# Application of Random Matrix Theory to High Dimensional Statistics

Joël Bun

Numéro NNT : 2016SACLS245

# THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
L'UNIVERSITE PARIS-SUD

ECOLE DOCTORALE N°564
Physique en Île de France

Spécialité de doctorat : Physique

Par

**M. Joël Bun**

## Application de la théorie des matrices aléatoires pour les statistiques en grande dimension

**Thèse présentée et soutenue à Orsay, le 6 Septembre 2016.**

**Composition du Jury :**

| | | |
|---|---|---|
| M. Hilhorst, Hendrik-Jan | Professeur, Paris-Saclay | Président du jury |
| M. Guhr, Thomas | Professeur, Universität Duisburg-Essen | Rapporteur |
| Mme. Péché, Sandrine | Professeur, Université Paris Diderot | Rapporteur |
| M., Bouchaud, Jean-Philippe | Chercheur, Capital Fund Management | Examinateur |
| M., Knowles, Antti | Maitre de Conférence, ETH Zürich | Examinateur |
| M. Majumdar, Satya | Directeur de recherche LPTMS | Directeur de thèse |
| M., Potters, Marc | Chercheur, Capital Fund Management | Co-encadrant de thèse |

# Résumé

Cette thèse porte sur l'étude statistique des systèmes en grande dimension grâce à la théorie des grandes matrices aléatoires. De nos jours, il est de plus en plus fréquent de travailler sur des bases de données de très grandes tailles dans plein de domaines différents. Cela ouvre la voie à de nouvelles possibilités d'exploitation ou d'exploration de l'information, et de nombreuses technologies numériques ont été crées récemment dans cette optique. D'un point de vue théorique, ce problème nous contraint à revoir notre manière d'analyser et de comprendre les données enregistrées. En effet, dans cet univers communément appelé *Big Data*, un bon nombre de méthodes traditionnelles d'inférence statistique multivariée deviennent inadaptées. Le but de cette thèse est donc de mieux comprendre ce phénomène appelé *fléau (ou malédiction) de la dimension*, et ensuite de proposer différents outils statistiques exploitant explicitement la dimension du problème et permettant d'extraire des informations fiables des données.

La première partie de thèse porte sur l'estimation de grande matrices de covariance (ou corrélation). Cet objet intervient dans de très nombreux problèmes pratiques (physique, finance, biologie, machine learning etc...) et il s'agit du sujet central de cette thèse. Ce problème est un exemple classique de l'impact considérable du bruit de mesure dans les systèmes en grande dimension. Le point de départ de ce travail est l'estimateur empirique de Pearson, qui est certainement le plus utilisé en pratique. Nous proposons une analyse détaillée de l'impact du fléau de la dimension sur cet estimateur grâce à la théorie des matrices aléatoires. Nous présenterons également comment étendre ces résultats à des modèles plus généraux de matrices de covariance.

Dans la seconde partie, nous nous intéressons aux modèles dans lesquels le vrai signal à extraire est corrompu par un bruit additif indépendant du signal. Nous considérons d'abord le cas où le bruit provient de l'addition d'une matrice Gaussienne, symétrique. Ensuite, nous abordons un modèle plus général inspiré de la théorie des probabilités libres permettant d'étudier des processus plus complexes.

Pour tous ces modèles, nous étudions les statistiques des valeurs propres mais surtout des vecteurs propres. Nous verrons par exemple que les vecteurs propres "mesurés", donc bruités, conservent très peu d'information concernant les vecteurs propres d'origine, c'est-à-dire ceux du vrai "signal". Même si ce résultat parait décevant aux premiers abords, nous montrerons qu'il est possible d'extraire de l'information significative en utilisant le caractère universel de nos résultats. En particulier, cela nous permettra de construire des estimateurs optimaux, observables, universels et cohérents avec le régime de grande dimension.

**Mots clés**: Statistiques en grande dimension, Covariance, Matrices aléatoires, Estimation, Décomposition Spectrale, Résolvante, Transformée de Stieltjes, Probabilités libres, Méthode des répliques, Mouvement Brownien de Dyson, Théorie de Markowitz, Bootstrap

# Abstract

This thesis focuses on the statistical analysis of high-dimensional systems using Random Matrix Theory (RMT). Nowadays, it is easy to get a lot of quantitative or qualitative data in a lot of different fields. This access to new data brought new challenges about data processing and there are now many different numerical tools to exploit database that may be of the order of teraoctet. In a theoretical standpoint, this framework appeals for new or refined results to deal with this amount of data. Indeed, it appears that most results of classical multivariate statistics become inaccurate in this era of *Big Data*. The aim of this thesis is twofold: the first one is to understand theoretically this so-called *curse of dimensionality* that describe phenomena which arise in high-dimensional space. Then, we shall see how we can use these tools to build reliable estimators that are consistent with the dimension of the problem.

Even if we will tackle different statistical problems in the following, the main problem that we will focus on is the estimation of high dimensional covariance matrices. This subject is the main thread of the first part of this thesis and finds very important applications in many practical problems, be it in physics, finance, machine learning or biology. This problem turns out to be a standard example of the increasing impact of the measurement noise as the dimension grows. We shall start from Pearson sample estimator which is the simplest way to estimate covariances. We then analyze how the curse of dimensionality affects this estimator using RMT. We will also discuss about possible extensions to more general model of covariance matrices.

