$, there exist isolated eigenvalues of $\mathbf{L}^\alpha$ in $(S_{m,+}^\alpha,S_{m+1,-}^\alpha)$, $m\in\{0,\ldots,M\}$, for all large n almost surely, if and only if there exists eigenvalues ℓ of $\bar{\mathbf{M}}$ such that

$$\lim_{x\downarrow S_{m,+}^\alpha} E_2^\alpha(x) < -\ell^{-1} < \lim_{x\uparrow S_{m+1,-}^\alpha} E_2^\alpha(x). \tag{A.18}$$

In particular, when $\mathcal{S}^\alpha=[S_-^\alpha,S_+^\alpha]$ is composed of a single connected component (as when $\mathcal{S}^\alpha$ is the support of the semi-circle law as well as most cases met in practice), then isolated eigenvalues of $\mathbf{L}^\alpha$ may only be found beyond S_+^α if $\ell>\lim_{x\downarrow S_+^\alpha}-\frac{1}{E_2^\alpha(x)}$ $(l>0)$ or below S_-^α if $\ell<\lim_{x\uparrow S_-^\alpha}-\frac{1}{E_2^\alpha(x)}$ $(l<0)$, for some non-zero eigenvalue ℓ of $\bar{\mathbf{M}}$. From the asymptotic spectrum of $\mathbf{L}_\alpha$, $S_-^\alpha=-S_+^\alpha$ as one can show that for any $z\in\mathbb{R}\setminus\mathcal{S}^\alpha$, $E_2^\alpha(-z)=-E_2^\alpha(z)$ so that both previous conditions reduce to $\mid\ell\mid>\lim_{x\downarrow S_+^\alpha}-\frac{1}{E_2^\alpha(x)}$. This proves Theorem 19.

The next section is advocated to the study of the eigenvectors associated to isolated eigenvalues of $\mathbf{L}_\alpha$.

A.4 Informative eigenvectors

In this section, in order to fully characterize the performances of Algorithm 3, we study in depth the normalized eigenvectors $\bar{\mathbf{n}}_i^\alpha$ used for the classification in the algorithm (step 3 of Algorithm 3). We consider here the eigenvectors corresponding to the eigenvalues for which $1+\theta^\alpha(\rho) \neq 0$ (when $1+\theta^\alpha(\rho)=0$, the corresponding eigenvectors do not contain any structural information about the classes; this case is treated in Section A.5). For technical reasons, we restrict ourselves here to those eigenpairs $(\lambda_i, \bar{\mathbf{n}}_i^\alpha)$'s for which there exists no $\lambda_j \neq \lambda_i$ such that, if $\lambda_i \to \rho$, $\lambda_j \to \rho$.

We recall that we may write $\bar{\mathbf{n}}_i^\alpha$ [2] as the "noisy plateaus" vector

$$\bar{\mathbf{n}}_i^\alpha = \sum_{a=1}^{K} \nu_i^a \frac{\mathbf{j}_a}{\sqrt{n_a}} + \sqrt{\sigma_{ii}^a}\mathbf{w}_i^a \tag{A.19}$$

where $\mathbf{w}_i^a \in \mathbb{R}^n$ is a random vector orthogonal to $\mathbf{j}_a$, of norm $\sqrt{n_a}$ and supported on the indices of $\mathcal{C}_a$ and

$$\nu_i^a = \frac{1}{\sqrt{n_a}} \left(\bar{\mathbf{v}}_i^\alpha\right)^\mathsf{T} \mathbf{j}_a = \frac{1}{\sqrt{n_a}} \frac{(\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{\alpha-1} \mathbf{j}_a}{\sqrt{(\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{2(\alpha-1)} \mathbf{u}_i^\alpha}} \tag{A.20}$$

$$\sigma_{ij}^a = \frac{(\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{\alpha-1} \mathcal{D}_a \mathbf{D}^{\alpha-1} \mathbf{u}_j^\alpha}{\sqrt{(\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{2(\alpha-1)} \mathbf{u}_i} \sqrt{(\mathbf{u}_j^\alpha)^\mathsf{T} \mathbf{D}^{2(\alpha-1)} \mathbf{u}_j^\alpha}} - \nu_i^a \nu_j^a \tag{A.21}$$

with $\mathcal{D}_a = \mathcal{D}(\mathbf{j}_a)$.

- We estimate the ν_i^a's by obtaining an estimator of the $K \times K$ matrix

  $$\frac{1}{n} \frac{\mathbf{J}^\mathsf{T} \mathbf{D}^{\alpha-1} \mathbf{u}_i^\alpha (\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{\alpha-1} \mathbf{J}}{(\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{2(\alpha-1)} \mathbf{u}_i^\alpha},$$

  the diagonal entries of which allow to estimate $|\nu_i^a|$ while the off-diagonal entries are used to decide on the signs of the ν_i^a's (up to a convention in the sign of $\mathbf{u}_i^\alpha$).

- Similarly, we first estimate the more involved object

  $$\frac{1}{n} \frac{\mathbf{J}^\mathsf{T} \mathbf{D}^{\alpha-1} \mathbf{u}_i^\alpha (\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{\alpha-1} \mathcal{D}_a \mathbf{D}^{\alpha-1} \mathbf{u}_j^\alpha (\mathbf{u}_j^\alpha)^\mathsf{T} \mathbf{D}^{\alpha-1} \mathbf{J}}{\left((\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{2(\alpha-1)} \mathbf{u}_i^\alpha\right) \left((\mathbf{u}_j^\alpha)^\mathsf{T} \mathbf{D}^{2(\alpha-1)} \mathbf{u}_j^\alpha\right)}$$

  from which $\frac{(\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{\alpha-1} \mathcal{D}_a \mathbf{D}^{\alpha-1} \mathbf{u}_j^\alpha}{\sqrt{(\mathbf{u}_i^\alpha)^\mathsf{T} \mathbf{D}^{2(\alpha-1)} \mathbf{u}_i^\alpha} \sqrt{(\mathbf{u}_j^\alpha)^\mathsf{T} \mathbf{D}^{2(\alpha-1)} \mathbf{u}_j^\alpha}}$ is retrieved by dividing any entry e, f of the former quantity by non-vanishing quantities $\nu_i^e \nu_i^f$. For the eigenvectors $\mathbf{u}_i^\alpha$ used for clustering, there is always at least one index f such that ν_i^f is non zero (otherwise, this eigenvector is of no use for clustering).

[2] Recall that the graph nodes were assumed labeled by class, and thus the entries of $\bar{\mathbf{n}}_i^\alpha$ are similarly sorted by class.

A.4.1 Evaluation of the class means ν_i^a's

The estimation of the ν_i^a's requires the evaluation of $\frac{1}{n}\frac{\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-\mathbf{1}}\mathbf{J}}{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\mathbf{2}(\alpha-\mathbf{1})}\mathbf{u}_i^\alpha}$ for $\mathbf{u}_i^\alpha$ eigenvector associated to a limiting isolated eigenvalue ρ with unit multiplicity of $\mathbf{L}_\alpha$. By residue calculus, we have that

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J} = -\frac{1}{2\pi i}\oint_{\Gamma_\rho}\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\left(\mathbf{L}_\alpha - z\mathbf{I}_n\right)^{-1}\mathbf{D}^{\alpha-1}\mathbf{J}\mathrm{d}z \tag{A.22}$$

for large n almost surely, where Γ_ρ is a complex (positively oriented) contour circling around the limiting eigenvalue ρ only. As from Theorem 15, $\mathbf{L}_\alpha = \mathbf{D}_q^{-\alpha}\frac{\mathbf{X}}{\sqrt{n}}\mathbf{D}_q^{-\alpha} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T} + o(1)$, we apply the Woodburry identity to the inverse in the previous integrand and we get

$$\begin{aligned}\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\left(\mathbf{L}_\alpha - z\mathbf{I}_n\right)^{-1}\mathbf{D}^{\alpha-1}\mathbf{J} &= \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{Q}_z^\alpha\mathbf{D}^{\alpha-1}\mathbf{J}\\ &\quad + \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}\left(\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}\right)^{-1}\mathbf{V}^\mathsf{T}\mathbf{Q}_z^\alpha\mathbf{D}^{\alpha-1}\mathbf{J} + o(1).\end{aligned}$$

The first right-hand side has asymptotically no residue when we integrate over the contour Γ_ρ (as per Proposition 38 there is no eigenvalues of $\bar{\mathbf{X}}$ in Γ_ρ for all large n almost surely). We are then left with the second right-most term. Using the block structure used in Section A.3, we may write

$$\left(\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}\right)^{-1} \xrightarrow{\text{a.s.}}$$
$$\begin{pmatrix}\mathbf{I}_K + e_{21}^\alpha(z)(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^\mathsf{T}) - \left(1 + \frac{z}{m_\mu}e_{10}^\alpha(z)\right)\mathbf{c}\mathbf{1}_K^\mathsf{T} & -e_{21}^\alpha(z)\mathbf{c}\\ \frac{z}{m_\mu^2}\left(v_\mu + ze_{0;-1}^\alpha(z)\right)\mathbf{1}_K^\mathsf{T} & -z\frac{e_{10}^\alpha(z)}{m_\mu}\end{pmatrix}^{-1}.$$

Let us write $\gamma(z) = \frac{z}{m_\mu^2}\left(v_\mu + ze_{0;-1}^\alpha(z)\right)$. We can now use a block inversion formula to write

$$\left(\mathbf{I}_{K+1} + \mathbf{V}^T\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}\right)^{-1} \xrightarrow{\text{a.s.}} \begin{pmatrix}(\underline{\mathbf{G}}_z^\alpha)^{-1} & -\frac{e_{10}^\alpha(z)\left[\underline{\mathbf{G}}_z^\alpha - \frac{\gamma(z)m_\mu e_{21}^\alpha(z)}{ze_{10}^\alpha(z)}\mathbf{c}\mathbf{1}_K^\mathsf{T}\right]^{-1}\mathbf{c}}{-\frac{ze_{21}^\alpha(z)}{m_\mu} + \gamma(z)e_{10}^\alpha(z)\mathbf{1}_K^\mathsf{T}\left[\underline{\mathbf{G}}_z^\alpha - \frac{\gamma(z)m_\mu e_{21}^\alpha(z)}{ze_{10}^\alpha(z)}\mathbf{c}\mathbf{1}_K^\mathsf{T}\right]^{-1}\mathbf{c}}\\ \frac{\gamma(z)m_\mu}{ze_{21}^\alpha(z)}\mathbf{1}_K^\mathsf{T}(\underline{\mathbf{G}}_z^\alpha)^{-1} & \frac{1}{-\frac{ze_{21}^\alpha(z)}{m_\mu} + \gamma(z)e_{10}^\alpha(z)\mathbf{1}_K^\mathsf{T}\left[\underline{\mathbf{G}}_z^\alpha - \frac{\gamma(z)m_\mu e_{21}^\alpha(z)}{ze_{10}^\alpha(z)}\mathbf{c}\mathbf{1}_K^\mathsf{T}\right]^{-1}\mathbf{c}}\end{pmatrix} \tag{A.23}$$

with $\underline{\mathbf{G}}_z^\alpha = \mathbf{I}_K + e_{21}^\alpha(z)\left(\mathcal{D}\left(\mathbf{c}\right) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}\left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right) + \theta^\alpha(z)\mathbf{c}\mathbf{1}_K^T$. The entries of the previous matrix seem to be cumbersome but as we will see, the residue calculus will greatly simplify. In fact, we have that $\mathbf{1}_K^T\underline{\mathbf{G}}_z^\alpha = (1 + \theta^\alpha(z))\,\mathbf{1}_K^T$ so that $\mathbf{1}_K^T(\underline{\mathbf{G}}_z^\alpha)^{-1} = \frac{1}{1+\theta^\alpha(z)}\mathbf{1}_K^T$ which is well defined since we are considering the case $1 + \theta^\alpha(z) \neq 0$. Similarly, we have that

$$\left[\underline{\mathbf{G}}_z^\alpha - \frac{\gamma(z)m_\mu e_{10}^\alpha(z)}{ze_{21}^\alpha(z)}\mathbf{c}\mathbf{1}_K^\mathsf{T}\right]\mathbf{c} = \left(-z\frac{e_{10}^\alpha(z)}{m_\mu}\right)\mathbf{c}$$

meaning that $\left[\underline{\mathbf{G}}_z^\alpha - \frac{\gamma(z)m_\mu e_{21}^\alpha(z)}{z e_{10}^\alpha(z)}\mathbf{c}\mathbf{1}_K^\mathsf{T}\right]^{-1}\mathbf{c} = -\frac{m_\mu}{z e_{10}^\alpha(z)}\mathbf{c}$. So finally, the terms $(1,2)$, $(2,1)$ and $(2,2)$ of $\left(\mathbf{I}_{K+1} + \mathbf{V}^T\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}\right)^{-1}$ do no longer depend on $(\underline{\mathbf{G}}_z^\alpha)^{-1}$ and thus do not have poles in the contour Γ_ρ. We can then write

$$\left(\mathbf{I}_{K+1} + \mathbf{V}^T\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}\right)^{-1} = \begin{pmatrix} (\underline{\mathbf{G}}_z^\alpha)^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \mathbf{R}_1(z)$$

with $\mathbf{R}_1(z)$ having no residue in the contour Γ_ρ. Thus, to perform the contour integration of the integrand in (A.22) around Γ_ρ, we just need to evaluate the top-left entries of $\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}$ and $\mathbf{V}^\mathsf{T}\mathbf{Q}_z^\alpha\mathbf{D}^{\alpha-1}\mathbf{J}$. Those are easily retrieved from the calculus in Section A.3.

We have in particular $(\frac{1}{\sqrt{n}}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega})_{11} \xrightarrow{\text{a.s.}} e_{00}^\alpha(z)(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^\mathsf{T}) - \beta^\alpha(z)\mathbf{c}\mathbf{1}_K^\mathsf{T}$ where $\beta^\alpha(z) = \frac{1}{m_\mu}\left[\int t^{2\alpha-1}\mu(\mathrm{d}t) + e_{-1;-1}^\alpha(z)\right]$ and similarly $(\mathbf{V}^\mathsf{T}\mathbf{Q}_z^\alpha\mathbf{D}^{\alpha-1}\mathbf{J})_{11} \xrightarrow{\text{a.s.}} e_{00}^\alpha(z)\mathcal{D}(\mathbf{c})$, so that finally

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J} \xrightarrow{\text{a.s.}}$$
$$-\frac{1}{2\pi i}\oint_{\Gamma_\rho}\left[\left(e_{00}^\alpha(z)(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^\mathsf{T}) - \beta^\alpha(z)\mathbf{c}\mathbf{1}_K^\mathsf{T}\right)(\underline{\mathbf{G}}_z^\alpha)^{-1} \times e_{00}^\alpha(z)\mathcal{D}(\mathbf{c}) + \mathbf{R}_2(z)\mathrm{d}z\right]$$

where $\mathbf{R}_2(z)$ is a matrix having no residue in the considered contour. Now, we are ready to compute the integral. From the Cauchy integral formula,

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J} \xrightarrow{\text{a.s.}}$$
$$\lim_{z\to\rho}(z-\rho)\left[e_{00}^\alpha(z)(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^\mathsf{T}) - \beta^\alpha(z)\mathbf{c}\mathbf{1}_K^\mathsf{T}\right](\underline{\mathbf{G}}_z^\alpha)^{-1} \times e_{00}^\alpha(z)\mathcal{D}(\mathbf{c}).$$

By writing $\underline{\mathbf{G}}_z^\alpha = \rho_z\mathbf{v}_{r,z}\mathbf{v}_{l,z}^T + \tilde{\mathbf{V}}_{r,z}\tilde{\mathbf{\Sigma}}_z\tilde{\mathbf{v}}_{l,z}^T$ where $\mathbf{v}_{r,z}$ and $\mathbf{v}_{l,z}$ are respectively right and left eigenvectors associated with the vanishing eigenvalue ρ_z of $\underline{\mathbf{G}}_z^\alpha$ when $z \to \rho$; $\tilde{\mathbf{V}}_{r,z} \in \mathbb{R}^{n\times\eta_\rho}$ and $\tilde{\mathbf{V}}_{l,z}\mathbb{R}^{n\times\eta_\rho}$ are respectively sets of right and left eigenspaces associated with non vanishing eigenvalues, we then have

$$\lim_{z\to\rho}(z-\rho)(\underline{\mathbf{G}}_z^\alpha)^{-1} \overset{(1)}{=} \lim_{z\to\rho}(z-\rho)\frac{\mathbf{v}_{r,z}\mathbf{v}_{l,z}^T}{\rho_z'}$$

where we have used the l'Hopital rule and the fact that the non vanishing eigenvalue part of $\underline{\mathbf{G}}_z^\alpha$ will produce zero in the limit $z \to \rho$. Using $\rho_z = \mathbf{v}_{l,z}^\mathsf{T}\underline{\mathbf{G}}_z^\alpha\mathbf{v}_{r,z}$, we obtain

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J} \xrightarrow{\text{a.s.}}$$
$$\left[e_{00}^\alpha(\rho)(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^\mathsf{T}) - \beta^\alpha(\rho)\mathbf{c}\mathbf{1}_K^\mathsf{T}\right]\frac{\mathbf{v}_{r,\rho}\mathbf{v}_{l,\rho}^\mathsf{T}}{\left(\mathbf{v}_{l,z}^\mathsf{T}\underline{\mathbf{G}}_z^\alpha\mathbf{v}_{r,z}\right)'_{z=\rho}} \times e_{00}^\alpha(\rho)\mathcal{D}(\mathbf{c}).$$

Since $(\mathbf{v}_{l,\rho})^\mathsf{T}\underline{\mathbf{G}}_\rho^\alpha = \underline{\mathbf{G}}_\rho^\alpha \mathbf{v}_{r,\rho} = 0$,

$$\begin{aligned}\left((\mathbf{v}_{l,z})^T\underline{\mathbf{G}}_z^\alpha\mathbf{v}_{r,z}\right)'_{z=\rho} &= ((\mathbf{v}_{l,z})^T)'_{z=\rho}\underline{\mathbf{G}}_\rho^\alpha\mathbf{v}_{r,\rho} + (\mathbf{v}_{l,\rho})^T\left(\underline{\mathbf{G}}_z^\alpha\right)'_{z=\rho}\mathbf{v}_{r,\rho} + (\mathbf{v}_{l,\rho})^T\underline{\mathbf{G}}_\rho^\alpha(\mathbf{v}_{r,z})'_{z=\rho}\\ &= (\mathbf{v}_{l,\rho})^T\left(\underline{\mathbf{G}}_z^\alpha\right)'_{z=\rho}\mathbf{v}_{r,\rho}\\ &= \left(e_{21}^\alpha(\rho)\right)'(\mathbf{v}_{l,\rho})^T\left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}\left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right)\mathbf{v}_{r,\rho}\end{aligned}$$

where the subscript $'$ denotes the first derivative with respect to z. Using the fact that $\mathbf{v}_{r,\rho}$ is orthogonal to $\mathbf{1}_K^T$, and $(\mathbf{v}_{r,\rho}, \mathbf{v}_{l,\rho})$ is also a pair of eigenvectors of $\left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}\left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right)$ associated with eigenvalue $-\frac{1}{e_{21}^\alpha(\rho)}$, we get

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J} \xrightarrow{\text{a.s.}} \frac{\left(e_{00}^\alpha(\rho)\right)^2}{e_{21}^\alpha(\rho)'}\frac{\mathbf{v}_{r,\rho}(\mathbf{v}_{l,\rho})^T}{\mathbf{v}_{l,\rho}^T\mathbf{v}_{r,\rho}}\mathcal{D}(\mathbf{c})\,. \tag{A.24}$$

By introducing $\mathbf{v}_\rho = \mathcal{D}(\mathbf{c})^{\frac{1}{2}}\mathbf{v}_{l,\rho} = \mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\mathbf{v}_{r,\rho}$ eigenvector of the symmetric matrix $\mathcal{D}(\mathbf{c})^{\frac{1}{2}}\left(\mathbf{I}_K - \mathbf{1}_K\mathbf{c}^\mathsf{T}\right)\mathbf{M}\left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right)\mathcal{D}(\mathbf{c})^{\frac{1}{2}}$, we obtain the final result

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J} \xrightarrow{\text{a.s.}} \frac{\left(e_{00}^\alpha(\rho)\right)^2}{e_{21}^\alpha(\rho)'}\mathcal{D}(\mathbf{c})^{1/2}\mathbf{v}_\rho(\mathbf{v}_\rho)^T\mathcal{D}(\mathbf{c})^{1/2}\,. \tag{A.25}$$

Next, we need to estimate the denominator term $(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha$ of $\frac{1}{n}\frac{\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J}}{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha}$ of ν_i^a. For $\mathbf{u}_i^\alpha$ an eigenvector of $\mathbf{L}_\alpha$ associated to an isolated eigenvalue converging to ρ asymptotically, we have

$$\begin{aligned}(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha &= \mathrm{tr}(\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)})\\ &= \mathrm{tr}\left(-\frac{1}{2\pi i}\oint_{\Gamma_\rho}(\mathbf{L}_\alpha - z\mathbf{I}_n)\,\mathbf{D}^{2(\alpha-1)}\mathrm{d}z\right).\end{aligned}$$

As in the previous section, by applying Woodburry idendity, this is equivalent to evaluating

$$\mathrm{tr}\left(-\frac{1}{2\pi i}\oint_{\Gamma_\rho}\left[\mathbf{V}^\mathsf{T}\mathbf{Q}_z^\alpha\mathbf{D}^{2(\alpha-1)}\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}\begin{pmatrix}(\underline{\mathbf{G}}_z^\alpha)^{-1} & 0\\ 0 & 0\end{pmatrix} + \mathbf{R}_3(z)\right]\mathrm{d}z\right),$$

where $\mathbf{R}_3(z)$ is a matrix having no residue in the considered contour.