The second part of the thesis is dedicated to models where the true signal is corrupted by an independent additive noise. More precisely, we first consider the case of a real symmetric Gaussian noise. Then, we will tackle a more general additive model inspired from free probability theory. For each model, we shall study the statistics of the eigenvalues and especially the eigenvectors. In particular, we will highlight that the empirical and noisy eigenvectors turn out to be unreliable estimators of the true ones in the high dimensional regime. Nevertheless, we may infer meaningful information about the true signal. This will help us to construct estimators that are optimal, universal, observable and consistent with the high dimensional framework.

**Keywords**: High-dimensional statistics, Covariance, Random matrices, Estimation, Spectral Decomposition, Resolvent, Stieltjes transform, Free probabiity, Replica methods, Dyson's Brownian Motion, Markowitz portfolio theory, Bootstrap

# Remerciements

Je souhaiterais tout d'abord remercier Jean-Philippe Bouchaud, Satya Majumdar et Marc Potters qui m'ont fait l'honneur de m'encadrer au cours de cette thèse.

Je voudrais exprimer toute ma gratitude envers Satya Majumdar, sans qui ce projet n'aurait jamais vu le jour. J'admire sa grande pédagogie et son enthousiasme durant nos séances de travail, qui m'ont beaucoup apporté. Je le remercie également pour son ouverture d'esprit concernant le déroulement de ma thèse.

Je tiens également à remercier chaleureusement Jean-Philippe Bouchaud et Marc Potters pour leur bienveillance, pour leur disponibilité et pour le temps qu'ils m'ont accordé depuis mon stage de master à CFM. Cela a été un véritable plaisir de collaborer avec eux sur ce sujet passionnant et de pouvoir profiter de leur expertise. J'ai beaucoup progressé à leur côtés et je leur en serai toujours reconnaissant.

Durant cette période, j'ai eu le privilège de travailler avec Romain Allez et Antti Knowles. Je remercie tout particulièrement Antti Knowles de m'avoir gentiment accueilli au sein de l'ETH Zürich. Cette visite, ainsi que notre collaboration, a surement été un point culminant de ma thèse. Je suis donc très heureux qu'il participe à mon jury. Je suis par ailleurs également honoré de compter Henk Hilhorst parmi les membres de ce jury.

Je voudrais remercier Martino Grasselli et Cyril Grunspan de m'avoir donné l'opportunité d'enseigner au Pôle Léonard de Vinci. Cette expérience a été très instructive et j'espère avoir été à la hauteur de la confiance qu'ils m'ont accordée. Merci également à Lakshithe Wagalath pour les nombreuses discussions enrichissantes que nous avons eues. J'en profite pour exprimer toute ma reconnaissance envers Daniel Gabay qui m'a permis d'arriver là où je suis aujourd'hui.

Mes premiers pas en matrices aléatoires ont été guidés par Philippe Very et Vincent Marzoli. Je tiens donc à les remercier pour m'avoir fait découvrir ce sujet passionnant, mais aussi pour les nombreux échanges éclairants durant mon stage à Natixis. Je leur dédie donc cette thèse.

L'enseignement de Sandrine Péché à l'Université Paris 7 a été un pilier de mon apprentissage de la théorie des matrices aléatoires. Au-delà de ce cours, elle a toujours été à l'écoute de mes (nombreuses) questions à ce sujet et c'est pourquoi je suis particulièrement ravi qu'elle ait accepté d'être rapporteur pour ma thèse.

*It is also a great honor for me that Thomas Guhr have accepted to write a report on my work.*

J'ai passé la majeure partie de ma thèse dans les locaux de CFM et cela a été une expérience extrêmement instructive et agréable. Je voudrais remercier tous les stagiaires/thésards/post-docs avec qui j'ai passé beaucoup de temps. Merci à Pierre, Stephen, Iacopo, Nicolas et Marc de m'avoir introduit au "café-poulet" lors de mon arrivé à CFM. Merci aussi à Yanis, que j'ai eu la chance de côtoyer durant le cours de matrices aléatoires ainsi qu'à CFM, pour les nombreuses discussions revigorantes sur des sujets divers et variés. Au delà des stagiaires/thésards/post-docs de CFM, j'ai également eu l'opportunité d'interagir avec de nombreux employés, notamment

Adam, Alexios, Charles-Albert, Emmanuel, Guillaume, Jean-Yves, Lam, Raphaël et Rémy. Je les remercie pour leur grande disponibilité. Je salue également tous les doctorants du LPTMS, notamment Igor, Pierre et Yasar.

J'ai aussi une pensée particulière pour mes soutiens les plus anciens. Merci à Eliada, Pierre, Thomas, Rodolphe, Bonheur, Aloun et Tarik pour nos interminables discussions sur le football; à Khammy et William pour nos mémorables parties de basket-ball ; à Anis pour m'avoir prouvé qu'il était possible d'avoir une thèse encore plus compliquée à lire que la mienne. Un immense merci à César pour nos nombreux fous rires mais aussi nos nombreux désarrois dans le fameux RER A. Je n'oublie pas non plus Philippe et Tony avec qui j'ai partagé énormément de bons moments durant ma scolarité.