Again here, we just need the top left entry of $\mathbf{V}^\mathsf{T}\mathbf{Q}_z^\alpha\mathbf{D}^{2(\alpha-1)}\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega}$ which is given from Theorem 15 by

$$(\mathbf{V}^\mathsf{T}\mathbf{Q}_z^\alpha\mathbf{D}^{2(\alpha-1)}\mathbf{Q}_z^\alpha\mathbf{V}\mathbf{\Omega})_{11} = \underbrace{\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{1-\alpha}\mathbf{Q}_z^\alpha\mathbf{D}^{2(\alpha-1)}\mathbf{Q}_z^\alpha\mathbf{D}^{1-\alpha}\mathbf{J}(\mathbf{I}_K - \mathbf{1}_K\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^\mathsf{T})}_{(\mathbf{I})} \tag{A.26}$$

$$- \underbrace{\frac{1}{\sqrt{n}(\mathbf{q}^\mathsf{T}\mathbf{1}_n)}\mathbf{D}^{1-\alpha}\mathbf{Q}_z^\alpha\mathbf{D}^{2(\alpha-1)}\mathbf{Q}_z^\alpha\mathbf{D}^{-\alpha}\mathbf{X}\mathbf{1}_n\mathbf{1}_K^\mathsf{T}}_{(\mathbf{II})}\,. \tag{A.27}$$

We can get rid of the term (**II**) since after residue calculus, we will get (similar to Equation A.24) $\mathbf{1}_K^\mathsf{T}\mathbf{v}_{r,\rho} = 0$ which cancels out the whole term. Let us now concentrate on the term (**I**). At this point, we need to introduce the following result which, for any deterministic vectors of bounded Euclidean norm $\mathbf{a}$, $\mathbf{b}$ and any deterministic diagonal matrix $\mathbf{\Xi}$, approximates the random quantity $\mathbf{a}^\mathsf{T}\mathbf{Q}_{z_1}^\alpha\mathbf{\Xi}\mathbf{Q}_{z_2}^\alpha\mathbf{b}$ by a deterministic equivalent.

Lemma 41 (Second deterministic equivalents)**.** *For all $z \in \mathbb{C}\setminus\mathcal{S}^\alpha$, we have the following deterministic equivalent*

$$\mathbf{Q}_{z_1}^\alpha\mathbf{\Xi}\mathbf{Q}_{z_2}^\alpha \leftrightarrow \bar{\mathbf{Q}}_{z_1}^\alpha\mathbf{\Xi}\bar{\mathbf{Q}}_{z_2}^\alpha + \bar{\mathbf{Q}}_{z_1}^\alpha\mathcal{D}\left[\left(\mathbf{I}_n - \mathbf{\Upsilon}_{z_1,z_2}\right)^{-1}\mathbf{\Upsilon}_{z_1,z_2}\operatorname{diag}\left(\mathbf{\Xi}\right)\right]\bar{\mathbf{Q}}_{z_2}^\alpha$$

where $\mathbf{\Xi}$ is any diagonal matrix, $\bar{\mathbf{Q}}_z^\alpha$ is given in Lemma 36 and

$$\mathbf{\Upsilon}_{z_1,z_2}(i,j) = \frac{1}{n}\frac{q_i^{1-2\alpha}q_j^{1-2\alpha}\left(1-q_iq_j\right)}{\left(-z_1 - e_{11}^\alpha(z_1)q_i^{1-2\alpha} + e_{21}^\alpha(z_1)q_i^{2-2\alpha}\right)\left(-z_2 - e_{11}^\alpha(z_2)q_j^{1-2\alpha} + e_{21}^\alpha(z_2)q_j^{2-2\alpha}\right)}.$$

The equivalence relation $\leftrightarrow$ is as defined in Lemma 36.

Thanks to Lemma 41 (proof provided in Appendix A.8), a deterministic approximation of the term (**I**) in Equation (A.26) can be obtained. We get in particular

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{1-\alpha}\mathbf{Q}_z^\alpha\mathbf{D}^{2(\alpha-1)}\mathbf{Q}_z^\alpha\mathbf{D}^{1-\alpha}\mathbf{J} = \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{1-\alpha}\bar{\mathbf{Q}}_z^\alpha\mathbf{D}^{2(\alpha-1)}\bar{\mathbf{Q}}_z^\alpha\mathbf{D}^{1-\alpha}\mathbf{J} \tag{A.28}$$

$$+ \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{1-\alpha}\bar{\mathbf{Q}}_z^\alpha\mathcal{D}\left[\left(\mathbf{I}_n - \mathbf{\Upsilon}_{z,z}\right)^{-1}\mathbf{\Upsilon}_{z,z}\mathbf{d}^\alpha\right]\bar{\mathbf{Q}}_z^\alpha\mathbf{D}^{1-\alpha}\mathbf{J} \tag{A.29}$$

where $\mathbf{d}^\alpha = \{q_i^{2(\alpha-1)}\}_{i=1}^n$ and $\mathbf{\Upsilon}_{z_1,z_2}$ was defined in Lemma 41. Using similar argument as in Equation (A.17), we can easily show that the first right hand side term of (A.28) converges almost surely to $e_{00;2}^\alpha\mathcal{D}(\mathbf{c})$. It then remains to estimate the second right-most term of (A.28). $\mathbf{\Upsilon}_{z_1,z_2}$ (as defined in Lemma 41) may be written as the sum of two rank-one matrices

$$\mathbf{\Upsilon}_{z_1,z_2} = \frac{1}{n}\left(\mathbf{a}_{z_1}\mathbf{a}_{z_2}^\mathsf{T} - \mathbf{b}_{z_1}\mathbf{b}_{z_2}^\mathsf{T}\right)$$

where $\mathbf{a}_z = \left\{\frac{q_j^{1-2\alpha}}{-z-q_j^{1-2\alpha}e_{11}^\alpha(z)+q_j^{2-2\alpha}e_{21}^\alpha(z)}\right\}_{j=1}^n$ and $\mathbf{b}_z = \left\{\frac{q_j^{2-2\alpha}}{-z-q_j^{1-2\alpha}e_{11}^\alpha(z)+q_j^{2-2\alpha}e_{21}^\alpha(z)}\right\}_{j=1}^n$.

The matrix $\mathbf{\Upsilon}_{z,z}$ can thus be further written $\mathbf{\Upsilon}_{z,z} = \frac{1}{n}\begin{pmatrix}\mathbf{a}_z & \mathbf{b}_z\end{pmatrix}\mathbf{I}_2\begin{pmatrix}\mathbf{a}_z^\mathsf{T}/n \\ -\mathbf{b}_z^\mathsf{T}/n\end{pmatrix}$. Using matrix inversion lemmas, we have

$$\left(\mathbf{I}_n - \mathbf{\Upsilon}_{z,z}\right)^{-1}\mathbf{\Upsilon}_{z,z}\mathbf{d}^\alpha = \begin{pmatrix}\mathbf{a}_z & \mathbf{b}_z\end{pmatrix}\begin{pmatrix}1-\frac{\mathbf{a}_z^\mathsf{T}\mathbf{a}_z}{n} & -\frac{\mathbf{a}_z^\mathsf{T}\mathbf{b}_z}{n} \\ \frac{\mathbf{b}_z^\mathsf{T}\mathbf{a}_z}{n} & 1+\frac{\mathbf{b}_z^\mathsf{T}\mathbf{b}_z}{n}\end{pmatrix}^{-1}\begin{pmatrix}\frac{\mathbf{a}_z^\mathsf{T}\mathbf{d}^\alpha}{n} \\ -\frac{\mathbf{b}_z^\mathsf{T}\mathbf{d}^\alpha}{n}\end{pmatrix}.$$

Using again the argument in Equation (A.17) , we can easily show that $\frac{\mathbf{a}_z^\mathsf{T}\mathbf{a}_z}{n}$, $\frac{\mathbf{a}_z^\mathsf{T}\mathbf{b}_z}{n}$, $\frac{\mathbf{b}_z^\mathsf{T}\mathbf{b}_z}{n}$, $\frac{\mathbf{a}_z^\mathsf{T}\mathbf{d}^\alpha}{n}$ and $\frac{\mathbf{b}_z^\mathsf{T}\mathbf{d}^\alpha}{n}$ converge for large n almost surely respectively to $e^\alpha_{22;2}(z)$, $e^\alpha_{32;2}(z)$, $e^\alpha_{42;2}(z)$, $e^\alpha_{-1;0}(z)$ and $e^\alpha_{00}(z)$ with $e^\alpha_{ij;2}$ defined in Equation (A.15). This given, we can show that

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{1-\alpha}\bar{\mathbf{Q}}_z^\alpha\mathcal{D}\left[\left(\mathbf{I}_n-\boldsymbol{\Upsilon}_{z,z}\right)^{-1}\boldsymbol{\Upsilon}_{z,z}\mathbf{d}^\alpha\right]\bar{\mathbf{Q}}_z^\alpha\mathbf{D}^{1-\alpha}\mathbf{J}\xrightarrow{\text{a.s.}}\chi^\alpha(z)\mathcal{D}(\mathbf{c})$$

with

$$\chi^\alpha(z)=\frac{\left[\left(1+e^\alpha_{42;2}(z)\right)e^\alpha_{-1,0}(z)-e^\alpha_{32;2}(z)e^\alpha_{00}(z)\right]e^\alpha_{32;3}(z)-\left[e^\alpha_{22;2}(z)e^\alpha_{-1,0}(z)+\left(1-e^\alpha_{22;2}(z)\right)e^\alpha_{00}(z)\right]e^\alpha_{42;3}(z)}{\left(1+e^\alpha_{42;2}(z)\right)\left(1-e^\alpha_{22;2}(z)\right)+\left[e^\alpha_{32;2}(z)\right]^2}$$

We thus have

$$(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha\xrightarrow{\text{a.s.}}\operatorname{tr}\left(\lim_{z\to\rho}\left(e^\alpha_{00;2}(z)+\chi^\alpha(z)\right)(\mathcal{D}(\mathbf{c})-\mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K-\mathbf{c}\mathbf{1}_K^\mathsf{T})(\underline{\mathbf{G}}_z^\alpha)^{-1}\right).$$

By applying l'Hopital rule to evaluate this limit as in the previous section, we obtain

$$(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha\xrightarrow{\text{a.s.}}\frac{e_{00;2}(\rho)+\chi^\alpha(\rho)}{\left(e^\alpha_{21}(\rho)\right)'}.$$

Finally,

$$\frac{1}{n}\frac{\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-\mathbf{1}}\mathbf{J}}{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha}\xrightarrow{\text{a.s.}}\frac{\left(e^\alpha_{00}(\rho)\right)^2}{e_{00;2}(\rho)+\chi^\alpha(\rho)}\mathcal{D}\left(\mathbf{c}\right)^{1/2}\mathbf{v}_\rho(\mathbf{v}_\rho)^T\mathcal{D}\left(\mathbf{c}\right)^{1/2}. \tag{A.30}$$

We recall that one goal of this section is to estimate $\nu_i^a=\frac{1}{\sqrt{n_a}}\frac{\mathbf{u}_i^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{j}_a}{\sqrt{\mathbf{u}_i^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i}}$, the square of which is $\frac{n}{n_a}\left[\frac{1}{n}\frac{\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}}{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha}\right]_{aa}$. From Equation (A.30), the former quantity is easily retrieved and we have

$$|\nu_i^a|^2=\frac{(e^\alpha_{00}(\rho_i))^2}{e^\alpha_{00;2}(\rho_i)+\chi^\alpha(\rho_i)}\left|v_a^i\right|^2. \tag{A.31}$$

This proves the following Theorem giving the limit of the empirical class means ν_i^a's.

Theorem 42 (Means)**.** *For each eigenpair $(\lambda(\bar{\mathbf{M}}),\mathbf{v})$ of $\mathcal{D}(\mathbf{c})^{\frac{1}{2}}\left(\mathbf{I}_K-\mathbf{1}_K\mathbf{c}^\mathsf{T}\right)\mathbf{M}\left(\mathbf{I}_K-\mathbf{c}\mathbf{1}_K^T\right)\mathcal{D}(\mathbf{c})^{\frac{1}{2}}$ of unit multiplicity, mapped to eigenpair $(\rho,\mathbf{u}_i^\alpha)$ of $\mathbf{L}_\alpha$ as defined in Corollary 19, under the conditions of Assumption 1 and for ν_i^a defined in* (A.20)*, we have almost surely as $n\to\infty$, $|(\nu_i^a)^2-(\nu_i^{a,\infty})^2|\to 0$ where*

$$(\nu_i^{a,\infty})^2\equiv\frac{\left[e^\alpha_{00}(\rho)\right]^2}{e^\alpha_{00;2}(\rho,\rho)+\chi^\alpha(\rho)}(v_a)^2$$

with
$\chi^\alpha(\rho)=\frac{\left[\left(1+e^\alpha_{42;2}(\rho)\right)e^\alpha_{-1,0}(\rho)-e^\alpha_{32;2}(\rho)e^\alpha_{00}(\rho)\right]e^\alpha_{32;3}(\rho)-\left[e^\alpha_{22;2}(\rho)e^\alpha_{-1,0}(\rho)+\left(1-e^\alpha_{22;2}(\rho)\right)e^\alpha_{00}(\rho)\right]e^\alpha_{42;3}(\rho)}{\left(1+e^\alpha_{42;2}(\rho)\right)\left(1-e^\alpha_{22;2}(\rho)\right)+\left[e^\alpha_{32;2}(\rho)\right]^2}$ *and v_a is the component a of $\mathbf{v}$.*

Using the definition of ν_a^i in (A.20) and of $\bar{\mathbf{v}}$, $\boldsymbol{\Pi}$ in Theorem 23, Theorem 23 unfolds easily since $\bar{\mathbf{v}}^\mathsf{T}\boldsymbol{\Pi}\bar{\mathbf{v}}=\sum_{a=1}^K(\nu_a^i)^2=\frac{\left[e^\alpha_{00}(\rho)\right]^2}{e^\alpha_{00;2}(\rho,\rho)+\chi^\alpha(\rho)}(v_a)^2$.

A.4.2 Evaluation of the class covariances σ_{ij}^a's

We have shown at the beginning of this section that to estimate the σ_{ij}^a's, we need to evaluate the more involved object

$$\frac{1}{n}\frac{\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathcal{D}_a\mathbf{D}^{\alpha-1}\mathbf{u}_j^\alpha(\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J}}{((\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha)\left((\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_j^\alpha\right)}.$$

Similarly to what was done previously for the estimation of $\frac{1}{n}\frac{\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-\mathbf{1}}\mathbf{J}}{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\mathbf{2}(\alpha-\mathbf{1})}\mathbf{u}_i^\alpha}$, we need here to evaluate

$$\left(\frac{1}{2\pi i}\right)^2\oint_{\Gamma_{\rho 1}}\oint_{\Gamma_{\rho 2}}\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\left(\mathbf{L}_\alpha-z_1\mathbf{I}_n\right)^{-1}\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\left(\mathbf{L}_\alpha-z_2\mathbf{I}_n\right)^{-1}\mathbf{D}^{\alpha-1}\mathbf{J}\mathrm{d}z_1\mathrm{d}z_2$$

where Γ_{ρ_1} and Γ_{ρ_2} are two positively oriented contours circling around some limiting isolated eigenvalues ρ_1 and ρ_2 respectively. We will use the same technique as in the proof of Theorem 42 to evaluate this integrand. Namely, by applying the Woodburry identity to each of the inverse in the integrand, we get

$$\begin{aligned}\left(\frac{1}{2\pi i}\right)^2\oint_{\Gamma_{\rho_1}}\oint_{\Gamma_{\rho_2}}&\frac{1}{n}\mathbf{J}^T\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_1}^\alpha\mathbf{V}\mathbf{\Omega}\left(\mathbf{I}_{K+1}+\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{V}\mathbf{\Omega}\right)^{-1}\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_2}^\alpha\mathbf{V}\\&\times\mathbf{\Omega}\left(\mathbf{I}_{K+1}+\mathbf{V}^T\mathbf{Q}_{z_2}^\alpha\mathbf{V}\mathbf{\Omega}\right)^{-1}\mathbf{V}^T\mathbf{Q}_{z_2}^\alpha\mathbf{D}^{\alpha-1}\mathbf{J}\mathrm{d}z_1\mathrm{d}z_2\end{aligned}$$

where we have used the fact that the cross-terms $\frac{1}{n}\mathbf{J}^T\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_i}^\alpha\mathbf{D}^{\alpha-1}\mathbf{J}$, $i=1,2$ will vanish asymptotically as the latter do not have poles in the considered contours.
By using the identity $\mathbf{\Omega}\left(\mathbf{I}_{K+1}+\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{V}\mathbf{\Omega}\right)^{-1}\mathbf{V}^T=\left(\mathbf{I}_{K+1}+\mathbf{\Omega}\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{V}\right)^{-1}\mathbf{\Omega}\mathbf{V}^T$, the previous integral writes

$$\begin{aligned}\left(\frac{1}{2\pi i}\right)^2\oint_{\Gamma_{\rho_1}}\oint_{\Gamma_{\rho_2}}&\frac{1}{n}\mathbf{J}^T\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_1}^\alpha\mathbf{V}\mathbf{\Omega}\left(\mathbf{I}_{K+1}+\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{V}\mathbf{\Omega}\right)^{-1}\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_2}^\alpha\mathbf{V}\\&\times\left(\mathbf{I}_{K+1}+\mathbf{\Omega}\mathbf{V}^T\mathbf{Q}_{z_2}^\alpha\mathbf{V}\right)^{-1}\mathbf{\Omega}\mathbf{V}^T\mathbf{Q}_{z_2}^\alpha\mathbf{D}^{\alpha-1}\mathbf{J}\mathrm{d}z_1\mathrm{d}z_2\end{aligned}$$

Most of those quantities have been evaluated in the evaluation of the ν_i^a's. We thus obtain

$$\begin{aligned}\left(\frac{1}{2\pi i}\right)^2\oint_{\Gamma_{\rho_1}}\oint_{\Gamma_{\rho_2}}&\left[\frac{1}{n}\mathbf{J}^T\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_1}^\alpha\mathbf{V}\mathbf{\Omega}\begin{pmatrix}(\underline{\mathbf{G}}_{z_1}^\alpha)^{-1}&0\\0&0\end{pmatrix}\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_2}^\alpha\mathbf{V}\right.\\&\left.\times\begin{pmatrix}\left((\underline{\mathbf{G}}_{z_2}^\alpha)^{-1}\right)^T&0\\0&0\end{pmatrix}\mathbf{\Omega}\mathbf{V}^T\mathbf{Q}_{z_2}^\alpha\mathbf{D}^{\alpha-1}\mathbf{J}+\mathbf{R}_4(z_1,z_2)\right]\mathrm{d}z_1\mathrm{d}z_2\end{aligned}$$

where $\mathbf{R}_4(z_1,z_2)$ has no poles in the considered contours. It is then sufficient to evaluate the top left entry of each of the matrices $\mathbf{J}^T\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_1}^\alpha\mathbf{V}\mathbf{\Omega}$, $\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_2}^\alpha\mathbf{V}$ and