Mes derniers remerciements vont bien évidemment à mes parents, à ma soeur, à mon frère, Stéphane ainsi qu'à ma cousine Julie. Ils m'ont tous été importants durant ces trois années de thèse. Enfin, mes pensées les plus affectueuses vont à celle qui est à mes côtés depuis plusieurs années et qui m'a toujours soutenu durant cette thèse. Merci pour tout, Claire.

# Contents

# Chapter 1

# Introduction générale et principales contributions

Cette thèse porte sur l'analyse statistique des systèmes complexes en grandes dimensions. La première partie est dédiée à l'estimation de grandes matrices de covariance et il s'agit du principal problème étudié dans ce manuscrit. La seconde partie concerne l'extension de ces résultats à des modèles de perturbations additifs.

Malgré cette distinction entre ces deux classes de modèles, il s'avère que le problème étudié est très similaire. Supposons que nous souhaitons estimer une matrice $\mathbf{C}$ de taille $N \times N$, déterministe et qui caractérise les dépendances entre les $N$ variables du système. En inférence statistique, la méthode traditionnelle pour estimer cette matrice $\mathbf{C}$ est de collecter un très grand nombre $T$ d'observations, idéalement indépendantes, afin de construire un estimateur *empirique* noté $\mathbf{E}$. Dans le cas où $T$ est très grand devant $N$, alors les estimateurs empiriques sont fiables dans le sens où l'erreur d'estimation devient très faible: on parle d'estimateur consistant [179]. De plus, nous voyons notamment que cette théorie ne suppose aucune structure particulière sur la matrice observée $\mathbf{E}$.

La question à laquelle nous allons tenter de répondre dans cette thèse est la suivante: est-ce que les estimateurs restent consistants lorsque le nombre de variables $N$ est du même ordre de grandeur que le nombre d'observations $T$? En d'autres termes, que pouvons-nous dire sur l'impact du bruit de mesure $q := N/T$ concernant la convergence des estimateurs empiriques? Ce cadre mathématique, connu sous le nom de *Big Data* de nos jours, est aujourd'hui fondamental d'un point de vue pratique pour de nombreuses disciplines scientifiques. Nous reviendrons sur les potentielles applications par la suite.

Cette question peut être étudiée en s'intéressant aux valeurs propres. En effet, considérons l'exemple des matrices de covariance: lorsque $N \ll T$, alors nous savons depuis les travaux d'Anderson que les valeurs propres observées, c'est-à-dire celles de $\mathbf{E}$, sont des estimateurs consistants des valeurs propres de $\mathbf{C}$ [10]. Par contre, lorsque $N$ est comparable à $T$, alors l'article fondateur de Marčenko et Pastur démontre que cette propriété n'est plus vérifiée [123]. Il s'agit d'un des résultats majeurs de la théorie des grandes matrices aléatoires dans la compréhension statistique de système de très grande dimension et le but de cette thèse est de comprendre un peu mieux ce fléau de la dimension grâce à cette théorie.

Dans ce mémoire, nous allons aborder ce problème à travers deux quantités distinctes. La première est l'analyse des vecteurs propres associés aux grandes matrices empiriques. Contrairement aux statistiques des valeurs propres qui disposent d'une littérature conséquente, le

comportement des vecteurs propres par rapport au bruit de mesure reste un problème relativement mal compris. Dans ce manuscrit, nous apportons des résultats permettant cette analyse en nous intéressant à l'espérance du produit scalaire entre les "vrais" vecteurs propres (ceux de $\mathbf{C}$) et les vecteurs propres bruités (ceux de $\mathbf{E}$).

Ensuite, la deuxième question que nous allons investiguer est la suivante: comment construire un estimateur qui soit consistant dans ce nouveau paradigme "big data"? Nous verrons en particulier que cet estimateur repose à la fois sur les statistiques des valeurs propres et des vecteurs propres de la matrice $\mathbf{E}$.

Comme indiqué dans le nom de ce mémoire, notre analyse repose sur la théorie des grandes matrices aléatoires qui s'avère extrêmement puissante pour caractériser les phénomènes se produisant sur des ensembles de très grande dimension (voir par exemple [1]). Plus précisément, depuis les travaux de Wigner sur les matrices symétriques Gaussiennes [189] et ensuite de Marčenko et Pastur en 1967 sur les matrices de covariance [123], cette théorie est à l'origine de nombreuses découvertes importantes en mécanique quantique, en physique statistique des systèmes désordonnés mais aussi dans les statistiques en grande dimension. Une des propriétés majeures de cette théorie est la possibilité d'exhiber des comportements universels à propos du spectre de grandes matrices. De plus, il existe une multitude de techniques analytiques permettant d'extraire ces informations et nous allons en présenter brièvement quelques une au cours de cette thèse dans cette introduction. Une description plus détaillée de chaque méthode se trouve dans les chapitres suivants.