$\mathbf{\Omega}\mathbf{V}^T\mathbf{Q}_{z_2}^\alpha\mathbf{D}^{\alpha-1}\mathbf{J}$ to compute the whole integrand. The first and the third of the latter matrices have been evaluated in the proof of Theorem 42. We are then left with the top left entry of $\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_2}^\alpha\mathbf{V}$ which is $\frac{1}{n}\mathbf{J}^T\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_1}^\alpha\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_2}^\alpha\mathbf{D}^{\alpha-1}\mathbf{J}$ from Theorem 15. The former quantity has already been evaluated in the previous section but for (z,z) replaced by (z_1,z_2) and the diagonal matrix between $\mathbf{Q}_{z_1}^\alpha$ and $\mathbf{Q}_{z_2}^\alpha$ being here $\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}$ instead of $\mathbf{D}^{2(\alpha-1)}$. We thus have

$$\left(\mathbf{V}^T\mathbf{Q}_{z_1}^\alpha\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{Q}_{z_2}^\alpha\mathbf{V}\right)_{11} \xrightarrow{\text{a.s.}} c_a\left(e_{00;2}^\alpha(z_1,z_2)\mathcal{D}\left(\boldsymbol{\delta}_{i=a}\right)_{i=1}^K + \chi^\alpha(z_1,z_2)\mathcal{D}(\mathbf{c})\right).$$

Finally, we are left to evaluate

$$\begin{aligned}\left(\frac{1}{2\pi i}\right)^2\oint_{\Gamma_{\rho_1}}\oint_{\Gamma_{\rho_2}}\frac{1}{n}&\left[e_{00}^\alpha(z_1)(\mathcal{D}(\mathbf{c})-\mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}(\mathbf{I}_K-\mathbf{c}\mathbf{1}_K^\mathsf{T})-\beta^\alpha(z_1)\mathbf{c}\mathbf{1}_K^\mathsf{T}\right](\underline{\mathbf{G}}_{z_1}^\alpha)^{-1}\\&\times c_a\left(e_{00;2}^\alpha(z_1,z_2)\mathcal{D}\left(\boldsymbol{\delta}_{i=a}\right)_{i=1}^K+\chi^\alpha(z_1,z_2)\mathcal{D}(\mathbf{c})\right)\\&\times\left((\underline{\mathbf{G}}_{z_2}^\alpha)^{-1}\right)^\mathsf{T}\left[e_{00}^\alpha(z_2)(\mathbf{I}_K-\mathbf{1}_K\mathbf{c}^\mathsf{T})\mathbf{M}(\mathcal{D}(\mathbf{c})-\mathbf{c}\mathbf{c}^\mathsf{T})-\beta^\alpha(z_2)\mathbf{1}_K\mathbf{c}^\mathsf{T}\right]\mathrm{d}z_1\mathrm{d}z_2.\end{aligned}$$

We can then perform a residue calculus similar to what was done in the proof of Theorem 42. Additionnaly, we use the fact that the eigenvectors $\mathbf{v}_{\rho_1}$ and $\mathbf{v}_{\rho_2}$ corresponding to distinct eigenvalues ρ_1 and ρ_2 of the symmetric matrix $\mathcal{D}(\mathbf{c})^{\frac{1}{2}}\left(\mathbf{I}_K-\mathbf{1}_K\mathbf{c}^\mathsf{T}\right)\mathbf{M}\left(\mathbf{I}_K-\mathbf{c}\mathbf{1}_K^T\right)\mathcal{D}(\mathbf{c})^{\frac{1}{2}}$ are orthogonals. All calculus done, we get

$$\begin{aligned}&\left(\frac{1}{n}\frac{\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{u}_j^\alpha(\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J}}{((\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha)\left((\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_j^\alpha\right)}\right)_{ef}\xrightarrow{\text{a.s.}}\\&\frac{e_{00}^\alpha(\rho_i)e_{00}^\alpha(\rho_j)}{\left(e_{00;2}^\alpha(\rho_i,\rho_i)+\chi^\alpha(\rho_i)\right)\left(e_{00;2}^\alpha(\rho_j,\rho_j)+\chi^\alpha(\rho_j)\right)}\\&\times\left[e_{00;2}^\alpha(\rho_i,\rho_j)\sqrt{c_ec_f}v_e^iv_f^jv_a^iv_a^j+\delta_{\rho_i=\rho_j}c_a\chi^\alpha(\rho_i)\sqrt{c_ec_f}v_e^iv_f^i\right].\end{aligned}\tag{A.32}$$

We are thus now ready to evaluate the σ_{ij}^a's. By definition,

$$\sigma_{ij}^a=\left[\frac{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{u}_j^\alpha}{\sqrt{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha}\sqrt{(\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_j^\alpha}}-\frac{1}{n_a}\frac{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{j}_a}{\sqrt{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha}}\frac{(\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{j}_a}{\sqrt{(\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_j^\alpha}}\right].\tag{A.33}$$

The first right hand side term is estimated by dividing $\left(\frac{1}{n}\frac{\mathbf{J}^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{u}_i^\alpha(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{D}_a\mathbf{D}^{\alpha-1}\mathbf{u}_j^\alpha(\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{J}}{\left((\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha\right)\left((\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_j^\alpha\right)}\right)_{ef}$ (Equation A.32) by $\frac{1}{\sqrt{n}}\frac{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{j}_e}{\sqrt{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha}}\neq 0$ and $\frac{1}{\sqrt{n}}\frac{(\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{j}_f}{\sqrt{(\mathbf{u}_j^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_j^\alpha}}\neq 0$ for any couple of indexes (e,f) such that the aforementioned quantities are non zeros. Indeed from the definition of ν_i^a and Equation (A.31), we get

$$\frac{1}{\sqrt{n}}\frac{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{\alpha-1}\mathbf{j}_e}{\sqrt{(\mathbf{u}_i^\alpha)^\mathsf{T}\mathbf{D}^{2(\alpha-1)}\mathbf{u}_i^\alpha}}\xrightarrow{\text{a.s.}}\sqrt{c_e}\frac{e_{00}^\alpha(\rho)}{\sqrt{e_{00;2}^\alpha(\rho,\rho)+\chi^\alpha(\rho)}}\left|v_e^i\right|.\tag{A.34}$$

The covariances σ_{ij}^a's are then found by combining the previous estimates (A.32) and (A.34) as per the Definition (A.33) of the σ_{ij}^a's. This proves the following theorem giving the limit of the empirical class covariances σ_{ij}^a's.

Theorem 43 (Covariances)**.** *For two unit multiplicity eigenpairs* $(\lambda_1(\bar{\mathbf{M}}), \mathbf{v}^1)$ *and* $(\lambda_2(\bar{\mathbf{M}}), \mathbf{v}^2)$ *of* $\mathcal{D}(\mathbf{c})^{\frac{1}{2}} \left(\mathbf{I}_K - \mathbf{1}_K \mathbf{c}^\mathsf{T}\right) \mathbf{M} \left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right) \mathcal{D}(\mathbf{c})^{\frac{1}{2}}$ *mapped respectively to* $(\rho_1, \mathbf{u}_i^\alpha)$ *and* $(\rho_2, \mathbf{u}_j^\alpha)$ *eigenpairs of* $\mathbf{L}_\alpha$ *and for* σ_{ij}^a *defined in* (A.21)*, we have almost surely as* $n \to \infty$, $\left|\sigma_{ij}^a - \sigma_{ij}^{a,\infty}\right| \to 0$ *where*

$$\sigma_{ij}^{a,\infty} \equiv \frac{\left[\left(e_{00;2}^\alpha(\rho_1, \rho_2) - e_{00}^\alpha(\rho_1) e_{00}^\alpha(\rho_2)\right) v_a^{\rho_1} v_a^{\rho_2} + \delta_{\rho_1}^{\rho_2} c_a \chi^\alpha(\rho_1)\right]}{\sqrt{e_{00;2}^\alpha(\rho_1) + \chi^\alpha(\rho_1)} \sqrt{e_{00;2}^\alpha(\rho_2) + \chi^\alpha(\rho_2)}}$$

where $\chi^\alpha(\rho)$ *is defined in Theorem 42.*

From Theorems 42 and 43, $\nu_i^{a,\infty}$ and $\sigma_{ij}^{a,\infty}$ depend on the e_{ij}'s (defined in Theorem 17), the normalized eigenvectors $\mathbf{v}$ of $\mathcal{D}(\mathbf{c})^{\frac{1}{2}} \left(\mathbf{I}_K - \mathbf{1}_K \mathbf{c}^\mathsf{T}\right) \mathbf{M} \left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right) \mathcal{D}(\mathbf{c})^{\frac{1}{2}}$ and the proportions c_a's of classes. Thanks to Lemma 24, the e_{ij}'s can consistently be estimated similarly to what was described in Proposition 25. Namely, the q_i's can be estimated using $\hat{q}_i = \frac{d_i}{\sqrt{\mathbf{d}^\mathsf{T} \mathbf{1}_n}}$ and replaced in Equations (A.14), (A.15), (A.16) to obtain consistent estimates for the e_{ij}'s. However, the eigenvectors $\mathbf{v}$ and the class proportions are not directly accessible in practice. Nevertheless, in the particular case of $K = 2$ classes, we know exactly $\mathbf{v}$.

Remark 44 ($K = 2$ classes)**.** *Here, only one isolated eigenvector is used for the classification. Since* $\mathbf{v}_r$ *(right eigenvector of* $\bar{\mathbf{M}}$*) is orthogonal to* $\mathbf{1}_2$*,* $\mathbf{v}_r$ *is necessarily the vector* $[1, -1]^\mathsf{T}$*. Hence, the normalized eigenvector* $\mathbf{v} = \frac{\mathcal{D}(\mathbf{c})^{-\frac{1}{2}} \mathbf{v}_r}{\|\mathcal{D}(\mathbf{c})^{-\frac{1}{2}} \mathbf{v}_r\|}$ *is* $\frac{1}{\sqrt{1/c_1 + 1/c_2}} \left[\frac{1}{\sqrt{c_1}}, -\frac{1}{\sqrt{c_2}}\right]^\mathsf{T}$.

We thus obtain from Theorems 42 and 43 along with Remark 44,

Corollary 45 (Means and covariances for $K = 2$ classes)**.** *For* $a = 1, 2$

$$(\nu^{a,\infty})^2 = \frac{\left[e_{00}^\alpha(\rho)\right]^2}{\left(e_{00;2}^\alpha(\rho, \rho) + \chi^\alpha(\rho)\right)\left(1 + \frac{c_a}{1 - c_a}\right)}$$

$$(\sigma^{a,\infty})^2 = \frac{\left[\frac{\left(e_{00;2}^\alpha(\rho,\rho) - e_{00}^\alpha(\rho)^2\right)}{\left(1 + \frac{c_a}{1-c_a}\right)} + c_a \chi^\alpha(\rho)\right]}{e_{00;2}^\alpha(\rho, \rho) + \chi^\alpha(\rho)}$$

for ρ *the unique isolated eigenvalue of* $\mathbf{L}_\alpha$ *(if it exists).*

A.5 Non informative eigenvectors

The objective of this section is to show that the eigenvectors $\tilde{\mathbf{u}}^\alpha$ of $\mathbf{L}_\alpha$ associated to the limiting eigenvalue $\tilde{\rho}$ for which $1 + \theta^\alpha(\tilde{\rho}) = 0$ (Remark 40) are not useful for the classification.

Let us write as in Section A.4

$$\tilde{\mathbf{u}}^\alpha = \sum_{a=1}^{K} \tilde{\nu}^a \frac{\mathbf{j}_a}{\sqrt{n_a}} + \sqrt{\tilde{\sigma}^a_{ii}}\mathbf{w}^a \tag{A.35}$$

where $\mathbf{w}^a \in \mathbb{R}^n$ is a random vector orthogonal to $\mathbf{j}_a$ of norm $\sqrt{n_a}$, supported on the indices of $\mathcal{C}_a$ with identically distributed entries. We shall show that $\tilde{\nu}^a$ is independent of class $\mathcal{C}_a$ and thus, any correct classification cannot be done using $\tilde{\mathbf{u}}^\alpha$. From (A.35), $\tilde{\nu}^a = \frac{(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{j}_a}{\sqrt{n_a}}$ which can be retrieved from the diagonal elements of $\frac{1}{n}\mathbf{J}^\mathsf{T}\tilde{\mathbf{u}}^\alpha(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{J}$. We will evaluate this object by using the same technique as in Section A.4. By the residue formula, we have

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\tilde{\mathbf{u}}^\alpha(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{J} = -\frac{1}{2\pi i}\oint_{\Gamma_{\tilde{\rho}}} \frac{1}{n}\mathbf{J}^\mathsf{T}\left(\mathbf{L}_\alpha - z\mathbf{I}_n\right)^{-1}\mathbf{J}\mathrm{d}z \tag{A.36}$$

$$= -\frac{1}{2\pi i}\oint_{\Gamma_{\tilde{\rho}}} \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{Q}^\alpha_z\mathbf{J}\mathrm{d}z + \frac{1}{2\pi i}\oint_{\Gamma_{\tilde{\rho}}} \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{Q}^\alpha_z\mathbf{V}\mathbf{\Omega}\left(\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}^\alpha_z\mathbf{V}\mathbf{\Omega}\right)^{-1}\mathbf{V}^\mathsf{T}\mathbf{Q}^\alpha_z\mathbf{J} \tag{A.37}$$

for large n almost surely, where $\Gamma_{\tilde{\rho}}$ is a complex (positively oriented) contour circling around the limiting eigenvalue $\tilde{\rho}$ only. The first integral $-\frac{1}{2\pi i}\oint_{\Gamma_{\tilde{\rho}}}\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{Q}^\alpha_z\mathbf{J}\mathrm{d}z$ is asymptotically zero since, from Proposition 38, the integrand has no poles in the contour $\Gamma_{\tilde{\rho}}$. We thus obtain similarly as in Section A.4

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\tilde{\mathbf{u}}^\alpha(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{J} = \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{Q}^\alpha_{\tilde{\rho}}\mathbf{V}\mathbf{\Omega}\left[\lim_{z\to\tilde{\rho}}(z-\tilde{\rho})(\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}^\alpha_z\mathbf{V}\mathbf{\Omega})^{-1}\right]\mathbf{V}^\mathsf{T}\mathbf{Q}^\alpha_{\tilde{\rho}}\mathbf{J}. \tag{A.38}$$

From (A.23), the entries $(1,2)$ and $(2,2)$ of $(\mathbf{I}_{K+1} + \mathbf{V}^\mathsf{T}\mathbf{Q}^\alpha_z\mathbf{V}\mathbf{\Omega})^{-1}$ do not contain $(\underline{\mathbf{G}}^\alpha_z)^{-1}$ since

$$\left[\underline{\mathbf{G}}^\alpha_z - \frac{\gamma(z)m_\mu e^\alpha_{21}(z)}{ze^\alpha_{10}(z)}\mathbf{c}\mathbf{1}_K^\mathsf{T}\right]^{-1}\mathbf{c} = -\frac{m_\mu}{ze^\alpha_{10}(z)}\mathbf{c}$$

and thus, the above limit will give zero for those entries. We thus get

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\tilde{\mathbf{u}}^\alpha(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{J} = \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{Q}^\alpha_{\tilde{\rho}}\mathbf{V}\mathbf{\Omega}\left[\lim_{z\to\tilde{\rho}}(z-\tilde{\rho})\begin{pmatrix}(\underline{\mathbf{G}}^\alpha_z)^{-1} & 0\\ \frac{\gamma(z)m_\mu}{ze^\alpha_{21}(z)}\mathbf{1}_K^\mathsf{T}(\underline{\mathbf{G}}^\alpha_z)^{-1} & 0\end{pmatrix}\right]\mathbf{V}^\mathsf{T}\mathbf{Q}^\alpha_{\tilde{\rho}}\mathbf{J}. \tag{A.39}$$

We recall that in the case under study $(1+\theta^\alpha(\tilde{\rho}) = 0)$, $\mathbf{1}_K$ and $\mathbf{c}$ are respectively left and right eigenvectors of $\underline{\mathbf{G}}^\alpha_z$ associated to the vanishing eigenvalue. We can thus write $\underline{\mathbf{G}}^\alpha_z = \rho_z\mathbf{c}\mathbf{1}_K^\mathsf{T} + \tilde{\mathbf{V}}_{r,z}\tilde{\mathbf{\Sigma}}_z\tilde{\mathbf{V}}_{l,z}^T$ where ρ_z is the vanishing eigenvalue when $z\to\tilde{\rho}$ and $\tilde{\mathbf{V}}_{r,z}$ and $\tilde{\mathbf{V}}_{l,z}$ are respectively sets of right and left eigenspaces associated with non vanishing eigenvalues. Hence, we have

$$\lim_{z\to\tilde{\rho}}(z-\tilde{\rho})(\underline{\mathbf{G}}^\alpha_z)^{-1} \overset{(1)}{=} \lim_{z\to\tilde{\rho}}\frac{\mathbf{c}\mathbf{1}_K^\mathsf{T}}{\rho_z'} \overset{(2)}{=} \lim_{z\to\tilde{\rho}}\frac{\mathbf{c}\mathbf{1}_K^\mathsf{T}}{\mathbf{1}_K^\mathsf{T}(\underline{\mathbf{G}}^\alpha_z)'\mathbf{c}} \overset{(3)}{=} \lim_{z\to\tilde{\rho}}\frac{\mathbf{c}\mathbf{1}_K^\mathsf{T}}{(\theta^\alpha(z))'} \overset{(4)}{=} \frac{\mathbf{c}\mathbf{1}_K^\mathsf{T}}{(\theta^\alpha(\tilde{\rho}))'} \tag{A.40}$$

where in (1) we have used the l'Hopital rule, in (2) we used the fact that ρ_z can be written $\rho_z = \mathbf{1}_K^\mathsf{T}\underline{\mathbf{G}}_z^\alpha\mathbf{c}$ and in (3) we have used $(\underline{\mathbf{G}}_{\tilde{\rho}}^\alpha)' = (e_{21}^\alpha(\tilde{\rho}))'\left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}\left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right) + (\theta^\alpha(\tilde{\rho}))'\mathbf{c}\mathbf{1}_K^T$ and $\mathbf{1}_K^\mathsf{T}\mathbf{c} = 1$. We then have

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\tilde{\mathbf{u}}^\alpha(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{J} = \frac{1}{n}(\mathbf{J}^\mathsf{T}\mathbf{Q}_{\tilde{\rho}}^\alpha\mathbf{V}\mathbf{\Omega})_{11}\frac{\mathbf{c}\mathbf{1}_K^\mathsf{T}}{(\theta^\alpha(\tilde{\rho}))'}(\mathbf{V}^\mathsf{T}\mathbf{Q}_{\tilde{\rho}}^\alpha\mathbf{J})_{11} \tag{A.41}$$

$$+ \frac{1}{n}(\mathbf{J}^\mathsf{T}\mathbf{Q}_{\tilde{\rho}}^\alpha\mathbf{V}\mathbf{\Omega})_{12}\frac{\gamma(\tilde{\rho})m_\mu\mathbf{1}_K^\mathsf{T}}{\tilde{\rho}e_{21}^\alpha(\tilde{\rho})(\theta^\alpha(\tilde{\rho}))'}(\mathbf{V}^\mathsf{T}\mathbf{Q}_{\tilde{\rho}}^\alpha\mathbf{J})_{11}. \tag{A.42}$$

All calculus done similarly as in Section A.4, we get

$$\begin{aligned}\frac{1}{n}\mathbf{J}^\mathsf{T}\tilde{\mathbf{u}}^\alpha(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{J} \xrightarrow{\text{a.s.}} \frac{e_{1,\frac{1}{2}}^\alpha(\tilde{\rho})}{(\theta^\alpha(\tilde{\rho}))'}\Big[&e_{1,\frac{1}{2}}^\alpha(\tilde{\rho})\left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}\left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right)\\ &-\frac{1}{m_\mu}\left(\int t^\alpha\mu(\mathrm{d}t) + e_{0,\frac{1}{2}}^\alpha(\tilde{\rho})\right)\mathbf{c}\mathbf{1}_K^\mathsf{T}\Big]\mathbf{c}\mathbf{c}^\mathsf{T}\\ &-(e_{1,\frac{1}{2}}^\alpha(\tilde{\rho}))^2\frac{\gamma(\tilde{\rho})m_\mu}{\tilde{\rho}e_{21}^\alpha(\tilde{\rho})(\theta^\alpha(\tilde{\rho}))'}\mathbf{c}\mathbf{c}^\mathsf{T}.\end{aligned}$$

Finally,

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\tilde{\mathbf{u}}^\alpha(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{J} \xrightarrow{\text{a.s.}} -\frac{e_{1,\frac{1}{2}}^\alpha(\tilde{\rho})}{m_\mu(\theta^\alpha(\tilde{\rho}))'}\left[\int t^\alpha\mu(\mathrm{d}t) + e_{0,\frac{1}{2}}^\alpha(\tilde{\rho}) + \frac{e_{1,\frac{1}{2}}^\alpha(\tilde{\rho})\gamma(\tilde{\rho})m_\mu^2}{\tilde{\rho}e_{21}^\alpha(\tilde{\rho})}\right]\mathbf{c}\mathbf{c}^\mathsf{T}. \tag{A.43}$$

By recalling that $\tilde{\nu}^a = \frac{(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{j}_a}{\sqrt{n_a}} = \sqrt{\frac{1}{n_a}\left[\frac{1}{n}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\mathbf{J}^\mathsf{T}\tilde{\mathbf{u}}^\alpha(\tilde{\mathbf{u}}^\alpha)^\mathsf{T}\mathbf{J}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\right]_{aa}}$, from (A.43) we deduce that

$$\tilde{\nu}^a \xrightarrow{\text{a.s.}} -\frac{e_{1,\frac{1}{2}}^\alpha(\tilde{\rho})}{m_\mu(\theta^\alpha(\tilde{\rho}))'}\left[\int t^\alpha\mu(\mathrm{d}t) + e_{0,\frac{1}{2}}^\alpha(\tilde{\rho}) + \frac{e_{1,\frac{1}{2}}^\alpha(\tilde{\rho})\gamma(\tilde{\rho})m_\mu^2}{\tilde{\rho}e_{21}^\alpha(\tilde{\rho})}\right]$$

which is independent of the class information (class proportions or inter-class affinities). This concludes the proof.

A.6 Proof of Lemma 24

Lemma 46. *Under Assumption 1,*

$$\max_{1\le i\le n}|q_i - \hat{q}_i| \to 0 \tag{A.44}$$

almost surely, where $\hat{q}_i = \frac{d_i}{\sqrt{\mathbf{d}^\mathsf{T}\mathbf{1}_n}}$.

We need to prove that $\sum_{n=1}^{\infty} \mathbb{P}\left(\max_{1\leq i\leq n} |q_i - \hat{q}_i| > \eta\right) < \infty$ for any $\eta > 0$ so that we can conclude from the first Borel Cantelli lemma (Theorem 4.3 in [Billingsley, 1995]) that

$$\mathbb{P}\left(\limsup_n \max_{1\leq i\leq n} |q_i - \hat{q}_i| > \eta\right) = 0$$

from which Lemma 46 unfolds. We have that

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq i\leq n} |q_i - \hat{q}_i| > \eta\right) &\leq \sum_{i=1}^{n} \mathbb{P}(|q_i - \hat{q}_i| > \eta) \\ &\leq \sum_{i=1}^{n} \mathbb{P}(\hat{q}_i - q_i > \eta) + \mathbb{P}(q_i - \hat{q}_i > \eta). \end{aligned} \tag{A.45}$$

Let us treat for instance the term $\mathbb{P}(\hat{q}_i - q_i > \eta)$ in the following. Since $A_{ij} = q_i q_j + q_i q_j \frac{M_{g_i g_j}}{\sqrt{n}} + X_{ij}$ with X_{ij} a zero mean random variable, we have $\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}A_{ij} \to q_i m_\mu$ and $\frac{1}{n^2}\sum_{i,j} \mathbb{E}A_{ij} \to m_\mu^2$ in the limit $n \to \infty$. For $\hat{q}_i = \frac{\sum_{j=1}^{n} A_{ij}}{\sqrt{\sum_{i,j} A_{ij}}}$, we can write

$$\begin{aligned}\hat{q}_i - q_i = &\underbrace{\frac{\frac{1}{n}\sum_{j=1}^{n}(A_{ij} - \mathbb{E}A_{ij})}{\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}}}}_{A} + \underbrace{\frac{\frac{1}{n}\sum_{j=1}^{n} A_{ij}}{\sqrt{\frac{1}{n^2}\sum_{i,j} A_{ij}}} - \frac{\frac{1}{n}\sum_{j=1}^{n} A_{ij}}{\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}}}}_{B} \\ &+ \underbrace{\frac{\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}A_{ij}}{\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}}} - q_i}_{C}\end{aligned}$$

Since A, B and C tend to zero in the limit $n \to \infty$, we will next use the fact that $\mathbb{P}(\hat{q}_i - q_i > \eta) \leq \mathbb{P}(A > \eta/3) + \mathbb{P}(B > \eta/3) + \mathbb{P}(C > \eta/3)$ and show that all those individual probabilities vanish asymptotically. Since the term C is deterministic and tends to zero in the limit $n \to \infty$, we have $\mathbb{P}(C > \eta/3) = 0$ for all large n. Let us then control $\mathbb{P}(A > \eta/3)$ and $\mathbb{P}(B > \eta/3)$. We have

$$\begin{aligned}\mathbb{P}(A > \eta/3) &= \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n}(A_{ij} - \mathbb{E}A_{ij}) > \frac{\eta m_\mu}{3} + o(1)\right) \\ &\leq \exp\left[-\frac{n\eta^2 m_\mu^2}{18\left(\sigma^2 + \eta m_\mu/9\right)} + o(1)\right] \end{aligned} \tag{A.46}$$

with $\sigma^2 = \limsup_n \max_{1\leq i\leq n} q_i(\sum_j q_j) - q_i^2(\sum_j q_j^2)$ and where in the last inequality of (A.46), we have used Bernstein's inequality (Theorem 3 in [Boucheron et al., 2013]) since the A_{ij}'s are independent Bernoulli random variables with variance $\sigma_{ij}^2 = q_i q_j(1 - q_i q_j) + \mathcal{O}(n^{-\frac{1}{2}})$.

For the term B we have

$$\begin{aligned}
\mathbb{P}(B > \eta/3) &= \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n} A_{ij}\frac{\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}} - \sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}}{\sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}} > \frac{\eta m_\mu}{3} + o(1)\right) \\
&\overset{(1)}{\leq} \mathbb{P}\left(\left|\frac{\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}} - \sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}}{\sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}}\right| > \frac{\eta m_\mu}{3} + o(1)\right) \\
&\overset{(2)}{\leq} \mathbb{P}\left(\left|\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}} - \sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}\right| > \frac{\eta(m_\mu + o(1))\sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}}{3}, \frac{1}{n^2}\sum_{i,j}A_{ij} > \psi\right) \\
&+ \mathbb{P}\left(\left|\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}} - \sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}\right| > \frac{\eta m_\mu\sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}}{3} + o(1), \frac{1}{n^2}\sum_{i,j}A_{ij} \leq \psi\right) \\
&\overset{(3)}{\leq} \underbrace{\mathbb{P}\left(\left|\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}} - \sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}\right| > \frac{\eta m_\mu\sqrt{\psi}}{3} + o(1)\right)}_{B_1} + \underbrace{\mathbb{P}\left(\frac{1}{n^2}\sum_{i,j}A_{ij} \leq \psi\right)}_{B_2}
\end{aligned} \tag{A.47}$$

where in the inequality (1) we have used the fact that $n^{-1}\sum_{j=1}^{n} A_{ij} \leq 1$; in the inequality (2) $\psi > 0$ is any constant smaller than m_μ^2 and in the inequality (3) we have used $\sqrt{n^{-2}\sum_{i,j}A_{ij}} > \sqrt{\psi}$ and the fact that the probability of the intersection between two events is always smaller than the probability of one of those events. It then remains to control B_1 and B_2. For B_2 we have

$$\mathbb{P}\left(\frac{1}{n^2}\sum_{i,j}A_{ij} \leq \psi\right) \leq \exp\left[-\frac{n(m_\mu^2 - \psi)^2}{2\left(\sigma^2 + (m_\mu^2 - \psi)/3\right)} + o(1)\right] \tag{A.48}$$

where the inequality follows from Bernstein's inequality with the similar arguments as

previously. Finally for the term B_1 we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}}-\sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}\right|>\frac{\eta m_\mu\sqrt{\psi}}{3}+o(1)\right)\\
&=\mathbb{P}\left(\left|\frac{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}-\frac{1}{n^2}\sum_{i,j}A_{ij}}{\sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}}+\sqrt{\frac{1}{n^2}\sum_{i,j}A_{ij}}}\right|>\frac{\eta m_\mu\sqrt{\psi}}{3}+o(1)\right)\\
&\overset{(1)}{\leq}\mathbb{P}\left(\left|\frac{1}{n^2}\sum_{i,j}\mathbb{E}A_{ij}-\frac{1}{n^2}\sum_{i,j}A_{ij}\right|>\frac{\eta m_\mu\sqrt{\psi}(m_\mu+\sqrt{\psi})}{3}+o(1)\right)+\mathbb{P}\left(\frac{1}{n^2}\sum_{i,j}A_{ij}\leq\psi\right)\\
&\overset{(2)}{\leq}\exp\left[-\frac{n\psi\left[\eta m_\mu(m_\mu+\sqrt{\psi})\right]^2}{18\left(\sigma^2+\eta m_\mu\sqrt{\psi}(m_\mu+\sqrt{\psi})/9\right)}+o(1)\right]+\exp\left[-\frac{n(m_\mu^2-\psi)^2}{2\left(\sigma^2+(m_\mu^2-\psi)/3\right)}+o(1)\right]
\end{aligned}\tag{A.49}$$

where in the inequality (1) of Equation (A.49) we have used the same arguments as in the inequalities $(2)-(3)$ of Equation (A.46) and in the inequality (2) we have used Bernstein's inequality along with Equation (A.48). From Equations (A.46)(A.48)(A.49), we conclude that $\sum_{n=1}^{\infty}\sum_{i=1}^{n}\mathbb{P}(\hat{q}_i-q_i>\eta)<\infty$ since $m_\mu^2-\psi>0$. It follows the same lines to show that $\sum_{n=1}^{\infty}\sum_{i=1}^{n}\mathbb{P}(q_i-\hat{q}_i>\eta)<\infty$ which concludes the proof.

A.7 Proof of Lemma 36 (First deterministic equivalents)

Let $\mathbf{Q}_z^\alpha=\left(\bar{\mathbf{X}}-z\mathbf{I}_n\right)^{-1}$ with $\bar{\mathbf{X}}$ a symmetric random matrix having independent entries $\bar{X}_{ij}$ which are Gaussian random variables with zero mean and variance $\frac{\sigma_{ij}^2}{n}$. For short, we shall denote $\mathbf{Q}_z^\alpha$ by $\mathbf{Q}$. We want to find a deterministic equivalent $\bar{\mathbf{Q}}$ of $\mathbf{Q}$ in the sense that $\frac{1}{n}\operatorname{tr}\mathbf{C}\mathbf{Q}-\frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}\to 0$ and $\mathbf{d}_1^\mathsf{T}(\mathbf{Q}-\bar{\mathbf{Q}})\mathbf{d}_2\to 0$ almost surely, for all deterministic Hermitian matrix $\mathbf{C}$ and deterministic vectors $\mathbf{d}_i$ of bounded norms (spectral norm for matrices and Euclidian norm for vectors). To this end, we will evaluate $\mathbb{E}(\mathbf{Q})$ since using Lemma 35, one can show that $n^{-1}\operatorname{tr}(\mathbf{C}\mathbf{Q})$ and $\mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{b}$ concentrate respectively around $n^{-1}\operatorname{tr}(\mathbf{A}\mathbb{E}Q)$ and $\mathbf{d}_1^\mathsf{T}\mathbb{E}\mathbf{Q}\mathbf{d}_2$ for all bounded norm matrix $\mathbf{C}$ and vectors $\mathbf{d}_1,\mathbf{d}_2$. For the computations, we use standard Gaussian calculus introduced in [Pastur and Shcherbina, 2011]. Using the resolvent identity (for two invertible matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A}^{-1}-\mathbf{B}^{-1}=-\mathbf{A}^{-1}(\mathbf{A}-\mathbf{B})\mathbf{B}^{-1}$), one has

$$\mathbf{Q}=\frac{1}{z}\bar{\mathbf{X}}\mathbf{Q}-\frac{1}{z}\mathbf{I}_n.\tag{A.50}$$

We then first compute $\mathbb{E}(\bar{\mathbf{X}}\mathbf{Q})$. By writing $\bar{X}_{il}=\frac{\sigma_{il}}{\sqrt{n}}Z_{il}$ where Z_{il} is a random variable with zero mean and unit variance, we thus have

$$\mathbb{E}(\bar{\mathbf{X}}\mathbf{Q})_{ij}=\sum_{l=1}^{n}\frac{\sigma_{il}}{\sqrt{n}}\mathbb{E}(Z_{il}Q_{lj}).$$

By applying Stein's Lemma (Lemma 34 in Section A.0.1), we have

$$\begin{aligned}\mathbb{E}(Z_{il}Q_{lj}) &= \mathbb{E}\left(\frac{\partial(\bar{\mathbf{X}}-z\mathbf{I})^{-1}_{lj}}{\partial Z_{il}}\right)\\ &= \mathbb{E}\left(-(\bar{\mathbf{X}}-z\mathbf{I})^{-1}\frac{\partial\bar{\mathbf{X}}}{\partial Z_{il}}(\bar{\mathbf{X}}-z\mathbf{I})^{-1}\right)_{lj}\\ &= \mathbb{E}\left(-(\bar{\mathbf{X}}-z\mathbf{I})^{-1}\frac{\sigma_{il}}{\sqrt{n}}(\mathbf{E}_{il}+\mathbf{E}_{li})(\bar{\mathbf{X}}-z\mathbf{I})^{-1}\right)_{lj}\end{aligned}$$

where $\mathbf{E}_{il}$ is the matrix with all entries equal to 0 but the entry (i,l) which is equal to 1. Using simple algebra, we have

$$\left((\bar{\mathbf{X}}-z\mathbf{I})^{-1}\mathbf{E}_{il}(\bar{\mathbf{X}}-z\mathbf{I})^{-1}\right)_{lj} = (\bar{\mathbf{X}}-z\mathbf{I})^{-1}_{li}(\bar{\mathbf{X}}-z\mathbf{I})^{-1}_{lj}$$

and

$$\left((\bar{\mathbf{X}}-z\mathbf{I})^{-1}\mathbf{E}_{li}(\bar{\mathbf{X}}-z\mathbf{I})^{-1}\right)_{lj} = (\bar{\mathbf{X}}-z\mathbf{I})^{-1}_{ll}(\bar{\mathbf{X}}-z\mathbf{I})^{-1}_{ij}.$$

We thus get

$$\mathbb{E}(\bar{\mathbf{X}}\mathbf{Q})_{ij} = \sum_{l=1}^{n} -\frac{\sigma_{il}^2}{n}\left(\mathbb{E}\left[Q_{li}Q_{lj}\right] + \mathbb{E}\left[Q_{ll}Q_{ij}\right]\right).$$

Going back to (A.50), we thus have

$$\begin{aligned}\mathbb{E}(Q_{ij}) &= -\frac{1}{z}\sum_{l=1}^{n}\frac{\sigma_{il}^2}{n}\mathbb{E}\left[Q_{li}Q_{lj}\right] - \frac{1}{z}\sum_{l=1}^{n}\frac{\sigma_{il}^2}{n}\mathbb{E}\left[Q_{ll}Q_{ij}\right] - \frac{1}{z}\delta_{ij}\\ &= -\frac{1}{z}\mathbb{E}\left[\mathbf{Q}\frac{\mathbf{\Sigma}_i}{n}\mathbf{Q}\right]_{ij} - \frac{1}{z}\mathbb{E}\left[Q_{ij}\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i\mathbf{Q}}{n}\right)\right] - \frac{1}{z}\delta_{ij} \qquad \text{(A.51)}\end{aligned}$$

where $\mathbf{\Sigma}_i = \mathcal{D}\left(\sigma_{ij}^2\right)_{j=1}^n$. Since the goal is to retrieve $\mathbb{E}(Q_{ij})$, the following lemma allows to split $\mathbb{E}\left[Q_{ij}\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i\mathbf{Q}}{n}\right)\right]$ into $\mathbb{E}\left[Q_{ij}\right]$ and $\mathbb{E}\left[\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i\mathbf{Q}}{n}\right)\right]$.

Lemma 47. *For $\mathbf{Q} = (\bar{\mathbf{X}} - z\mathbf{I}_n)^{-1}$ and $\mathbf{\Sigma}_i = \mathcal{D}(\sigma_{ij}^2)_{j=1}^n$, where $\bar{\mathbf{X}}$ is a symmetric random matrix having independent entries (up to the symmetry) of zero mean and variance $\frac{\sigma_{ij}^2}{n}$, we have*

$$\mathbb{E}\left[Q_{ij}\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i\boldsymbol{Q}}{n}\right)\right] = \mathbb{E}\left[Q_{ij}\right]\mathbb{E}\left[\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i\boldsymbol{Q}}{n}\right)\right] + o(1).$$

Proof. For two real random variables x and y, by Cauchy-Shwarz's inequality,

$$\left|\mathbb{E}\left[(x-\mathbb{E}(x))(y-\mathbb{E}(y))\right]\right| \leq \sqrt{\operatorname{Var}(x)}\sqrt{\operatorname{Var}(y)}$$

which, for $x = \operatorname{tr}\left(\frac{\mathbf{\Sigma}_i\mathbf{Q}}{n}\right)$ and $y = Q_{ij} - \mathbb{E}(Q_{ij})$ gives

$$\left|\mathbb{E}\left[Q_{ij}\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i\mathbf{Q}}{n}\right)\right] - \mathbb{E}\left[Q_{ij}\right]\mathbb{E}\left[\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i\mathbf{Q}}{n}\right)\right]\right| \leq \sqrt{\operatorname{Var}(x)}\sqrt{\operatorname{Var}(y)}$$

since $\mathbb{E}(y)$ is equal to 0 in that case. Using Nash Poincaré inequality (Lemma 35 in Section A.0.1), one can show that $\mathrm{Var}(x) = \mathcal{O}\left(\frac{1}{n^2}\right)$[Hachem et al., 2007]. Additionally, $\forall i, j$ and $z \in \mathbb{C}^+$, $|Q_{ij}| \leq \frac{1}{|\Im(z)|}$. This finally implies that $\mathbb{E}\left[Q_{ij} \operatorname{tr}\left(\frac{\mathbf{\Sigma}_i \mathbf{Q}}{n}\right)\right] - \mathbb{E}\left[Q_{ij}\right] \mathbb{E}\left[\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i \mathbf{Q}}{n}\right)\right] = \mathcal{O}(n^{-1})$. □

Since $\Im(-z - \mathbb{E}\operatorname{tr}(\frac{\mathbf{\Sigma}_i \mathbf{Q}}{n})) < -\Im(z)$ for $z \in \mathbb{C}^+$, $-z - \mathbb{E}\operatorname{tr}(\frac{\mathbf{\Sigma}_i \mathbf{Q}}{n})$ does not vanish asymptotically. Going back to $\mathbb{E}(Q_{ij})$ in Equation (A.51), we may then write

$$\mathbb{E}(Q_{ij}) = \frac{\mathbb{E}\left[\mathbf{Q}\frac{\mathbf{\Sigma}_i}{n}\mathbf{Q}\right]_{ij} + \delta_{ij}}{-z - \mathbb{E}\left[\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i \mathbf{Q}}{n}\right)\right]} + \mathcal{O}(n^{-1}). \tag{A.52}$$

Multiplying Equation (A.52) by $\frac{\sigma_{ki}^2}{n}$, taking $j = i$, summing over i and scaling by n, we get

$$\operatorname{tr}\mathbb{E}\left(\frac{\mathbf{\Sigma}_k \mathbf{Q}}{n}\right) = \sum_{i=1}^{n} \left(\frac{\mathbb{E}\left[\frac{\mathbf{\Sigma}_k}{n}\mathbf{Q}\frac{\mathbf{\Sigma}_i}{n}\mathbf{Q}\right]_{ii} + \frac{\sigma_{ki}^2}{n}}{-z - \mathbb{E}\left[\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i \mathbf{Q}}{n}\right)\right]} \right) + \mathcal{O}(n^{-1}).$$

Using a similar approach to the proof of Lemma 47, we can show that $\sum_{i=1}^{n} \mathbb{E}\left[\frac{\mathbf{\Sigma}_k}{n}\mathbf{Q}\frac{\mathbf{\Sigma}_i}{n}\mathbf{Q}\right]_{ii} = \mathcal{O}(n^{-1})$. We thus have

$$\frac{1}{n}\operatorname{tr}\mathbb{E}\left(\mathbf{\Sigma}_k \mathbf{Q}\right) = \sum_{i=1}^{n} \frac{\frac{\sigma_{ki}^2}{n}}{-z - \frac{1}{n}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{\Sigma}_i \mathbf{Q}\right)\right]} + o(1).$$