Le but de cette introduction générale est de proposer aux lecteurs un résumé détaillé des différents travaux considérés durant ma thèse. Tout d'abord, nous allons motiver le problème statistique mentionné précédemment à travers quelques problèmes concrets. Comme indiqué ci-dessus, la motivation principale provient des statistiques en grande dimension et cela nous donnera l'occasion de récapituler les résultats existants dans la littérature. Ensuite, nous présenterons de manière succincte quelques méthodes de calculs utilisées: la théorie des probabilités libres, le mouvement Brownien de Dyson et la méthode des répliques. Cette présentation sera surtout l'occasion de fixer les notations et nous terminerons cette introduction par un bref énoncé des résultats obtenus et de quelques problèmes ouverts qui en découlent.

## 1.1 Estimation de grandes matrices de covariance

**1.1.1. Motivations.** La nouvelle ère liée au "Big Data" nous impose de reconsidérer les outils statistiques pour analyser les données de très grande dimension. Il s'agit de nos jours d'un problème récurrent dans quasiment toutes les disciplines scientifiques: physique, traitement d'image, génomique, épidémiologie, ingénierie, économie ou finance par exemple.

Une approche naturelle pour considérer des problèmes en grande dimension est d'identifier des *facteurs* communs qui expliquent la dynamique jointe des $N$ variables. Ces variables peuvent être le rendement journalier de différents stocks (le S&P 500 par exemple), la variation de températures dans différents endroits de la planète, la vitesse de particules dans un support granuleux ou différents indicateurs biologiques dans une population donnée (pression sanguine, cholestérol, etc.). L'objet mathématique le plus couramment utilisé pour quantifier les similarités entre différents observables est la matrice de corrélation, notée $\mathbf{C}$, qui est donc de taille $N \times N$. En effet, ses valeurs propres et ses vecteurs propres peuvent être utilisés pour caractériser les modes propres les plus importants, définis comme les combinaisons linéaires des variables originales ayant la plus grande variance (ou le plus grand pouvoir prédictif). Cette méthode est connue

sous le nom d'Analyse en Composante Principale (ACP, ou PCA en anglais).

Formellement, notons par $\mathbf{y} \in \mathbb{R}^N$ l'ensemble des variables que nous supposerons centrées, réduites[1] et présentant une interdépendance possiblement non triviale. Alors, pour mesurer l'intéraction entre ces variables, l'approche classique consiste à calculer les corrélations définies par:

$$\mathbf{C}_{ij} = \mathbb{E}\big[y_i y_j\big], \qquad i,j \in [\![1, N]\!]\,, \tag{1.1.1}$$

où l'espérance est prise sur la probabilité jointe caractérisant ces $N$ variables. Dans la suite, la matrice $\mathbf{C}$ sera appelée la vraie matrice de corrélation[2].

Le principal problème dans l'équation (2.1.1) est que l'espérance ne peut quasiment jamais être calculée explicitement du fait que la probabilité jointe des $N$ variables est souvent inconnue. Une façon pour remédier à ce problème est de collecter un grand nombre de réalisations *indépendantes* de ces variables pour former une matrice des données notée $\mathbf{Y} := (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T) \in \mathbb{R}^{N\times T}$. Dans le cas où le nombre de réalisations $T$ est suffisamment grand par rapport à $N$, une solution simple (mais naïve) est de construire l'estimateur empirique $\mathbf{E}$ de $\mathbf{C}$, définit par:

$$E_{ij} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}\, Y_{jt} := \frac{1}{T}\left(\mathbf{Y}\mathbf{Y}^*\right)_{ij}, \tag{1.1.2}$$

où $Y_{it}$ est la $t-$ème réalisation ($t = 1, \ldots, T$) de la $i-$ème variable ($i = 1, \ldots, N$ ) et nous supposons que les entrées de cette matrice sont centrées et réduites. Cet estimateur est connu sous le nom de l'estimateur de Pearson dans la littérature. Afin d'illustrer nos propos, nous pouvons voir la matrice $\mathbf{Y}$ comme la matrice des rendements de $N$ actifs financiers observés sur $T$ jours de trading. En biologie, $T$ peut être vu comme la taille de l'échantillon de population considérée pour la mesure de différents indicateurs de santé.

Lorsque $N \ll T$, alors nous pouvons utiliser les résultats classiques de statistiques multivariées pour établir que $\mathbf{E}$ converge (presque sûrement) vers $\mathbf{C}$ [179]. Par contre, lorsque $N$ devient grand, nous devons estimer simultanément $N(N-1)/2$ éléments de $\mathbf{C}$ en utilisant $NT$ observations, et nous voyons que cela devient problématique lorsque $T$ n'est pas très grand devant $N$. Bien que ce raisonnement est plus une heuristique qu'une véritable preuve, nous pouvons conclure qu'il est primordial de distinguer la vraie matrice $\mathbf{C}$ de son estimateur empirique $\mathbf{E}$ lorsque l'on travaille avec des objets de très grande dimension. Un des objectifs de cette thèse est justement de caractériser cette différence entre $\mathbf{E}$ et $\mathbf{C}$ dans ce régime où $N$ et $T$ deviennent très grands mais avec un ratio de dimension $q := N/T$ qui n'est pas arbitrairement petit. Nous discuterons également d'une méthode permettant de reconstruire du mieux possible la matrice $\mathbf{C}$ à partir de $\mathbf{E}$ tout en encodant le fait qu'il s'agit d'un estimateur bruité de $\mathbf{C}$.

L'estimation de matrice de covariance/corrélation est un point crucial dans de nombreux problèmes en statistiques. Voici une liste non-exhaustive de problèmes classiques liés à cet objet:

(i) Moindres carrés généralisés (MCG): Dans ce problème, nous essayons de décrire $\mathbf{y}$ en utilisant un modèle linéaire de la forme:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1.3}$$

où $\mathbf{X}$ est la matrice des régresseurs (ou facteurs) de taille $N \times k$ ($k \geqslant 1$), $\boldsymbol{\beta}$ symbolise les coefficients de régression à ces $k$ facteurs, et $\boldsymbol{\varepsilon}$ est le vecteur des résidus. Typiquement,

[1] Cette hypothèse à priori inoffensive est discutée dans le chapitre 4.

[2] Nous utiliserons plutôt la terminologie des corrélations plutôt que des covariances durant cette thèse.

nous cherchons le vecteur $\boldsymbol{\beta}$ qui explique au mieux les données et c'est exactement ce que nous permet les MCG. En effet, supposons que $\mathbf{E}[\boldsymbol{\varepsilon}|X] = 0$ et $\mathbb{V}[\boldsymbol{\varepsilon}|X] = \mathbf{C}$ qui est la matrice de covariance des résidus. Alors, estimer $\boldsymbol{\beta}$ par MCG nous donne la solution suivante: (voir [7] pour plus de précision):

$$\widehat{\boldsymbol{\beta}} = (X^* \mathbf{C} X)^{-1} X^* \mathbf{C}^{-1} \mathbf{y}. \tag{1.1.4}$$

(ii) Méthode des moments généralisée (MMG): Supposons que l'on souhaite estimer un ensemble de paramètres $\Theta$ d'un modèle fixé sur les données observées. Le principe de la méthode est de comparer les $k$ moments théoriques (qui dépendent de $\Theta$) avec leurs contreparties empiriques. Pour une estimation parfaite, ces $k$ différences sont donc toutes égales à zéro. La distance par rapport à zéro dépend des covariances entre ces $k$ fonctions et donc une bonne estimation de cette matrice de covariance augmente l'efficacité de la méthode (voir [88] pour plus de détails). Il s'avère que MCG est un cas spécial de MMG.

(iii) Analyse discriminante linéaire (ADL) [79]: Supposons que nous cherchons à classifier les variables $\mathbf{y}$ entre deux populations Gaussiennes de moyennes différentes $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$, à priori $\pi_1$ et $\pi_2$, mais de même matrice de covariance $\mathbf{C}$. Alors, l'ADL classifie $\mathbf{y}$ dans la classe 2 si:

$$\mathbf{x}^* \mathbf{C}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \frac{1}{2} (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^* \mathbf{C}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \log(\pi_2/\pi_1)\,. \tag{1.1.5}$$

(iv) Optimisation de portefeuille en présence d'un grand nombre d'actifs [125]: Nous souhaitons investir dans un ensemble d'actifs financiers $\mathbf{y}$ de façon à minimiser le risque moyen du portefeuille pour une performance future espérée $\nu > 0$. Si nous suivons la théorie de Markowitz, le portefeuille optimal est caractérisé par le vecteur de pondérations $\mathbf{w} := (w_1, \ldots, w_N)^*$ qui est solution d'un problème d'optimisation quadratique où l'on minimise la variance de la stratégie $\langle \mathbf{w}\,, \mathbf{C}\mathbf{w} \rangle$ sous contrainte d'un rendement futur moyen $\langle \mathbf{w}\,, \mathbf{g} \rangle \geqslant \mu$, avec $\mathbf{g}$ un vecteur de prédicteurs[3]. La stratégie optimale est donnée par:

$$\mathbf{w} = \nu \frac{\mathbf{C}^{-1}\mathbf{g}}{\mathbf{g}^* \mathbf{C}^{-1} \mathbf{g}}. \tag{1.1.6}$$

Nous insistons sur le fait qu'il s'agit ici d'une liste non-exhaustive de problèmes où la matrice de covariance joue un rôle prépondérant, d'autres problèmes pratiques sont mentionnés dans la revue de Paul & Aue [145], la thèse de Bartz [16] ou le livre de Couillet et Debbah [57].

Nous verrons dans les chapitres qui suivent que l'utilisation de l'estimateur empirique $\mathbf{E}$ peut mener à des performances futures (ou réalisées) désastreuses. Plus précisément, il est possible de montrer, sous certaines hypothèses techniques, que le risque réalisé dans des problèmes semblables à (i) et (iv) est donné par $\operatorname{Tr} \mathbf{E}^{-1}\mathbf{C}\mathbf{E}^{-1}$. Or cette quantité est un bon estimateur de l'optimum $\operatorname{Tr} \mathbf{C}^{-1}$ uniquement dans la limite $q \to 0$. Dans le cadre mathématique qui nous intéresse, c'est-à-dire $N$ et $T$ sont comparables, la théorie des matrices aléatoires nous dit que:

$$\operatorname{Tr} \mathbf{E}^{-1}\mathbf{C}\mathbf{E}^{-1} = \frac{\operatorname{Tr} \mathbf{C}^{-1}}{1-q}, \qquad q < 1\,, \tag{1.1.7}$$

pour une grande variété de processus. En d'autres termes, le risque réalisé, qui est celui qui nous intéresse en pratique, s'éloigne considérablement du risque optimal pour tout $q > 0$ et peut