By using standard techniques [Hachem et al., 2007], one can show that the unique solution $e_i(z)$ to $e_i(z) = \frac{1}{n}\sum_{j=1}^{n} \frac{\sigma_{ij}^2}{-z - e_j(z)}$ is such that $\frac{1}{n}\operatorname{tr}\mathbb{E}\left(\mathbf{\Sigma}_i \mathbf{Q}\right) - e_i(z) \xrightarrow{\text{a.s.}} 0$. Going back to Equation (A.52), we can thus write for large n

$$\mathbb{E}\left[(-z\mathbf{I} - \mathcal{D}(e_i(z)))\mathbf{Q}\right]_{ij} = \mathbb{E}\left[\mathbf{Q}\frac{\mathbf{\Sigma}_i}{n}\mathbf{Q}\right]_{ij} + \delta_{ij} + o(1). \tag{A.53}$$

Let us denote $\mathbf{\Xi} = -z\mathbf{I} - \mathcal{D}(e_i(z))$. Since $-z - \mathbb{E}\operatorname{tr}\left(\frac{\mathbf{\Sigma}_i \mathbf{Q}}{n}\right)$ is away from zero for $z \in \mathbb{C}^+$ so is $-z - e_i(z)$ and thus $\mathbf{\Xi}$ is invertible and bounded. For large n, we can write for a given deterministic matrix $\mathbf{C}$ of bounded norm

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\operatorname{tr}\mathbf{C}\mathbf{Q}\right] &= \frac{1}{n}\sum_{i,j}(\mathbf{C}\mathbf{\Xi}^{-1})_{ji}\mathbb{E}\left(\mathbf{\Xi}\mathbf{Q}\right)_{ij} \\
&\overset{(1)}{=} \frac{1}{n}\sum_{i,j}(\mathbf{C}\mathbf{\Xi}^{-1})_{ji}\left(\mathbb{E}\left[\mathbf{Q}\frac{\mathbf{\Sigma}_i}{n}\mathbf{Q}\right]_{ij} + \delta_{ij}\right) + o(1) \\
&= \frac{1}{n}\operatorname{tr}\mathbb{E}\left(\mathbf{C}\mathbf{\Xi}^{-1}\mathbf{Q}\frac{\mathbf{\Sigma}_i}{n}\mathbf{Q}\right) + \frac{1}{n}\operatorname{tr}(\mathbf{C}\mathbf{\Xi}^{-1}) + o(1)
\end{aligned}$$

where (1) follows from Equation (A.53). We can then prove that $\frac{1}{n}\operatorname{tr}\mathbb{E}\left(\mathbf{C}\mathbf{\Xi}^{-1}\mathbf{Q}\frac{\mathbf{\Sigma}_i}{n}\mathbf{Q}\right) = \mathcal{O}(n^{-1})$ using a similar approach to the proof of Lemma 47. Hence for large n

$$\mathbb{E}\left[\frac{1}{n}\operatorname{tr}\mathbf{C}\mathbf{Q}\right] = \frac{1}{n}\operatorname{tr}(\mathbf{C}\mathbf{\Xi}^{-1}) + o(1). \tag{A.54}$$

Similarly, for any vectors $\mathbf{a}$, $\mathbf{b}$ of bounded norms, we may write

$$\begin{aligned}\mathbb{E}\left[\mathbf{a}^*\mathbf{Q}\mathbf{b}\right] &= \sum_{i,j}(\mathbf{a}^*\boldsymbol{\Xi}^{-1})_i\mathbb{E}\left(\boldsymbol{\Xi}\mathbf{Q}\right)_{ij}\mathbf{b}_j \\ &= \sum_{i,j}(\mathbf{a}^*\boldsymbol{\Xi}^{-1})_i\mathbb{E}\left[\mathbf{Q}\frac{\boldsymbol{\Sigma}_i}{n}\mathbf{Q}\right]_{ij}\mathbf{b}_j + \mathbf{a}^*\boldsymbol{\Xi}^{-1}\mathbf{b} + o(1).\end{aligned}$$

We also have that $\sum_{i,j}(\mathbf{a}^*\boldsymbol{\Xi}^{-1})_i\mathbb{E}\left[\mathbf{Q}\frac{\boldsymbol{\Sigma}_i}{n}\mathbf{Q}\right]_{ij}\mathbf{b}_j = \mathcal{O}(n^{-1})$. This can be proved similarly to the proof of Lemma 47. Hence,

$$\mathbb{E}\left[\mathbf{a}^*\mathbf{Q}\mathbf{b}\right] = \mathbf{a}^*\boldsymbol{\Xi}^{-1}\mathbf{b} + o(1). \tag{A.55}$$

A.8 Proof of Lemma 41 (Second deterministic equivalents)

Our goal is to find a deterministic equivalent to the random quantity $\mathbf{Q}_{z_1}^{\alpha}\boldsymbol{\Xi}\mathbf{Q}_{z_2}^{\alpha}$ for any diagonal deterministic matrix $\boldsymbol{\Xi}$ where we recall that $\mathbf{Q}_{z_1}^{\alpha} = \left(\frac{\bar{\mathbf{X}}}{\sqrt{n}} - z_1\mathbf{I}_n\right)^{-1}$ with $\bar{\mathbf{X}}$ defined previously in Appendix A.7. The proof follows the same techniques as the proof of the first deterministic equivalent $\mathbf{Q}_z^{\alpha}$ in Appendix A.7 but here, the resolvent identity is either applied on $\mathbf{Q}_{z_1}^{\alpha}$ or $\mathbf{Q}_{z_2}^{\alpha}$. The technical details will be omitted as the key techniques have already been developped in Appendix A.7. For the sake of readability, we will denote $\mathbf{Q}_{z_1}^{\alpha} \equiv \mathbf{Q}_1$ and $\mathbf{Q}_{z_2}^{\alpha} \equiv \mathbf{Q}_2$. As in Appendix A.7, we will evaluate $\mathbb{E}(\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)$. By the resolvent identity, we have

$$\begin{aligned}\mathbb{E}(\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)_{ij} &= -\frac{1}{z_1}\mathbb{E}(\boldsymbol{\Xi}\mathbf{Q}_2)_{ij} + \frac{1}{z_1}\mathbb{E}(\mathbf{X}\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)_{ij} \\ &= -\frac{1}{z_1}\Xi_{ii}\mathbb{E}(\mathbf{Q}_2)_{ij} + \frac{1}{z_1}\mathbb{E}\sum_{k,l}X_{ik}(Q_1)_{kl}\Xi_{ll}(Q_2)_{lj}.\end{aligned}$$

We have from Lemma 34, $\mathbb{E}\sum_{k,l}X_{ik}(Q_1)_{kl}\Xi_{ll}(Q_2)_{lj} = \sum_{k,l}\frac{\sigma_{ik}}{\sqrt{n}}\Xi_{ll}\mathbb{E}\frac{\partial[(Q_1)_{kl}(Q_2)_{lj}]}{\partial Z_{ik}}$. By expanding all terms and all calculus done, we obtain

$$\begin{aligned}\mathbb{E}(\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)_{ij} = -\frac{1}{z_1}\mathbb{E}(\boldsymbol{\Xi}\mathbf{Q}_2)_{ij} - \frac{1}{z_1}\sum_{k,l}\frac{\sigma_{ik}^2}{n}\Xi_{ll}\mathbb{E}\Bigg[&\underbrace{(Q_1)_{ki}(Q_1)_{kl}(Q_2)_{lj}}_{(1)} + \underbrace{(Q_1)_{kk}(Q_1)_{il}(Q_2)_{lj}}_{(2)} \\ &+ \underbrace{(Q_1)_{kl}(Q_2)_{li}(Q_2)_{kj}}_{(3)} + \underbrace{(Q_1)_{kl}(Q_2)_{lk}(Q_2)_{ij}}_{(4)}\Bigg].\end{aligned}$$

Asymptotically, the non vanishing terms are (2) and (4) so that

$$\begin{aligned}\mathbb{E}(\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)_{ij} &= -\frac{1}{z_1}\mathbb{E}(\boldsymbol{\Xi}\mathbf{Q}_2)_{ij} - \frac{1}{z_1}\sum_{k,l}\frac{\sigma_{ik}^2}{n}\Xi_{ll}\mathbb{E}\left[(Q_1)_{kk}(Q_1)_{il}(Q_2)_{lj} + (Q_1)_{kl}(Q_2)_{lk}(Q_2)_{ij}\right] \\ &\quad + o(1) \\ &= -\frac{1}{z_1}\mathbb{E}(\boldsymbol{\Xi}\mathbf{Q}_2)_{ij} - \frac{1}{z_1}\frac{1}{n}\mathbb{E}\left[\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_1)(\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)_{ij}\right] - \frac{1}{z_1}\frac{1}{n}\mathbb{E}\left[\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_2\boldsymbol{\Xi}\mathbf{Q}_1)(\mathbf{Q}_2)_{ij}\right] \\ &\quad + o(1). \end{aligned} \tag{A.56}$$

Similarly to what was done in the proof of of Lemma 47, we can show that $\mathbb{E}\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_1)\mathbb{E}(\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)_{ij} = \mathbb{E}\left(\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_1)\right)\mathbb{E}\left((\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)_{ij}\right)+o(1)$. We can then write from (A.56)

$$\begin{aligned}&\mathbb{E}\left(\left(\mathbf{I}_n + \frac{1}{z_1}\mathcal{D}\left(\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_1)\right)_{i=1}^n\right)\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2\right) = \\ &-\frac{1}{z_1}\mathbb{E}\left(\boldsymbol{\Xi} + \mathcal{D}\left(\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_2\boldsymbol{\Xi}\mathbf{Q}_1)\right)_{i=1}^n\right)\mathbf{Q}_2 + o(1). \end{aligned} \tag{A.57}$$

From (A.57) and the result of Lemma 36, this entails

$$\mathbb{E}\left(\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2\right) \longleftrightarrow \bar{\mathbf{Q}}_1\boldsymbol{\Xi}\bar{\mathbf{Q}}_2 + \bar{\mathbf{Q}}_1\mathcal{D}\left(\mathbb{E}\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_2\boldsymbol{\Xi}\mathbf{Q}_1)\right)_{i=1}^n\bar{\mathbf{Q}}_2. \tag{A.58}$$

Every object in (A.58) is known but $\mathbb{E}\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_2\boldsymbol{\Xi}\mathbf{Q}_1)$ which we need to evaluate now. By left-multiplying (A.58) by $\boldsymbol{\Sigma}_i$ and taking the normalized trace, we get

$$\mathbb{E}\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_2\boldsymbol{\Xi}\mathbf{Q}_1) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_i\bar{\mathbf{Q}}_1\boldsymbol{\Xi}\bar{\mathbf{Q}}_2\right) + \mathbb{E}\frac{1}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_i\bar{\mathbf{Q}}_1\mathcal{D}\left(\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_2\boldsymbol{\Xi}\mathbf{Q}_1)\right)_{i=1}^n\bar{\mathbf{Q}}_2\right). \tag{A.59}$$

By denoting $f_i = \frac{1}{n}\mathbb{E}\left(\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_2\boldsymbol{\Xi}\mathbf{Q}_1)\right)$, Equation (A.59) leads to

$$\mathbf{f} = \left\{\frac{1}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_i\bar{\mathbf{Q}}_1\boldsymbol{\Xi}\bar{\mathbf{Q}}_2\right)\right\}_{i=1}^n + \frac{1}{n}\left\{\left(\bar{\mathbf{Q}}_2\boldsymbol{\Sigma}_i\bar{\mathbf{Q}}_1\right)_{jj}\right\}_{i,j=1}^n\mathbf{f}$$

which finally entails

$$\mathbf{f} = \left(\mathbf{I}_n - \frac{1}{n}\left\{\left(\bar{\mathbf{Q}}_2\boldsymbol{\Sigma}_i\bar{\mathbf{Q}}_1\right)_{jj}\right\}_{i,j=1}^n\right)^{-1}\frac{1}{n}\left\{\bar{\mathbf{Q}}_2\boldsymbol{\Sigma}_i\bar{\mathbf{Q}}_1\right\}_{i,j=1}^n\operatorname{diag}(\boldsymbol{\Xi}).$$

To complete the proof of Lemma 41, we need to show that $\operatorname{Var}\left(\frac{1}{n}\operatorname{tr}(\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)\right)$ and $\operatorname{Var}\left(\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_2\boldsymbol{\Xi}\mathbf{Q}_1)\right)$ are asymptotically summable so that by the Borell Cantelli Lemma, $\frac{1}{n}\operatorname{tr}(\mathbf{Q}_1\boldsymbol{\Xi}\mathbf{Q}_2)$ and $\frac{1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_i\mathbf{Q}_2\boldsymbol{\Xi}\mathbf{Q}_1)$ converge respectively almost surely to their expectations. Those follow directly by using Nash Poincare inequality (Lemma 35) similarly to what was done in the proof of Lemma 47.

Appendix B

Supplementary material Chapter 5

B.1 Intermediary result

We show that

$$\log \mathbb{P}(\mathbf{A}^{(\mathbf{l})}) = \mathcal{G}(q) + D_{KL}(q||p).$$

We can write successively

$$\begin{aligned}
\log \mathbb{P}(\mathbf{A}^{(\mathbf{l})}) &= \int_{\boldsymbol{\theta}^{(l)}} \sum_{\mathbf{g}^{(l)}} q(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)}) \mathrm{d}\boldsymbol{\theta}^{(l)} \log \mathbb{P}(\mathbf{A}^{(\mathbf{l})}) \\
&= \int_{\boldsymbol{\theta}^{(l)}} \sum_{\mathbf{g}^{(l)}} q(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)}) \mathrm{d}\boldsymbol{\theta}^{(l)} \log \frac{\mathbb{P}(\mathbf{A}^{(\mathbf{l})}, \mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)})}{\mathbb{P}(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)} | \mathbf{A}^{(\mathbf{l})})} \mathrm{d}\boldsymbol{\theta}^{(l)} \\
&= \int_{\boldsymbol{\theta}^{(l)}} \sum_{\mathbf{g}^{(l)}} q(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)}) \mathrm{d}\boldsymbol{\theta}^{(l)} \log \frac{\mathbb{P}(\mathbf{A}^{(\mathbf{l})}, \mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)})}{q(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)})} \mathrm{d}\boldsymbol{\theta}^{(l)} \\
&- \int_{\boldsymbol{\theta}^{(l)}} \sum_{\mathbf{g}^{(l)}} q(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)}) \mathrm{d}\boldsymbol{\theta}^{(l)} \log \frac{\mathbb{P}(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)} | \mathbf{A}^{(\mathbf{l})})}{q(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)})} \mathrm{d}\boldsymbol{\theta}^{(l)} \\
&= \mathbb{E}_q \log \mathbb{P}(\mathbf{A}^{(\mathbf{l})}, \mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)}) + D_{KL}(q||p) \\
&= \underbrace{\mathbb{E}_q \log \mathbb{P}(\mathbf{A}^{(\mathbf{l})} | \mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)}) + \mathbb{E}_q \log \frac{\mathbb{P}(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)})}{q(\mathbf{g}^{(l)}, \boldsymbol{\theta}^{(l)})}}_{\mathcal{G}(q)} + D_{KL}(q||p)
\end{aligned}$$

B.2 Updates of variational parameters

We give the detailed steps to derive updates for the variational parameters $\tau^{(l)}$ and $\boldsymbol{\mu}^{(l)}$.

Using the likelihood

$$\mathbb{P}(A_{ij}^{(l)}|g_i^{(l)}, g_j^{(l)}, \theta^{(l)}_{g_i^{(l)} g_j^{(l)}}) \propto \exp\left\{T^{(l)}(A_{ij}^{(l)})\eta^{(l)}(\theta^{(l)}_{g_i^{(l)} g_j^{(l)}})\right\}, \tag{B.1}$$

we can write

$$\begin{aligned}
\log \mathbb{P}(\mathbf{A^{(l)}}|\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \boldsymbol{\theta}^{(l)}) &= \sum_{r_1^{(l)}} \sum_{ij} \sum_{(g_i^{(l)}, g_j^{(l)}) = r_1^{(l)}} T^{(l)}(A_{ij}^{(l)})\eta^{(l)}(\theta^{(l)}_{r_1^{(l)}}) \\
&+ \sum_{r_2^{(l)}} \sum_{ij} \sum_{(g_i^{(l)}, g_j^{(l)}) = r_2^{(l)}} T^{(l)}(A_{ij}^{(l)})\eta^{(l)}(\theta^{(l)}_{r_2^{(l)}}) + \sum_{r_3^{(l)}} \sum_{ij} \sum_{(g_i^{(l)}, g_j^{(l)}) = r_3^{(l)}} T^{(l)}(A_{ij}^{(l)})\eta^{(l)}(\theta^{(l)}_{r_3^{(l)}}) \\
&+ \sum_{r_4^{(l)}} \sum_{ij} \sum_{(g_i^{(l)}, g_j^{(l)}) = r_4^{(l)}} T^{(l)}(A_{ij}^{(l)})\eta^{(l)}(\theta^{(l)}_{r_4^{(l)}}).
\end{aligned} \tag{B.2}$$

Using the *dependency cases* stated in Chapter 5, Equation (B.2) (for $l = 1$) can be written

$$\begin{aligned}
\log \mathbb{P}(\mathbf{A^{(1)}}|\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \boldsymbol{\theta}^{(1)}) &= \sum_{r_1^{(1)}} \sum_{ij} \sum_{(g_i^{(2)}, g_j^{(2)}) = r_1^{(1)}} T^{(1)}(A_{ij}^{(1)})\eta^{(1)}(\theta^{(1)}_{r_1^{(1)}}) \\
&+ \sum_{r_2^{(1)}} \sum_{ij} \sum_{(g_i^{(2)}, g_j^{(1)}) = r_2^{(1)}} T^{(1)}(A_{ij}^{(1)})\eta^{(1)}(\theta^{(1)}_{r_2^{(1)}}) \\
&+ \sum_{r_3^{(1)}} \sum_{ij} \sum_{(g_i^{(1)}, g_j^{(2)}) = r_3^{(1)}} T^{(1)}(A_{ij}^{(1)})\eta^{(1)}(\theta^{(1)}_{r_3^{(1)}}) \\
&+ \sum_{r_4^{(1)}} \sum_{ij} \sum_{(g_i^{(1)}, g_j^{(1)}) = r_4^{(1)}} T^{(1)}(A_{ij}^{(1)})\eta^{(1)}(\theta^{(1)}_{r_4^{(1)}})
\end{aligned} \tag{B.3}$$

By similarly using the the *dependency cases* arguments, Equation (B.2) writes for $l = 2$

$$\begin{aligned}
\log \mathbb{P}(\mathbf{A^{(2)}}|\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \boldsymbol{\theta}^{(2)}) &= \sum_{r_1^{(2)}} \sum_{ij} \sum_{(g_i^{(1)}, g_j^{(1)}) = r_1^{(2)}} T^{(2)}(A_{ij}^{(2)})\eta^{(2)}(\theta^{(2)}_{r_1^{(2)}}) \\
&+ \sum_{r_2^{(2)}} \sum_{ij} \sum_{(g_i^{(1)}, g_j^{(2)}) = r_2^{(2)}} T^{(2)}(A_{ij}^{(2)})\eta^{(2)}(\theta^{(2)}_{r_2^{(2)}}) \\
&+ \sum_{r_3^{(2)}} \sum_{ij} \sum_{(g_i^{(2)}, g_j^{(1)}) = r_3^{(2)}} T^{(2)}(A_{ij}^{(2)})\eta^{(2)}(\theta^{(2)}_{r_3^{(2)}}) \\
&+ \sum_{r_4^{(2)}} \sum_{ij} \sum_{(g_i^{(2)}, g_j^{(2)}) = r_4^{(2)}} T^{(2)}(A_{ij}^{(2)})\eta^{(2)}(\theta^{(2)}_{r_4^{(2)}})
\end{aligned} \tag{B.4}$$