[3] D'autres contraintes peuvent être intégrés à ce problème

même diverger pour $q \to 1$. Il est également important de préciser que ces résultats sont vrais pour d'autres mesures de risque [49,53]. En finance, une configuration typique dans les marchés actions est donnée par $N = 500$ et $T = 2500$, correspondant à dix ans de données journalières, ce qui est déjà suffisamment grand comparé à la durée de vie d'un stock en particulier. Dans ce cas, nous avons $q = 0.2$ et nous pouvons alors conclure d'après (1.1.7) que le risque réalisé sera 1.25 fois plus grand que le risque optimal bien que la taille des échantillons soit très grande. Si l'on s'intéresse aux indicateurs macroéconomiques tels que l'inflation par exemple, 20 ans de données mensuelles produisent uniquement $T = 240$ observations, tandis que le nombre de secteurs d'activités pour lesquels nous sauvegardons l'inflation est autour de $N = 30$, ce qui donne $q = 0.125$. Nous voyons ainsi que la compréhension de l'effet induit par un ratio $q$ supérieur à zéro est fondamental dans de nombreuses applications concrètes.

**1.1.2. État de l'art.** L'estimation de matrices de covariance est un problème classique en statistique multivariée. Un des résultats les plus influents à ce sujet remonte à l'année 1928 lorsque John Wishart étudia la distribution de la matrice de covariance empirique $\mathbf{E}$ dans le cas où la matrice des données $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ est caractérisée par une loi Gaussienne multivariée avec observations indépendantes et identiquement distribuées (i.i.d) [192]. Plus précisément, Wishart obtint une expression explicite pour la distribution de $\mathbf{E}$ sachant $\mathbf{C}$:

$$\mathcal{P}_W(\mathbf{E}|\mathbf{C}) = \frac{T^{NT/2}}{2^{NT/2}\Gamma_N(T/2)} \frac{\det(\mathbf{E})^{\frac{T-N-1}{2}}}{\det(\mathbf{C})^{T/2}} e^{-\frac{T}{2}\mathrm{Tr}\mathbf{C}^{-1}\mathbf{E}}, \tag{1.1.8}$$

où $\Gamma_N(\cdot)$ est la fonction Gamma multivariée de paramètre $N$[4]. Une propriété important est $\mathbb{E}[\mathbf{E}] = \mathbf{C}$, c'est-à-dire que l'estimateur est non biaisé [192]. En statistique, on dit que $\mathbf{E}$ suit une distribution Wishart$(N, T, \mathbf{C}/T)$. Au delà de la distribution de $\mathbf{E}$, il existe également une formule explicite pour la densité de probabilité marginale des valeurs propres pour tout $N$ and $T$ bornés [9]:

$$\rho_N(\lambda) = \frac{1}{N} \sum_{k=0}^{N-1} \frac{k!}{T-N+k} \left[L_k^{T-N}(\lambda)\right]^2 \lambda^{T-N} e^{-\lambda}, \tag{1.1.9}$$

sous l'hypothèse $T > N$ et avec les $L_k^l$ qui dénotent les polynômes de Laguerre [5].

Cette découverte de Wishart est souvent mentionnée comme un des premiers résultats de la théorie des matrices aléatoires [65]. Néanmoins, cela ne répond pas entièrement au problème qui nous intéresse, à savoir le comportement de la matrice empirique $\mathbf{E}$ en fonction du nombre de variables $N$ qui peut être arbitrairement grand. Cela a été compris bien plus tard avec le travail précurseur de Charles Stein en 1956 [96,165]. La contribution principale de Stein (et James) peut être résumée ainsi: quand le nombre de variables $N \geqslant 3$, alors il existe un meilleur estimateur que $\mathbf{E}$ en terme d'erreur quadratique moyen (voir l'article d'Efron [72] pour une présentation complète). Cet estimateur a la propriété de dépendre d'une information extérieure[6] et cela a donné naissance au paradoxe de Stein: l'estimateur empirique devient de moins en moins précis lorsque la dimension du système $N$ augmente. Ce résultat va en fait bien au-delà des matrices de covariance: supposons que nous voulons estimer la moyenne d'un vecteur Gaussien, alors l'estimateur empirique classique (équivalent ici à l'estimateur de maximum de vraisemblance) est moins précis que l'estimateur biaisé ("shrinkage") de James-Stein en terme d'erreur quadratique

[4] $\Gamma_N(u) = \pi^{N(N-1)/4} \prod_{j=1}^{N} \Gamma(u + (1-j)/2)$.
[5] $L_k^l(\lambda) = \frac{e^\lambda}{k!\lambda^l} \frac{d^k}{d\lambda^k}(e^{-\lambda}\lambda^{k+l})$.
[6] Nous reviendrons sur cette caractérisation dans le chapitre 6.