The derivation steps to evaluate $\mathcal{G}(q)$ are similar to the ones in [Aicher et al., 2014]. We get

$$\begin{aligned}\mathcal{G}(q) = \sum_{l=1}^{2} \Bigg[&\sum_{r_1^{(l)}} (\bar{T}^{(l)}_{r_1^{(l)}} + \tau_0^{(l)} - \tau_{r_1^{(l)}}) \bar{\eta}^{(l)}_{r_1^{(l)}} + \sum_{r_2^{(l)}} (\bar{T}^{(l)}_{r_2^{(l)}} + \tau_0^{(l)} - \tau^{(l)}_{r_2^{(l)}}) \bar{\eta}^{(l)}_{r_2^{(l)}} \\ &+ \sum_{r_3^{(l)}} (\bar{T}^{(l)}_{r_3^{(l)}} + \tau_0^{(l)} - \tau^{(l)}_{r_3^{(l)}}) \bar{\eta}^{(l)}_{r_3^{(l)}} + \sum_{r_4^{(l)}} (\bar{T}^{(l)}_{r_4^{(l)}} + \tau_0^{(l)} - \tau^{(l)}_{r_4^{(l)}}) \bar{\eta}^{(l)}_{r_4^{(l)}} \\ &+ \sum_i \sum_{g_i^{(l)}} \mu^{(l)}_{i,g_i^{(l)}} \log \frac{(\mu_0^{(l)})_{i,(g_i^{(l)})}}{\mu^{(l)}_{i,g_i^{(l)}}} + \sum_{r^{(l)}} \log \frac{Z^{(l)}(\tau^{(l)}_{r^{(l)}})}{Z^{(l)}(\tau_0^{(l)})} \Bigg] \end{aligned} \tag{B.5}$$

where we define

$$\bar{T}^{(1)}_{r_1^{(1)}} = \sum_{ij} \sum_{(g_i^{(1)}, g_j^{(1)}) = r_1^{(1)}} T^{(1)}(A_{ij}^{(1)}) \mu^{(1)}_{i,g_i^{(1)}} \mu^{(1)}_{j,g_j^{(1)}} \tag{B.6}$$

$$= \sum_{ij} \sum_{(g_i^{(2)}, g_j^{(2)}) = r_1^{(1)}} T^{(1)}(A_{ij}^{(1)}) \mu^{(2)}_{i,g_i^{(2)}} \mu^{(2)}_{j,g_j^{(2)}} \tag{B.7}$$

$$\bar{T}^{(1)}_{r_2^{(1)}} = \sum_{ij} \sum_{(g_i^{(1)}, g_j^{(1)}) = r_2^{(1)}} T^{(1)}(A_{ij}^{(1)}) \mu^{(1)}_{i,g_i^{(1)}} \mu^{(1)}_{j,g_j^{(1)}} \tag{B.8}$$

$$= \sum_{ij} \sum_{(g_i^{(2)}, g_j^{(1)}) = r_2^{(1)}} T^{(1)}(A_{ij}^{(1)}) \mu^{(2)}_{i,g_i^{(2)}} \mu^{(1)}_{j,g_j^{(1)}} \tag{B.9}$$

$$\bar{T}^{(1)}_{r_3^{(1)}} = \sum_{ij} \sum_{(g_i^{(1)}, g_j^{(1)}) = r_3^{(1)}} T^{(1)}(A_{ij}^{(1)}) \mu^{(1)}_{i,g_i^{(1)}} \mu^{(1)}_{j,g_j^{(1)}} \tag{B.10}$$

$$= \sum_{ij} \sum_{(g_i^{(1)}, g_j^{(2)}) = r_3^{(1)}} T^{(1)}(A_{ij}^{(1)}) \mu^{(1)}_{i,g_i^{(1)}} \mu^{(2)}_{j,g_j^{(2)}} \tag{B.11}$$

$$\bar{T}^{(1)}_{r_4^{(1)}} = \sum_{ij} \sum_{(g_i^{(1)}, g_j^{(1)}) = r_4^{(1)}} T^{(1)}(A_{ij}^{(1)}) \mu^{(1)}_{i,g_i^{(1)}} \mu^{(1)}_{j,g_j^{(1)}} \tag{B.12}$$

$$\bar{\eta}^{(1)}_{r^{(1)}} = \frac{\partial \log Z^{(1)}}{\partial \tau^{(1)}}\Big|_{\tau^{(1)} = r^{(1)}} \tag{B.13}$$

and

$$\bar{T}^{(2)}_{r_1^{(2)}} = \sum_{ij} \sum_{(g_i^{(2)},g_j^{(2)})=r_1^{(2)}} T^{(2)}(A^{(2)}_{ij})\mu^{(2)}_{i,g_i^{(2)}}\mu^{(2)}_{j,g_j^{(2)}} \tag{B.14}$$

$$= \sum_{ij} \sum_{(g_i^{(1)},g_j^{(1)})=r_1^{(2)}} T^{(2)}(A^{(2)}_{ij})\mu^{(1)}_{i,g_i^{(1)}}\mu^{(1)}_{j,g_j^{(1)}} \tag{B.15}$$

$$\bar{T}^{(2)}_{r_2^{(2)}} = \sum_{ij} \sum_{(g_i^{(2)},g_j^{(2)})=r_2^{(2)}} T^{(2)}(A^{(2)}_{ij})\mu^{(2)}_{i,g_i^{(2)}}\mu^{(2)}_{j,g_j^{(2)}} \tag{B.16}$$

$$= \sum_{ij} \sum_{(g_i^{(1)},g_j^{(2)})=r_2^{(2)}} T^{(2)}(A^{(2)}_{ij})\mu^{(1)}_{i,g_i^{(1)}}\mu^{(2)}_{j,g_j^{(2)}} \tag{B.17}$$

$$\bar{T}^{(2)}_{r_3^{(2)}} = \sum_{ij} \sum_{(g_i^{(2)},g_j^{(2)})=r_3^{(2)}} T^{(2)}(A^{(2)}_{ij})\mu^{(2)}_{i,g_i^{(2)}}\mu^{(2)}_{j,g_j^{(2)}} \tag{B.18}$$

$$= \sum_{ij} \sum_{(g_i^{(2)},g_j^{(1)})=r_3^{(2)}} T^{(2)}(A^{(2)}_{ij})\mu^{(2)}_{i,g_i^{(2)}}\mu^{(1)}_{j,g_j^{(1)}} \tag{B.19}$$

$$\bar{T}^{(2)}_{r_4^{(2)}} = \sum_{ij} \sum_{(g_i^{(2)},g_j^{(2)})=r_4^{(2)}} T^{(2)}(A^{(2)}_{ij})\mu^{(2)}_{i,g_i^{(2)}}\mu^{(2)}_{j,g_j^{(2)}} \tag{B.20}$$

$$\bar{\eta}^{(2)}_{r^{(2)}} = \frac{\partial \log Z^{(2)}}{\partial \tau^{(2)}}\Big|_{\tau^{(2)}=r^{(2)}} \tag{B.21}$$

Now, in order to find the stationary points of $\mathcal{G}(q)$, we need to differentiate the latter with respect to $\tau^{(1)}$ and $\tau^{(2)}$ respectively. We obtain

$$\frac{\partial \mathcal{G}(q)}{\partial \tau^{(l)}} = \sum_{r^{(l)}} (\bar{T}^{(l)}_{r^{(l)}} + \tau_0^{(l)} - \tau^{(l)}_{r^{(l)}}) \frac{\partial \bar{\eta}^{(l)}_{r^{(l)}}}{\partial \tau^{(l)}_{r^{(l)}}} \tag{B.22}$$

Setting the Equation (B.22) to zero, we get the update equations for $\tau^{(l)}$ as follows

$$\tau^{(l)}_{r^{(l)}} = \tau_0^{(l)} + \bar{T}^{(l)}_{r^{(l)}}, r^{(l)} = 1, \cdots, (K^{(l)})^2 \tag{B.23}$$

To obtain the update equations for the group labels parameters $\mu^{(l)}_{i,k}$, we need to compute $\frac{\partial \mathcal{G}(q)}{\partial \mu^{(l)}_{i,k}}$ and include Lagrange multipliers to enforce the conditions $\sum_k \mu^{(l)}_{i,k} = 1$ for each node i and each layer l such that $\mu^{(l)}_{i,k}$ is a valid probability measure.

We need to be careful here as can be seen in Equations (B.6)-(B.14), $\bar{T}^{(2)}$ depends on $\mu^{(1)}_{i,g_i^{(1)}}$ when $g_i^{(1)} \in \{1, \ldots, K\}$ i.e., $g_i^{(1)}$ is an element in the indexings $r_1^{(1)}, r_2^{(1)}$. Similarly, $\bar{T}^{(1)}$ depends on $\mu^{(2)}_{i,g_i^{(2)}}$ when $g_i^{(2)} \in \{1, \ldots, K\}$ i.e., $g_i^{(2)}$ is an element in the indexings $r_1^{(2)}, r_2^{(2)}$. To this end we shall thus differentiate the cases $g_i^{(1)}, g_i^{(2)} \in [1, K]$ and $g_i^{(1)}, g_i^{(2)} \notin [1, K]$.

Using this fact and after differentiation, we get the following updates
For $k \in \{1, \ldots, K\}$, $\mu_{i,k}^{(2)} = \mu_{i,k}^{(1)}$ with

$$\mu_{i,k}^{(1)} \propto \exp\left\{\frac{1}{2}\sum_{l=1}^{L}\left[\sum_{\substack{r_1^{(l)}, j\neq i \\ (k,g_j^{(l)})=r_1^{(l)}}} T^{(l)}(A_{ij}^{(l)})\mu_{j,g_j^{(l)}}^{(l)}\bar{\eta}_{r_1^{(l)}}^{(l)} + \sum_{\substack{r_2^{(l)}, j\neq i \\ (k,g_j^{(l)})=r_2^{(l)}}} T^{(l)}(A_{ij}^{(l)})\mu_{j,g_j^{(l)}}^{(l)}\bar{\eta}_{r_2^{(l)}}^{(l)}\right]\right\} \tag{B.24}$$

For $k \notin \{1, \ldots, K\}$

$$\mu_{i,k}^{(l)} \propto \exp\left\{\sum_{\substack{r_3^{(l)}, j\neq i \\ (k,g_j^{(l)})=r_3^{(l)}}} T^{(l)}(A_{ij}^{(l)})\mu_{j,g_j^{(l)}}^{(l)}\bar{\eta}_{r_3^{(l)}}^{(l)} + \sum_{\substack{r_4^{(l)}, j\neq i \\ (k,g_j^{(l)})=r_4^{(l)}}} T^{(l)}(A_{ij}^{(l)})\mu_{j,g_j^{(l)}}^{(l)}\bar{\eta}_{r_4^{(l)}}^{(l)}\right\}. \tag{B.25}$$

Appendix C

Supplementary material Chapter 6

C.1 Proof of Theorem 30

Under Assumptions 1, as $p \to \infty$, for $i \neq j$,

$$\frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} = \mathcal{O}(p^{-\frac{1}{2}}).$$

Since for $i \neq j$, $\frac{(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ}{p} \to 0$ as $p \to \infty$, we shall write a Taylor expansion of f around 0 for the entries K_{ij} $(i \neq j)$.

For $\mathbf{x}_i \in \mathcal{C}_a$ and $\mathbf{x}_j \in \mathcal{C}_b$, we have that

$$\frac{1}{p}(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ = (\mathbf{w}_i^\circ)^T \mathbf{w}_j^\circ + \frac{1}{\sqrt{p}}(\boldsymbol{\mu}_a^\circ)^T \mathbf{w}_j^\circ + \frac{1}{\sqrt{p}}(\mathbf{w}_i^\circ)^T \boldsymbol{\mu}_b^\circ + \frac{1}{p}(\boldsymbol{\mu}_a^\circ)^T \boldsymbol{\mu}_b^\circ$$

Taylor-expanding the non-diagonal elements of $\mathbf{K}$ around their limiting values yields:

$$\begin{aligned}
&f(\frac{1}{p}(\mathbf{x}_i^\circ)^T \mathbf{x}_j^\circ) = f(0)\boldsymbol{\delta}_{i\neq j} + f'(0)\left[(\mathbf{w}_i^\circ)^T\mathbf{w}_j^\circ + \frac{1}{\sqrt{p}}(\boldsymbol{\mu}_a^\circ)^T \mathbf{w}_j^\circ + \frac{1}{\sqrt{p}}(\mathbf{w}_i^\circ)^T(\boldsymbol{\mu}_b^\circ) + \frac{1}{p}(\boldsymbol{\mu}_a^\circ)^T \boldsymbol{\mu}_b^\circ\right] \\
&+ \frac{f''(0)}{2}\left((\mathbf{w}_i^\circ)^T\mathbf{w}_j^\circ\right)^2 \boldsymbol{\delta}_{i\neq j} + f''(0)\frac{1}{\sqrt{p}}(\boldsymbol{\mu}_a^\circ)^T \mathbf{w}_j^\circ(\mathbf{w}_i^\circ)^T\mathbf{w}_j^\circ\boldsymbol{\delta}_{i\neq j} + f''(0)(\mathbf{w}_i^\circ)^T\mathbf{w}_j^\circ\frac{1}{\sqrt{p}}(\mathbf{w}_i^\circ)^T \boldsymbol{\mu}_b^\circ\boldsymbol{\delta}_{i\neq j} \\
&+ f''(0)(\mathbf{w}_i^\circ)^T\mathbf{w}_j^\circ\frac{1}{p}(\boldsymbol{\mu}_a^\circ)^T \boldsymbol{\mu}_b^\circ\boldsymbol{\delta}_{i\neq j} + f(\tau)\boldsymbol{\delta}_{i=j} + f'(\tau)\left((\mathbf{w}_i^\circ)^T\mathbf{w}_i^\circ) - \tau\right)\boldsymbol{\delta}_{i=j} \\
&+ O(n^{-2})
\end{aligned}$$

Hence,

$$\mathbf{K} = f(0)\mathbf{1}\mathbf{1}^T - f(0)\mathbf{I} + \left\{\frac{f''(0)}{2}\left((\mathbf{w}_i^\circ)^T\mathbf{w}_j^\circ\right)^2\boldsymbol{\delta}_{i\neq j}\right\}_{i,j=1}^n + f'(0)\left[\left[(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p}\right]_{a,b=1}^K\right.$$
$$+\left[\mathbf{1}_{n_a}\frac{(\boldsymbol{\mu}_a^\circ)^\mathsf{T}}{\sqrt{p}}(\mathbf{WP})_b\right]_{a,b=1}^K + \left[(\mathbf{WP})_a^\mathsf{T}\frac{(\boldsymbol{\mu}_b^\circ)}{\sqrt{p}}\mathbf{1}_{n_b}^\mathsf{T}\right]_{a,b=1}^K + \mathbf{PW}^\mathsf{T}\mathbf{WP} - \mathcal{D}\{\mathbf{PW}^\mathsf{T}\mathbf{WP}\}\Big]$$
$$+\frac{f''(0)}{\sqrt{p}}\left\{\mathcal{D}\left\{(\mathbf{WP})_a^T\boldsymbol{\mu}_b^\circ\right\}(\mathbf{WP})_a^T(\mathbf{WP})_b\right\}_{a,b=1}^K + \frac{f''(0)}{\sqrt{p}}\left\{(\mathbf{WP})_a^T(\mathbf{WP})_b\mathcal{D}\left\{(\mathbf{WP})_b^T\boldsymbol{\mu}_a^\circ\right\}\right\}_{a,b=1}^k$$
$$f''(0)\mathbf{PWW}^T\mathbf{P}\circ\left\{\frac{1}{p}(\boldsymbol{\mu}_a^\circ)^T\boldsymbol{\mu}_b^\circ\mathbf{1}_{n_a}\mathbf{1}_{n_b}^T\right\}_{a,b=1}^K + O_{1n}(n^{-1})$$

We can see that the spectral norm of the three last matrices in the right-hand side of the above equation are $O_{1n}(n^{-1})$. We thus obtain:

$$\mathbf{K} = f(0)\mathbf{1}\mathbf{1}^T - f(0)\mathbf{I} + \left\{\frac{f''(0)}{2}\left((\mathbf{w}_i^\circ)^T\mathbf{w}_j^\circ\right)^2\boldsymbol{\delta}_{i\neq j}\right\}_{i,j=1}^n + f'(0)\left[\left[(\boldsymbol{\mu}_a^\circ)^\mathsf{T}\boldsymbol{\mu}_b^\circ\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p}\right]_{a,b=1}^K\right.$$
$$+\left[\mathbf{1}_{n_a}\frac{(\boldsymbol{\mu}_a^\circ)^\mathsf{T}}{\sqrt{p}}(\mathbf{WP})_b\right]_{a,b=1}^K + \left[(\mathbf{WP})_a^\mathsf{T}\frac{(\boldsymbol{\mu}_b^\circ)}{\sqrt{p}}\mathbf{1}_{n_b}^\mathsf{T}\right]_{a,b=1}^K + \mathbf{PW}^\mathsf{T}\mathbf{WP} - \mathcal{D}\{\mathbf{PW}^\mathsf{T}\mathbf{WP}\}\Big]$$
$$+O_{1n}(n^{-1})$$

Defining $\boldsymbol{\Phi} = \left\{((\mathbf{w}_i^\circ)^T\mathbf{w}_j^\circ)^2\boldsymbol{\delta}_{i\neq j}\right\}_{i,j=1}^n - \left\{\frac{1}{p^2}\operatorname{tr}\mathbf{C}_a\mathbf{C}_b\mathbf{1}_{n_a}\mathbf{1}_{n_b}^T\right\}_{a,b=1}^K$ and using $(\mathbf{PW}^\mathsf{T}\mathbf{WP})_{ii} \to \tau$, we get

$$\mathbf{K} = f'(0)\mathbf{PW}^\mathsf{T}\mathbf{WP} + \frac{f''(0)}{2}\boldsymbol{\Phi} - f(0)\mathbf{I}_n - \tau f'(0)\mathbf{I}_n + f(0)\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \mathbf{V}\tilde{\boldsymbol{\Omega}}\mathbf{V}^\mathsf{T}$$

where

$$\mathbf{V} = \left[\frac{\mathbf{J}}{\sqrt{p}}, \mathbf{PW}^\mathsf{T}\mathbf{M}\right]$$
$$\tilde{\boldsymbol{\Omega}} = \begin{pmatrix}\mathbf{M}^\mathsf{T}\mathbf{M} + \frac{f''(0)}{2f'(0)}\left[\frac{1}{p}\operatorname{tr}\mathbf{C}_a\mathbf{C}_b\right]_{a,b=1}^K & \mathbf{I}_K\\ \mathbf{I}_K & \mathbf{0}_K\end{pmatrix}$$
$$F(\tau) = \frac{f(\tau) - f(0) - \tau f'(\tau)}{f'(0)}.$$

Since $f'(0) = \frac{\alpha}{\sqrt{p}}$, we should multiply $\mathbf{K}$ by $\sqrt{p}$ to get a bulk of order $\mathcal{O}(1)$. We shall also consider a recentering of the data in the high dimensional feature space i.e., assuming that $\mathbf{K} = f\left(\mathbf{PX}^\mathsf{T}\mathbf{XP}\right) = \Phi(\tilde{\mathbf{X}})^\mathsf{T}\Phi(\tilde{\mathbf{X}})$, we consider the recentered kernel $\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\mathbf{K}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}\right)$. By denoting $\mathbf{P} = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}\right)$, we get the expected result

$$\sqrt{p}\mathbf{P}\left(\mathbf{K} + (f(0) + \tau f'(0))\mathbf{I}_n\right)\mathbf{P} = \alpha\mathbf{PW}^\mathsf{T}\mathbf{WP} + \frac{f''(0)}{2}\sqrt{p}\boldsymbol{\Phi} + \mathbf{V}\boldsymbol{\Omega}\mathbf{V}^\mathsf{T} \tag{C.1}$$

C.2 Proof of Theorem 31

Let us denote by $\bar{\mathbf{X}} \triangleq \alpha \mathbf{P}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{P} + \frac{f''(0)}{2}\mathbf{P}(\sqrt{p}\mathbf{\Phi})\mathbf{P}$ the random noise matrix.

We know from Theorem 30 that $\hat{\mathbf{K}} = \bar{\mathbf{X}} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$ is equivalent to an additive spiked random matrix (see [Chapon et al., 2012] and Section 3.2.4). By using Weyl interlacing formula, the e.s.d. of $\hat{\mathbf{K}}$ converges weakly to the e.s.d of the noise zero mean matrix $\bar{\mathbf{X}}$ since $\hat{\mathbf{K}}$ differs from the former by a low rank matrix. We will thus study the e.s.d. of $\bar{\mathbf{X}}$, a matrix with zero mean and finite variance. We show that its empirical spectral distribution (e.s.d.) $\tilde{\pi}$ converges weakly to $\bar{\pi}^\alpha$ with Stieljes transform $m(z) = \int (t-z)^{-1} \mathrm{d}\bar{\pi}^\alpha(t)$ for $z \in \mathbb{C}^+$.