dès lors que $N \geqslant 3$ [96]. En ce qui concerne les matrices de covariance, le paradoxe de Stein peut être observé de manière précise pour tout $N \geqslant 3$ en utilisant les propriétés de la distribution de Wishart (3.1.38) ainsi que la notion de *famille conjuguée* de la théorie Bayésienne (voir le Chapitre 6). Plus précisément, ce paradoxe a été démontré en premier lieu pour l'estimation de la matrice de "précision" $\mathbf{C}^{-1}$ dans les articles [71, 85] et ensuite pour la matrice de covariance $\mathbf{C}$ dans [87]. C'est d'ailleurs dans l'article de Haff qu'apparaît pour la première fois le célèbre estimateur de shrinkage linéaire:

$$\mathbf{\Xi} = (1 - \alpha_s)\mathbf{E} + \alpha_s \mathbf{I}_N, \tag{1.1.10}$$

où $\Xi$ représentera pour toute la suite un estimateur quelconque de $\mathbf{C}$ et $\alpha_s \in (0, 1)$ est appelé l'intensité de shrinkage. Nous voyons dans l'équation (2.1.9) que le shrinkage linéaire interpole entre la matrice empirique "bruitée" $\mathbf{E}$ (pas de shrinkage, $\alpha_s = 0$) et l'hypothèse nulle $\mathbf{I}_N$ (shrinkage extrême, $\alpha_s = 1$). Il est facile de voir que cet exemple illustre parfaitement l'idée qu'un estimateur qui utilise de l'information supplémentaire offre de meilleures performances lorsque la dimension du système grandit. Dans [87], Haff propose d'estimer l'intensité $\alpha_s$ en utilisant la distribution de probabilité marginale des observations $\mathbf{Y}$. L'amélioration offerte par cet estimateur par rapport à $\mathbf{E}$ lorsque $N \to \infty$ a été quantifiée précisément bien plus tard en 2004 avec l'article de Ledoit et Wolf [115].

Il est intéressant de noter que le premier résultat sur le comportement de $\mathbf{E}$ dans la limite des grandes dimensions ne provient pas de la communauté des statistiques, mais plutôt des mathématiques avec l'article de Marčenko and Pastur en 1967 [123]. Dans cet article, les deux auteurs obtiennent une équation auto-cohérente (appelée équation de Marčenko-Pastur) pour la densité des valeurs propres de $\mathbf{E}$ sachant $\mathbf{C}$ pour $N, T \to \infty$ avec un ratio $q = N/T$ possiblement d'ordre 1. D'ailleurs, c'est précisément grâce à ce résultat que nous pouvons étudier en détail l'influence du paramètre $q$ dans la qualité d'estimation de $\mathbf{E}$ par rapport à $\mathbf{C}$. Par exemple, il est possible de montrer que pour $q \to 0$, nous retrouvons alors la convergence "classique" des valeurs propres de $\mathbf{E}$ vers celles de $\mathbf{C}$ démontrée par Anderson [10]. Par contre, pour tout $q > 0$, le résultat de Marčenko et Pastur illustre parfaitement le fait que les valeurs propres empiriques (celles de $\mathbf{E}$) sont des estimateurs bruités des vraies valeurs propres (celles de $\mathbf{C}$) peu importe la valeur $T$: ce phénomène caractérise le fléau de la dimension. Afin d'illustrer nos propos sur l'influence du ratio $q$ par rapport à la qualité d'estimation de $\mathbf{E}$, prenons l'exemple simple où $\mathbf{C}$ est la matrice identité. Dans ce cas, il est clair que la densité des vraies valeurs propres est donnée par une masse de Dirac en 1. Or, si nous calculons la densité des valeurs propres de $\mathbf{E}$ pour un $q > 0$, nous observons dans la Figure 2.1.1 que le spectre de $\mathbf{E}$ pour $q = 0.25$ s'écarte significativement de la masse de Dirac en 1 ($q = 0$). Cet effet est encore plus exacerbé lorsque nous prenons une plus grande valeur de $q = 0.5$.

Le résultat de Marčenko et Pastur (MP par la suite) a eu un impact considérable ces deux dernières décennies dans notre compréhension du fléau de la dimension mais également dans le développement de la théorie des matrices aléatoires. En effet, cela a ouvert la voie à la conception de nouveaux objets mathématiques remettant en cause une bonne partie des résultats classiques des statistiques multivariées. Tout d'abord, il a été compris en 1986 puis en 1995 que le résultat de MP est universel dans la limite des grandes dimensions, c'est-à-dire qu'il est valide pour une large classe de processus stochastiques et pour une vraie matrice $\mathbf{C}$ quelconque [157, 161, 195]. La notion d'universalité est un des attraits majeurs de la théorie des matrices aléatoires d'un point de vu théorique. Dans le même temps, des études empiriques ont illustré la pertinence de ces résultats sur les données financières [110, 151] qui sont réputées pour être fortement non