As stated in Chapter 3, the strategy consists in finding the deterministic limit $m(z)$ for the random quantity $\frac{1}{n}\operatorname{tr}\left(\bar{\mathbf{X}} - z\mathbf{I}_n\right)^{-1}$ which is the Stieljes transform of the e.s.d. $\tilde{\pi}^\alpha$. We use Gaussian calculus to find the aforemetioned deterministic limit. We do not provide details of the Gaussian calculus here, similar calculations being provided in [Kammoun and Couillet, 2017]. The result is as follows

Lemma 48. *Let $\mathbf{Q}_z = (\bar{\mathbf{X}} - z\mathbf{I}_n)^{-1}$. Then, for all $z \in \mathbb{C}^+$, we have the following deterministic equivalents*

$$\mathbf{Q}_z \leftrightarrow m(z)\mathbf{I}_n \tag{C.2}$$

almost surely where $m(z)$ the unique solution of

$$m(z) = \frac{1}{-z + \frac{\alpha}{n}\operatorname{tr}\left(\mathbf{I} + \frac{\alpha}{c_0}m(z)\mathbf{C}^\circ\right)^{-1}\mathbf{C}^\circ - \frac{2\beta^2}{c_0}\omega^2 m(z)}$$

with $\omega = \lim_{p\to\infty}\frac{1}{p}\operatorname{tr}(\mathbf{C}^\circ)^2$

From Lemma 48, we get directly $\frac{1}{n}\operatorname{tr}\mathbf{Q}_z - m(z) \xrightarrow{\text{a.s.}} 0$. This proves Theorem 31.

C.3 Proof of Theorem 32

The proof of Theorem 32 follows similar arguments as the isolated eigenvalues treatment in Section A.3. From Theorem 31, the e.s.d. of $\mathbf{K}$ converges weakly to the limiting law of the noise matrix $\alpha\mathbf{P}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{P} + \frac{f''(0)}{2}\mathbf{P}(\sqrt{p}\mathbf{\Phi})\mathbf{P}$, the two matrices only differing by the low rank matrix $\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}$. The latter matrix shall then induce isolated eigenvalues which are of interest for clustering since their associated eigenvectors contain the relevant information. According to *Non eigenvalue outside the support* result that we do not prove here (similar proof being conducted in for e.g., [Kammoun and Couillet, 2017]), the matrix $\alpha\mathbf{P}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{P} + \frac{f''(0)}{2}\mathbf{P}(\sqrt{p}\mathbf{\Phi})\mathbf{P}$ has asymptotically no eigenvalue outside its main support $\mathcal{S}$. To determine the position of the isolated eigenvalues, we use as in Section A.3, the standard techniques [Benaych-Georges and Nadakuditi, 2012, Baik and Silverstein, 2006].

The eigenvalues ρ of $\mathbf{K}$ falling at a macroscoping distance from the main support $\mathcal{S}$ solve almost surely for large n

$$0 = \det\left(\mathbf{K} - \rho \mathbf{I}_n\right) \tag{C.3}$$

with $\rho \notin \mathcal{S}$. Solving Equation (C.3) is asymptotically equivalent to solving for large n,

$$0 = \det\left(\alpha \mathbf{P}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{P} + \frac{f''(0)}{2}\mathbf{P}(\sqrt{p}\mathbf{\Phi})\mathbf{P} + \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T} - \rho \mathbf{I}_n\right) \tag{C.4}$$

from the approximation of Theorem 30. Let us denote by $\bar{\mathbf{X}} \triangleq \alpha \mathbf{P}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{P} + \frac{f''(0)}{2}\mathbf{P}(\sqrt{p}\mathbf{\Phi})\mathbf{P}$ the random noise matrix and by $\mathbf{Q}_\rho = (\bar{\mathbf{X}} - \rho \mathbf{I}_n)^{-1}$ (for $\rho \notin \mathcal{S}$) its resolvent. Since there is asymptotically no eigenvalue of $\bar{\mathbf{X}}$ outside $\mathcal{S}$, for $\rho \notin \mathcal{S}$, Equation (C.4) is asymptotically equivalent to

$$0 = \det\left(\mathbf{I} + \mathbf{Q}_\rho \mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}\right) = \det\left(\mathbf{I} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho \mathbf{V}\mathbf{\Omega}\right). \tag{C.5}$$

From the argument principle ([Chapon et al., 2012]), the roots of $\mathbf{I} + \mathbf{V}^\mathsf{T}\mathbf{Q}_\rho\mathbf{V}\mathbf{\Omega}$ are asymptotically the same as for its deterministic limit; we will thus next provide this deterministic limit.

Using Theorem 30, we have

$$\mathbf{V}^T\mathbf{Q}_\rho\mathbf{V} = \begin{bmatrix} \frac{1}{p}\mathbf{J}^T\mathbf{Q}_\rho\mathbf{J} & \frac{1}{\sqrt{p}}\mathbf{J}^T\mathbf{Q}_\rho\mathbf{P}\mathbf{W}^T\mathbf{M} \\ \frac{1}{\sqrt{p}}\mathbf{M}\mathbf{W}\mathbf{P}\mathbf{Q}_\rho\mathbf{J} & \mathbf{M}\mathbf{W}\mathbf{P}\mathbf{Q}_\rho\mathbf{P}\mathbf{W}^T\mathbf{M} \end{bmatrix} \tag{C.6}$$

Lemma 49 (Deterministic equivalents)**.** *Let* $\mathbf{Q}_z = (\bar{\mathbf{X}} - z\mathbf{I}_n)^{-1}$. *Then, for all* $z \in \mathbb{C}^+$, *we have the following deterministic equivalents*

$$\mathbf{M}\mathbf{W}\mathbf{P}\mathbf{Q}_z\mathbf{P}\mathbf{W}^T\mathbf{M} = \frac{1}{p}\operatorname{tr}\left[\left(\mathbf{I}_p + \frac{\alpha}{c_0}m(z)\mathbf{C}^\circ\right)^{-1}\mathbf{C}^\circ\right] c_0^{-1}m(z)\mathbf{M}^T\mathbf{M} + o(1) \tag{C.7}$$

$$\mathbf{M}\mathbf{W}\mathbf{P}\mathbf{Q}_z\mathbf{J} = o(1) \tag{C.8}$$

almost surely where $m(z)$ *the unique solution of*

$$m(z) = \frac{1}{-z + \frac{\alpha}{n}\operatorname{tr}\left(\mathbf{I} + \frac{\alpha}{c_0}m(z)\mathbf{C}^\circ\right)^{-1}\mathbf{C}^\circ - \frac{2\beta^2}{c_0}\omega^2 m(z)}.$$

Lemma 49 provides the necessary ingredients for computing deterministic limits for the entries of $\mathbf{V}^T\mathbf{Q}_\rho\mathbf{V}$ in Equation (C.6).

We get

$$\mathbf{V}^\mathsf{T}\mathbf{Q}_\rho\mathbf{V} = \begin{pmatrix} c_0^{-1}m(\rho)\mathcal{D}(\mathbf{c}) & 0 \\ 0 & \frac{1}{p}\operatorname{tr}\left[\left(\mathbf{I}_p + \frac{\alpha}{c_0}m(\rho)\mathbf{C}^\circ\right)^{-1}\mathbf{C}^\circ\right] c_0^{-1}m(\rho)\mathbf{M}^T\mathbf{M}. \end{pmatrix}.$$

Right multiplying by $\mathbf{\Omega}$, we obtain

$$\mathbf{I}+\mathbf{V}^T\mathbf{Q}_\rho\mathbf{V}\mathbf{\Omega}=\begin{bmatrix}\mathbf{I}_K+m(\rho)\mathcal{D}(\mathbf{c})c_0^{-1}\left[\alpha\mathbf{M}^T\mathbf{M}+\beta\mathbf{T}\right] & \alpha m(\rho)\mathcal{D}(\mathbf{c})c_0^{-1}\\ \frac{\alpha}{p}\operatorname{tr}\left[(\mathbf{I}_K+\frac{\alpha}{c_0}m(\rho)\mathbf{C}^\circ)^{-1}\mathbf{C}^\circ\right]\mathbf{M}^T\mathbf{M} & \mathbf{I}_K\end{bmatrix}+o(1)$$

Hence

$$\begin{aligned}\det\left[\mathbf{I}+\mathbf{V}^T\mathbf{Q}_\rho\mathbf{V}\mathbf{\Omega}\right]&=\det\left\{\left[\alpha m(\rho)\mathcal{D}(\mathbf{c})c_0^{-1}-\alpha^2m^2(\rho)\mathcal{D}(\mathbf{c})c_0^{-2}\frac{1}{p}\operatorname{tr}\left[\mathbf{I}_K+\frac{\alpha}{c_0}m(\rho)\mathbf{C}^\circ\right]^{-1}\mathbf{C}^\circ\right]\mathbf{M}^T\mathbf{M}\right.\\&\left.+m(\rho)\mathcal{D}(\mathbf{c})c_0^{-1}\beta\mathbf{T}+\mathbf{I}\right\}\\&=\det\left[\alpha m(\rho)c_0^{-1}\frac{1}{p}\operatorname{tr}\left[\mathbf{I}_K+\frac{\alpha}{c_0}m(\rho)\mathbf{C}^\circ\right]^{-1}\mathcal{D}(\mathbf{c})\mathbf{M}^T\mathbf{M}+m(\rho)\mathcal{D}(\mathbf{c})c_0^{-1}\beta\mathbf{T}+\mathbf{I}_K\right]\end{aligned}$$

Let $\mathbf{G}_\rho$ be given by:

$$\mathbf{G}_\rho=\mathbf{I}_K+\frac{m(\rho)}{c_0}\mathcal{D}(\mathbf{c})\left[\alpha g(\rho)\mathbf{M}^T\mathbf{M}+\beta\mathbf{T}\right]$$

with $g(\rho)=\frac{1}{p}\operatorname{tr}(\mathbf{I}_p+\frac{\alpha m(\rho)}{c_0}\mathbf{C}^\circ)^{-1}$. The isolated eigenvalues are thus the values $\rho\notin\mathcal{S}$ for which $\mathbf{G}_\rho$ has zero eigenvalues.

C.4 Proof of Theorem 33

Let λ be an eigenvalue of $\hat{\mathbf{K}}$ converging to ρ with multiplicity $\ell_\rho\geq 1$. Define $\mathbf{\Pi}_\rho$ a projector on the eigenspace associated with all eigenvalues of $\hat{\mathbf{K}}$ converging to ρ. We wish to investigate the limit of the matrix $\frac{1}{p}\mathbf{J}^T\mathbf{\Pi}_\rho\mathbf{J}$, the entries of which are up to scaling the desired quantities $\eta_i^a\eta_i^b$ when $\ell_\rho=1$. By residue calculus, we have

$$\frac{1}{p}\mathbf{J}^T\mathbf{\Pi}_\rho\mathbf{J}=-\frac{1}{2\pi\imath}\oint_{\mathcal{C}_\rho}\frac{1}{p}\mathbf{J}^T\left(\hat{\mathbf{K}}-zI_n\right)^{-1}\mathbf{J}\mathrm{d}z$$

for all large n almost surely where $\mathcal{C}_\rho$ is a complex (positively oriented and with winding number one) contour circling around ρ only. By Woodbury's matrix inverse identity, we have:

$$\frac{1}{p}\mathbf{J}^T\left(\hat{\mathbf{K}}-zI_n\right)^{-1}J=\frac{1}{p}\mathbf{J}^T\mathbf{Q}_z\mathbf{J}-\frac{1}{p}\mathbf{J}^T\mathbf{Q}_z\mathbf{V}\mathbf{\Omega}\left(\mathbf{I}_k+\mathbf{V}^T\mathbf{Q}_z\mathbf{V}\mathbf{\Omega}\right)^{-1}\mathbf{V}^T\mathbf{Q}_zJ$$

where

$$\frac{1}{\sqrt{p}}\mathbf{V}^T\mathbf{Q}_z\mathbf{J} = \left[\frac{1}{p}\mathbf{J}^T\mathbf{Q}_z\mathbf{J}\ \ 0\right]^T + o(1) = c_0^{-1}m(z)\mathcal{D}(\mathbf{c}) + o(1)$$

$$\frac{1}{\sqrt{p}}\mathbf{J}^T\mathbf{Q}_z\mathbf{V}\mathbf{\Omega} = \left[c_0^{-1}m(z)\mathcal{D}(\mathbf{c})\left(\alpha\mathbf{M}^T\mathbf{M} + \beta\mathbf{T}\right)\ \ \alpha c_0^{-1}m(z)\mathcal{D}(\mathbf{c})\right] + o(1)$$

$$\left(I_k + \mathbf{V}^T\mathbf{Q}_z\mathbf{V}\mathbf{\Omega}\right)^{-1} = \begin{bmatrix} \mathbf{G}_z^{-1} & * \\ -\frac{\alpha c_0^{-1}m(z)}{p}\operatorname{tr}\left[\mathbf{I} + \frac{\alpha}{c_0}m(z)\mathbf{C}^\circ\right]^{-1}\mathbf{C}^\circ\mathbf{M}^T\mathbf{M}\mathbf{G}_z^{-1} & * \end{bmatrix}$$

Using these approximations, we ultimately find,

$$\begin{aligned}
&\frac{1}{p}\mathbf{J}^T\mathbf{Q}_z\mathbf{V}\mathbf{\Omega}\left(\mathbf{I}_k + \mathbf{V}^T\mathbf{Q}_z\mathbf{V}\mathbf{\Omega}\right)^{-1}\mathbf{V}^T\mathbf{Q}_z\mathbf{J} = c_0^{-2}m^2(z)\mathcal{D}(\mathbf{c})\beta\mathbf{T}\mathbf{G}_z^{-1}\mathcal{D}(\mathbf{c}) \\
&+ \alpha c_0^{-2}m^2(z)\mathcal{D}(\mathbf{c})\mathbf{M}^T\mathbf{M}\mathbf{G}_z^{-1}\mathcal{D}(\mathbf{c})\frac{1}{p}\operatorname{tr}(\mathbf{I} + \frac{\alpha}{c_0}m(z)\mathbf{C}^\circ)^{-1} \\
&= c_0^{-2}m^2(z)\mathcal{D}(\mathbf{c})\left[\beta\mathbf{T} + \frac{\alpha}{p}\operatorname{tr}\left(\mathbf{I}_p + \frac{\alpha}{c_0}m(z)\mathbf{C}^\circ\right)^{-1}\mathbf{M}^T\mathbf{M}\right]\mathbf{G}_z^{-1}\mathcal{D}(\mathbf{c}) \\
&= c_0^{-1}m(z)\mathcal{D}(\mathbf{c}) - c_0^{-1}m(z)\mathbf{G}_z^{-1}\mathcal{D}(\mathbf{c})
\end{aligned}$$

The first term gives zero residue asymptotically. Hence,

$$\frac{1}{p}\mathbf{J}^T\hat{\mathbf{\Pi}}_\rho\mathbf{J} = -\frac{1}{2\pi\imath}\oint_{\mathcal{C}_\rho} c_0^{-1}m(z)\mathbf{G}_z^{-1}\mathcal{D}(\mathbf{c})\mathrm{d}z + o(1)$$

Letting $g(z) = \frac{1}{p}\operatorname{tr}\left(\mathbf{I}_p + \frac{\alpha}{c_0}m(z)\mathbf{C}^\circ\right)^{-1}$,

$$\frac{1}{p}\mathbf{J}^T\hat{\mathbf{\Pi}}_\rho\mathbf{J} = -\frac{1}{2\pi\imath}\oint_{\mathcal{C}_\rho} c_0^{-1}m(z)\left[(D(\mathbf{c}))^{-1} + \frac{m(z)}{c_0}\left[\alpha g(z)\mathbf{M}^T\mathbf{M} + \beta\mathbf{T}\right]\right]^{-1}\mathrm{d}z + o(1)$$

Let $\mathcal{G} = \left[(D(\mathbf{c}))^{-1} + \frac{m(z)}{c_0}\left[\alpha g(z)\mathbf{M}^T\mathbf{M} + \beta\mathbf{T}\right]\right]$ and $\mathbf{V}_\rho$ its eigenvectors associated with the eigenvalue 0. Then following the same reasoning for the residue calculation as in Appendix A.4,

$$\frac{1}{p}\mathbf{J}^T\hat{\mathbf{\Pi}}_\rho\mathbf{J} = -c_0^{-1}m(\rho)\sum_{i=1}^{\ell_\rho}\frac{\mathbf{V}_{\rho,i}\mathbf{V}_{\rho,i}^T}{\mathbf{V}_{\rho,i}^T\mathcal{G}_\rho'\mathbf{V}_{\rho,i}} + o(1)$$

To complete the proof, we simplify the above expression for $c_1 = c_2 = \frac{1}{2}$. In that case, $\mathcal{G}$ has zero eigenvalue at:

$$2 + \frac{m(\rho)}{c_0}\left(\alpha g(\rho)\operatorname{tr}\mathbf{M}^T\mathbf{M} + \beta\operatorname{tr}\mathbf{T}\right) = 2 + \frac{m(\rho)}{c_0}\left[\alpha g(\rho)\frac{\|\ \mathring{\boldsymbol{\mu}}_1 - \mathring{\boldsymbol{\mu}}_2\ \|^2}{2} + \frac{\beta}{2\sqrt{p}}\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2\right]$$

Let $\delta = \| \mathring{\boldsymbol{\mu}}_1 - \mathring{\boldsymbol{\mu}}_2 \|^2$ and $\theta = \frac{1}{\sqrt{p}} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$. Then, ρ satisfies:

$$2 + \frac{m(\rho)}{c_0} \left[\alpha g(\rho) \frac{\delta}{2} + \frac{\beta\theta}{2} \right] = 0.$$

Returning back to $\frac{1}{p}\mathbf{J}^T \hat{\mathbf{\Pi}}_\rho \mathbf{J}$. We have:

$$\begin{aligned}
\mathbf{V}_{\rho,i}^T \mathcal{G}'(\rho) \mathbf{V}_{\rho,i} &= m'(\rho) c_0^{-1} \left[\alpha g(\rho) \operatorname{tr} \mathbf{M}^T \mathbf{M} + \beta \operatorname{tr} \mathbf{T} \right] + m(\rho) c_0^{-1} \alpha g'(\rho) \operatorname{tr} \mathbf{M}^T \mathbf{M} \\
&= \frac{m'(\rho)}{c_0} \left(\frac{\alpha}{4} g(\rho) \delta + \frac{\beta}{4} \theta \right) + \frac{m(\rho)}{c_0} \alpha g'(\rho) \frac{\delta}{4} = -\frac{m'(\rho)}{m(\rho)} + \frac{m(\rho)}{c_0} \alpha g'(\rho) \frac{\delta}{4}
\end{aligned}$$

Finally,

$$\frac{1}{p}\mathbf{J}^T \hat{\mathbf{\Pi}}_\rho \mathbf{J} = \frac{c_0^{-1} m^2(\rho) \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}}{m'(\rho) \left[1 - \frac{m^2(\rho)}{m'(\rho)} \alpha g'(\rho) \frac{\delta}{4} \right]} + o(1)$$

and we get the expected result

$$\eta_i^2 = \frac{m^2(\rho)}{m'(\rho) \left[1 - \frac{m^2(\rho)}{m'(\rho)} \alpha g'(\rho) \frac{\delta}{4} \right]} + o(1).$$

Bibliography

[Adamic and Glance, 2005] Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM.

[Aicher et al., 2014] Aicher, C., Jacobs, A. Z., and Clauset, A. (2014). Learning latent block structure in weighted networks. *Journal of Complex Networks*, page cnu026.

[Ajanki et al., 2015] Ajanki, O., Erdos, L., and Krüger, T. (2015). Quadratic vector equations on complex upper half-plane. *arXiv preprint arXiv:1506.05095*.

[Amelio and Pizzuti, 2014] Amelio, A. and Pizzuti, C. (2014). Community detection in multidimensional networks. In *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*, pages 352–359. IEEE.

[Andrzejak et al., 2001] Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907.

[Arnold et al., 1994] Arnold, L., Gundlach, V. M., and Demetrius, L. (1994). Evolutionary formalism for products of positive random matrices. *The Annals of Applied Probability*, pages 859–901.

[Avrachenkov et al., 2015] Avrachenkov, K., Cottatellucci, L., and Kadavankandy, A. (2015). Spectral properties of random matrices for stochastic block model. In *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2015 13th International Symposium on*, pages 537–544. IEEE.

[Bai, 2008] Bai, Z. (2008). Circular law. In *Advances In Statistics*, pages 128–163. World Scientific.

[Bai and Silverstein, 2010] Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer.

[Bai and Silverstein, 1998] Bai, Z.-D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Annals of probability*, pages 316–345.

[Baik et al., 2005] Baik, J., Ben Arous, G., and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, pages 1643–1697.

[Baik and Silverstein, 2006] Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408.

[Barbillon et al., 2017] Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2017). Stochastic block models for multiplex networks: an application to a multilevel network of researchers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1):295–314.

[Benaych-Georges and Couillet, 2016] Benaych-Georges, F. and Couillet, R. (2016). Spectral analysis of the gram matrix of mixture models. *ESAIM: Probability and Statistics*, 20:217–237.

[Benaych-Georges and Nadakuditi, 2012] Benaych-Georges, F. and Nadakuditi, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135.

[Billingsley, 1995] Billingsley, P. (1995). Probability and measure. wiley series in probability and mathematical statistics.

[Bińkowski et al., 2018] Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

[Blei et al., 2006] Blei, D. M., Jordan, M. I., et al. (2006). Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143.

[Boden et al., 2012] Boden, B., Günnemann, S., Hoffmann, H., and Seidl, T. (2012). Mining coherent subgraphs in multi-layer graphs with edge labels. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1258–1266. ACM.

[Bordenave et al., 2015] Bordenave, C., Lelarge, M., and Massoulié, L. (2015). Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[Boulos et al., 2017] Boulos, R. E., Tremblay, N., Arneodo, A., Borgnat, P., and Audit, B. (2017). Multi-scale structural community organisation of the human genome. *BMC bioinformatics*, 18(1):209.

[Brandes et al., 2007] Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2007). On finding graph clusterings with maximum modularity. In *International Workshop on Graph-Theoretic Concepts in Computer Science*, pages 121–132. Springer.

[Brodka and Grecki, 2012] Brodka, P. and Grecki, T. (2012). *mLFR Benchamark: Testing Community Detection Algorithms in Multilayer, Multiplex and Multiple Social Networks*.

[Chapelle et al., 2003] Chapelle, O., Weston, J., and Schölkopf, B. (2003). Cluster kernels for semi-supervised learning. In *Advances in neural information processing systems*, pages 601–608.

[Chapon et al., 2012] Chapon, F., Couillet, R., Hachem, W., and Mestre, X. (2012). The outliers among the singular values of large rectangular random matrices with additive fixed rank deformation. *arXiv preprint arXiv:1207.0471*.

[Chen et al., 2015a] Chen, H., Chen, J., et al. (2015a). Functional organization of the human 4d nucleome. *Proceedings of the National Academy of Sciences*, 112(26):8002–8007.

[Chen et al., 2016] Chen, J., Hero, A. O., and Rajapakse, I. (2016). Spectral identification of topological domains. *Bioinformatics*, page btw221.

[Chen and Yuan, 2006] Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290.

[Chen et al., 2015b] Chen, Y., Li, X., and Xu, J. (2015b). Convexified modularity maximization for degree-corrected stochastic block models. *arXiv preprint arXiv:1512.08425*.

[Choromanska et al., 2015] Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2015). The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204.

[Chung, 1997] Chung, F. R. (1997). *Spectral graph theory*, volume 92. American Mathematical Soc.

[Clauset et al., 2004] Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6):066111.

[Cline et al., 2007] Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using cytoscape. *Nature protocols*, 2(10):2366.

[Coja-Oghlan and Lanka, 2009] Coja-Oghlan, A. and Lanka, A. (2009). Finding planted partitions in random graphs with general degree distributions. *SIAM Journal on Discrete Mathematics*, 23(4):1682–1714.

[Comar et al., 2012] Comar, P. M., Tan, P.-N., and Jain, A. K. (2012). A framework for joint community detection across multiple related networks. *Neurocomputing*, 76(1):93–104.

[Couillet and Benaych-Georges, 2016] Couillet, R. and Benaych-Georges, F. (2016). Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454.

[Couillet and Debbah, 2011] Couillet, R. and Debbah, M. (2011). *Random matrix methods for wireless communications*. Cambridge University Press.

[Couillet et al., 2018] Couillet, R., Liao, Z., and Mai, X. (2018). Classification asymptotics in the random matrix regime. In *European Signal Processing Conference (EUSIPCO'18)*.

[Couillet et al., 2016a] Couillet, R., Wainrib, G., Sevi, H., and Tiomoko Ali, H. (2016a). The asymptotic performance of linear echo state neural networks. *Journal of Machine Learning Research*, 17(178):1–35.

[Couillet et al., 2016b] Couillet, R., Wainrib, G., Sevi, H., and Tiomoko Ali, H. (2016b). Training performance of echo state neural networks. In *Statistical Signal Processing Workshop (SSP), 2016 IEEE*, pages 1–4. IEEE.

[Couillet et al., 2016c] Couillet, R., Wainrib, G., Tiomoko Ali, H., and Sevi, H. (2016c). A random matrix approach to echo-state neural networks. In *International Conference on Machine Learning*, pages 517–525.

[Dauphin et al., 2014] Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941.

[De Bacco et al., 2017] De Bacco, C., Power, E. A., Larremore, D. B., and Moore, C. (2017). Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4):042317.

[De Domenico et al., 2015a] De Domenico, M., Lancichinetti, A., Arenas, A., and Rosvall, M. (2015a). Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Physical Review X*, 5(1):011027.

[De Domenico et al., 2015b] De Domenico, M., Nicosia, V., Arenas, A., and Latora, V. (2015b). Structural reducibility of multilayer networks. *Nature communications*, 6:6864.

[Decelle et al., 2011a] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011a). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106.

[Decelle et al., 2011b] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011b). Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701.

[Dekker et al., 2017] Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., O'shea, C. C., Park, P. J., Ren, B., et al. (2017). The 4d nucleome project. *Nature*, 549(7671):219.

[Dixon et al., 2015] Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331.

[Dixon et al., 2012] Dixon, J. R., Selvaraj, S., Yue, F., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

[Duch and Arenas, 2005] Duch, J. and Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical review E*, 72(2):027104.

[El Karoui et al., 2010] El Karoui, N. et al. (2010). The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50.

[Fortunato, 2010] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

[Gamerman and Lopes, 2006] Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC.

[Gao et al., 2016] Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2016). Community detection in degree-corrected block models. *arXiv preprint arXiv:1607.06993*.

[Girko, 2001] Girko, V. L. (2001). *Theory of stochastic canonical equations*, volume 2. Springer Science & Business Media.

[Girko, 2012] Girko, V. L. (2012). *Theory of random determinants*, volume 45. Springer Science & Business Media.

[Goldenberg et al., 2010] Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M., et al. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233.

[Gretton et al., 2012] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

[Guimera et al., 2005] Guimera, R., Mossa, S., Turtschi, A., and Amaral, L. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799.

[Guimera et al., 2004] Guimera, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101.

[Gulikers et al., 2015] Gulikers, L., Lelarge, M., and Massoulié, L. (2015). A spectral method for community detection in moderately-sparse degree-corrected stochastic block models. *arXiv preprint arXiv:1506.08621*.

[Gulikers et al., 2016] Gulikers, L., Lelarge, M., and Massoulié, L. (2016). Non-backtracking spectrum of degree-corrected stochastic block models. *arXiv preprint arXiv:1609.02487*.

[Hachem et al., 2013] Hachem, W., Loubaton, P., Mestre, X., Najim, J., and Vallet, P. (2013). A subspace estimator for fixed rank perturbations of large random matrices. *Journal of Multivariate Analysis*, 114:427–447.

[Hachem et al., 2007] Hachem, W., Loubaton, P., Najim, J., et al. (2007). Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930.

[Han et al., 2015] Han, Q., Xu, K., and Airoldi, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning*, pages 1511–1520.

[Hartigan and Wong, 1979] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

[Hero and Rajaratnam, 2011] Hero, A. and Rajaratnam, B. (2011). Large-scale correlation screening. *Journal of the American Statistical Association*, 106(496):1540–1552.

[Hoydis et al., 2011] Hoydis, J., Couillet, R., and Debbah, M. (2011). Iterative deterministic equivalents for the performance analysis of communication systems. *arXiv preprint arXiv:1112.4167*.

[Jin et al., 2015] Jin, J. et al. (2015). Fast community detection by score. *The Annals of Statistics*, 43(1):57–89.

[Jordan et al., 1999] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

[Kammoun and Couillet, 2017] Kammoun, A. and Couillet, R. (2017). Subspace kernel spectral clustering of large dimensional data. *(submitted to) Annals of Applied Probability, 2017*.

[Karrer and Newman, 2011] Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107.

[Kawamoto and Kabashima, 2015] Kawamoto, T. and Kabashima, Y. (2015). Limitations in the spectral method for graph partitioning: Detectability threshold and localization of eigenvectors. *Physical Review E*, 91(6):062803.

[Kim et al., 2017] Kim, J., Lee, J.-G., and Lim, S. (2017). Differential flattening: A novel framework for community detection in multi-layer graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):27.

[Krzakala et al., 2013] Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., and Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940.

[Laloux et al., 2000] Laloux, L., Cizeau, P., Potters, M., and Bouchaud, J.-P. (2000). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(03):391–397.

[Lancichinetti et al., 2008] Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110.

[LeCun, 1998] LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

[Lei et al., 2015] Lei, J., Rinaldo, A., et al. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237.

[Lesieur et al., 2015] Lesieur, T., Krzakala, F., and Zdeborová, L. (2015). Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 680–687. IEEE.

[Lesieur et al., 2017] Lesieur, T., Krzakala, F., and Zdeborová, L. (2017). Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403.

[Liao et al.,] Liao, Z., Chitour, Y., and Couillet, R. Almost global convergence to global minima for gradient descent in deep linear networks.

[Liao and Couillet, 2017] Liao, Z. and Couillet, R. (2017). A large dimensional analysis of least squares support vector machines. *arXiv preprint arXiv:1701.02967*.

[Liao and Couillet, 2018] Liao, Z. and Couillet, R. (2018). On the spectrum of random features maps of high dimensional data. *arXiv preprint arXiv:1805.11916*.

[Lieberman-Aiden et al., 2009] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.

[Linden et al., 2003] Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80.

[Louart et al., 2018] Louart, C., Liao, Z., Couillet, R., et al. (2018). A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248.

[Lyzinski et al., 2014] Lyzinski, V., Sussman, D. L., Tang, M., Athreya, A., Priebe, C. E., et al. (2014). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905–2922.

[Mai and Couillet, 2017] Mai, X. and Couillet, R. (2017). A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *arXiv preprint arXiv:1711.03404*.

[Marčenko and Pastur, 1967] Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457.

[Marchenko and Pastur, 1967] Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536.

[Marcotte et al., 1999] Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753.

[Massoulié, 2014] Massoulié, L. (2014). Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM.

[Mehta, 2004] Mehta, M. L. (2004). *Random matrices*, volume 142. Elsevier.

[Minka, 2001] Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

[Mossel et al., 2013] Mossel, E., Neeman, J., and Sly, A. (2013). A proof of the block model threshold conjecture. *Combinatorica*, pages 1–44.

[Mossel et al., 2015] Mossel, E., Neeman, J., and Sly, A. (2015). Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461.

[Mucha et al., 2010] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878.

[Nadakuditi and Newman, 2012] Nadakuditi, R. R. and Newman, M. E. (2012). Graph spectra and the detectability of community structure in networks. *Physical review letters*, 108(18):188701.

[Newman, 2010] Newman, M. (2010). *Networks: an introduction*. Oxford university press.

[Newman, 2013] Newman, M. (2013). Spectral community detection in sparse networks. *arXiv preprint arXiv:1308.6494*.

[Newman, 2016] Newman, M. (2016). Community detection in networks: Modularity optimization and maximum likelihood are equivalent. *arXiv preprint arXiv:1606.02319*.

[Newman, 2004] Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133.

[Newman, 2006a] Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104.

[Newman, 2006b] Newman, M. E. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

[Ng et al., 2002] Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856.

[Ng et al., 2012] Ng, S. K., Krishnan, T., and McLachlan, G. J. (2012). The em algorithm. In *Handbook of computational statistics*, pages 139–172. Springer.

[Nicosia and Latora, 2015] Nicosia, V. and Latora, V. (2015). Measuring and modeling correlations in multiplex networks. *Physical Review E*, 92(3):032805.

[Opper and Saad, 2001] Opper, M. and Saad, D. (2001). *Advanced mean field methods: Theory and practice*. MIT press.

[Oselio et al., 2015] Oselio, B., Kulesza, A., and Hero, A. (2015). Information extraction from large multi-layer social networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5451–5455. IEEE.

[Oselio et al., 2014] Oselio, B., Kulesza, A., and Hero, A. O. (2014). Multi-layer graph analysis for dynamic social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):514–523.

[Pastur and Shcherbina, 2011] Pastur, L. A. and Shcherbina, M. (2011). *Eigenvalue distribution of large random matrices*, volume 171. American Mathematical Society Providence, RI.

[Paul and Chen, 2015] Paul, S. and Chen, Y. (2015). Community detection in multi-relational data with restricted multi-layer stochastic blockmodel. *arXiv preprint arXiv:1506.02699*.

[Paul and Chen, 2016] Paul, S. and Chen, Y. (2016). Null models and modularity based community detection in multi-layer networks. *arXiv preprint arXiv:1608.00623*.

[Peixoto, 2015] Peixoto, T. P. (2015). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807.

[Pennington and Worah, 2017] Pennington, J. and Worah, P. (2017). Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646.

[Pizzuti and Socievole, 2017] Pizzuti, C. and Socievole, A. (2017). Many-objective optimization for community detection in multi-layer networks. In *Evolutionary Computation (CEC), 2017 IEEE Congress on*, pages 411–418. IEEE.

[Qin and Rohe, 2013] Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128.

[Reyes and Rodriguez, 2016] Reyes, P. and Rodriguez, A. (2016). Stochastic blockmodels for exchangeable collections of networks. *arXiv preprint arXiv:1606.05277*.

[Saade et al., 2014] Saade, A., Krzakala, F., and Zdeborová, L. (2014). Spectral clustering of graphs with the bethe hessian. In *Advances in Neural Information Processing Systems*, pages 406–414.

[Schölkopf, 2001] Schölkopf, B. (2001). The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307.

[Scholkopf and Smola, 2001] Scholkopf, B. and Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[Silverstein and Bai, 1995] Silverstein, J. W. and Bai, Z. (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192.

[Stanley et al., 2016] Stanley, N., Shai, S., Taylor, D., and Mucha, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE transactions on network science and engineering*, 3(2):95–105.

[Sweet et al., 2014] Sweet, T. M., Thomas, A. C., and Junker, B. W. (2014). Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. *Handbook on mixed membership models and their applications*, pages 463–488.

[Tang et al., 2009] Tang, L., Wang, X., and Liu, H. (2009). Uncoverning groups via heterogeneous interaction analysis. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 503–512. IEEE.

[Tang et al., 2012] Tang, L., Wang, X., and Liu, H. (2012). Community detection via heterogeneous interaction analysis. *Data mining and knowledge discovery*, 25(1):1–33.

[Taylor et al., 2016] Taylor, D., Shai, S., Stanley, N., and Mucha, P. J. (2016). Enhanced detectability of community structure in multilayer networks through layer aggregation. *Physical review letters*, 116(22):228301.

[Telatar, 1999] Telatar, E. (1999). Capacity of multi-antenna gaussian channels. *Transactions on Emerging Telecommunications Technologies*, 10(6):585–595.

[Tiomoko Ali and Couillet, 2016a] Tiomoko Ali, H. and Couillet, R. (2016a). community detection in heterogeneous networks. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 1385–1389. IEEE.

[Tiomoko Ali and Couillet, 2016b] Tiomoko Ali, H. and Couillet, R. (2016b). Performance analysis of spectral community detection in realistic graph models. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4548–4552. IEEE.

[Tiomoko Ali and Couillet, 2016c] Tiomoko Ali, H. and Couillet, R. (2016c). Performance analysis of spectral community detection in realistic graph models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*.

[Tiomoko Ali and Couillet, 2018] Tiomoko Ali, H. and Couillet, R. (2018). Improved spectral community detection in large heterogeneous networks. *Journal of Machine Learning Research*, 18:1–49.

[Tiomoko Ali et al., 2018a] Tiomoko Ali, H., Kammoun, A., and Couillet, R. (2018a). Random matrix asymptotic of inner product kernel spectral clustering. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*. IEEE.

[Tiomoko Ali et al., 2018b] Tiomoko Ali, H., Kammoun, A., and Couillet, R. (2018b). Random matrix-improved kernels for large dimensional spectral clustering. In *Statistical Signal Processing Workshop (SSP), 2016 IEEE*, pages 1–4. IEEE.

[Tiomoko Ali et al., 2018c] Tiomoko Ali, H., Liu, S., Yilmaz, Y., Hero, A., Couillet, R., and Rajapakse, I. (2018c). Latent heterogeneous multilayer community detection.

[Valles-Catala et al., 2016] Valles-Catala, T., Massucci, F. A., Guimera, R., and Sales-Pardo, M. (2016). Multilayer stochastic block models reveal the multilayer structure of complex networks. *Physical Review X*, 6(1):011036.

[Von Luxburg et al., 2008] Von Luxburg, U., Belkin, M., and Bousquet, O. (2008). Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586.

[Wainwright et al., 2008] Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

[Wei and Cheng, 1989] Wei, Y.-C. and Cheng, C.-K. (1989). Towards efficient hierarchical designs by ratio cut partitioning. In *Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers., 1989 IEEE International Conference on*, pages 298–301. IEEE.

[Wigner, 1993] Wigner, E. P. (1993). Characteristic vectors of bordered matrices with infinite dimensions ii. In *The Collected Works of Eugene Paul Wigner*, pages 541–545. Springer.

[Wilson et al., 2017] Wilson, J. D., Palowitch, J., Bhamidi, S., and Nobel, A. B. (2017). Community extraction in multilayer networks with heterogeneous community structure. *The Journal of Machine Learning Research*, 18(1):5458–5506.

[Wishart, 1928] Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52.

[Xiang et al., 2012] Xiang, J., Hu, X.-G., Zhang, X.-Y., Fan, J.-F., Zeng, X.-L., Fu, G.-Y., Deng, K., and Hu, K. (2012). Multi-resolution modularity methods and their limitations in community detection. *The European Physical Journal B*, 85(10):352.

[Yedidia et al., 2003] Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2003). Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239.

[Zeng et al., 2006] Zeng, Z., Wang, J., Zhou, L., and Karypis, G. (2006). Coherent closed quasi-clique discovery from large dense graph databases. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 797–802. ACM.

[Zhang and Zhou, 2017] Zhang, A. Y. and Zhou, H. H. (2017). Theoretical and computational guarantees of mean field variational inference for community detection. *arXiv preprint arXiv:1710.11268*.

[Zhang et al., 2013] Zhang, Z., Li, Q., Zeng, D., and Gao, H. (2013). User community discovery from multi-relational networks. *Decision Support Systems*, 54(2):870–879.

[Zhao et al., 2012] Zhao, Y., Levina, E., Zhu, J., et al. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292.

université PARIS-SACLAY | ÉCOLE DOCTORALE Sciences et technologies de l'information et de la communication (STIC)

Titre : Nouvelles méthodes pour l'apprentissage non-supervisé en grandes dimensions.

Mots clés : Apprentissage non supervisé, Classification de données en grandes dimensions, Détection de communautés, Théorie des matrices aléatoires, Inférence Bayésienne.

Résumé : Motivée par les récentes avancées dans l'analyse théorique des performances des algorithmes d'apprentissage automatisé, cette thèse s'intéresse à l'analyse de performances et à l'amélioration de la classification non-supervisée de données et graphes en grande dimension. Spécifiquement, dans la première grande partie de cette thèse, en s'appuyant sur des outils avancés de la théorie des grandes matrices aléatoires, nous analysons les performances de méthodes spectrales sur des modèles de graphes réalistes et denses ainsi que sur des données en grandes dimensions en étudiant notamment les valeurs propres et vecteurs propres des matrices d'affinités de ces données. De nouvelles méthodes améliorées sont proposées sur la base de cette analyse théorique et démontrent à travers de nombreuses simulations que leurs performances sont meilleures comparées aux méthodes de l'état de l'art. Dans la seconde partie de la thèse, nous proposons un nouvel algorithme pour la détection de communautés hétérogènes entre plusieurs couches d'un graphe à plusieurs types d'interaction. Une approche bayésienne variationnelle est utilisée pour approximer la distribution apostériori des variables latentes du modèle. Toutes les méthodes proposées dans cette thèse sont utilisées sur des bases de données synthétiques et sur des données réelles et présentent de meilleures performances en comparaison aux approches standard de classification dans les contextes susmentionnés.

Title: New methods for large-scale unsupervised learning.

Keywords: Unsupervised learning, High dimensional data clustering, Community detection, Random matrix theory, Bayesian inference.

Abstract: Spurred by recent advances on the theoretical analysis of the performances of the data-driven machine learning algorithms, this thesis tackles the performance analysis and improvement of high dimensional data and graph clustering. Specifically, in the first bigger part of the thesis, using advanced tools from random matrix theory, the performance analysis of spectral methods on dense realistic graph models and on high dimensional kernel random matrices is performed through the study of the eigenvalues and eigenvectors of the similarity matrices characterizing those data. New improved methods are proposed and are shown to outperform state-of-the-art approaches. In a second part, a new algorithm is proposed for the detection of heterogeneous communities from multi-layer graphs using variational Bayes approaches to approximate the posterior distribution of the sought variables. The proposed methods are successfully applied to synthetic benchmarks as well as real-world datasets and are shown to outperform standard approaches to clustering in those specific contexts.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France