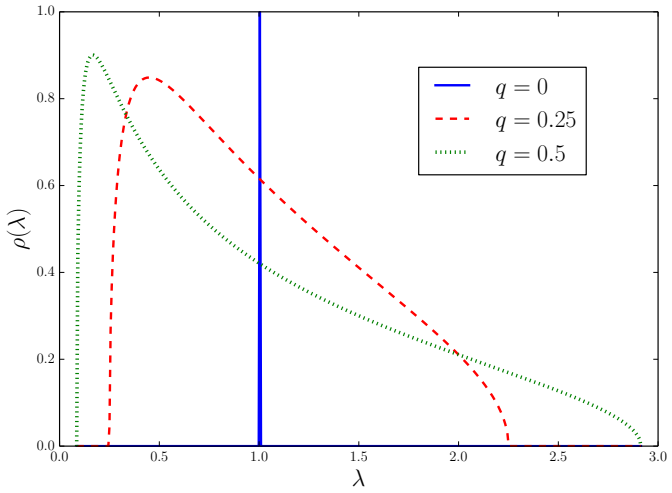



FIGURE 1.1.1. Illustration de la densité des valeurs propres de $\mathbf{E}$ issues de l'équation de Marčenko et Pastur avec $\mathbf{C} = \mathbf{I}_N$ et $N = 500$. La ligne bleue ($q = 0$) correspond à une estimation parfaite des valeurs propres de $\mathbf{C}$. Plus le ratio d'observation $q$ est élevé, plus le spectre des valeurs propres de $\mathbf{E}$ est large. Par exemple, pour $T = 4N$ (courbe en rouge), l'écart par rapport aux vraies valeurs propres est déjà considérable.

Gaussiens [30]. Ces travaux suggèrent que le "bulk" des valeurs propres de $\mathbf{E}$ pour des données financières coïncident avec les prédictions théoriques de MP pour $\mathbf{C} = \mathbf{I}_N$, tandis qu'un nombre fini de valeurs propres isolées ("outliers" ou "spikes" en anglais) se situent à une distance significative du bulk. Cette observation est à l'origine du modèle "spiked covarance matrix" de Iain Johnstone [100] et possède de nombreuses applications pour l'ACP. En particulier, Johnstone démontra un autre caractère universel de cette théorie: la plus grande valeur propre du bulk est régit par la distribution de Tracy & Widom [99, 175]. En outre, cela illustre que le bord de spectre des valeurs propres est *rigide* dans le sens où leurs positions fluctuent très peu, de l'ordre de $T^{-3/2}$. Cette observation permet alors de construire une méthode simple pour "nettoyer" les valeurs propres de $\mathbf{E}$: toute valeur propre se situant à une distance significative (par rapport $T^{-3/2}$) du bulk peut être interprétée comme contenant un signal non trivial. Autrement dit, toutes celles qui se situent dans la prédiction théorique de MP proviennent uniquement du bruit de mesure et ne peuvent donc pas être considérées comme fiables [111, 151]. Cette méthode est appelée "clipping" des valeurs propres et offre des performances bien supérieures en terme de prédiction des risques financiers comparée à la matrice empirique [29]. En conclusion, cela démontre que la notion de nettoyage (ou de régularisation) est primordiale lorsque l'on manipule des données de très grande dimension.

Bien que ce modèle permet d'améliorer significativement l'estimation de la matrice $\mathbf{C}$ par rapport à l'estimateur empirique $\mathbf{E}$, nous pouvons nous demander si une telle hypothèse de modélisation des vraies corrélations (ou covariances) serait réaliste en pratique. Étant donné que le résultat de Marčenko et Pastur nous permet de travailler avec n'importe quelle vraie matrice $\mathbf{C}$, il est donc tentant de vouloir reconstruire la distribution des valeurs propres de $\mathbf{C}$ à partir de la distribution observée des valeurs propres de $\mathbf{E}$. Malheureusement, ce problème est particulièrement complexe, car l'équation de Marčenko-Pastur est numériquement stable dans l'autre sens: connaissant le spectre de $\mathbf{C}$, on peut calculer le spectre de $\mathbf{E}$. Par conséquent, de

nombreuses propositions pour "inverser" cette équation sont apparues dans la littérature à partir de l'année 2008 [29, 104, 133, 194]. La méthode de [29] propose de paramétriser la densité des vraies valeurs propres et d'estimer les paramètres sur les valeurs propres observées. L'avantage d'une telle méthode est qu'elle est très efficace d'un point de vue numérique mais elle suppose que nous avons un à priori cohérent sur la densité des vraies valeurs propres. Il est donc clair qu'une telle méthode souffre de son manque de généralité et la méthode d'El Karoui est, en ce sens, plus robuste [104]. En effet, l'hypothèse de base est de dire que la densité des valeurs propres peut être vue comme une somme discrète de masses de Dirac et le but est de trouver les poids associés à ces masses. Cette méthode offre donc un plus grand degré de liberté que la méthode paramétrique mais elle est extrêmement dépendante de la position des masses de Dirac, ce qui pose problème en pratique. La méthode de Mestre est complètement différente et présuppose que le spectre de $\mathbf{C}$ est composé d'un nombre fini $n \ll N$ de valeurs propres distinctes [133]. Sous cette hypothèse, Xavier Mestre propose une formule analytique pour estimer ces $n$ valeurs propres en utilisant celles de $\mathbf{E}$. Cependant nous voyons que l'entrée cruciale de cette méthode est le nombre $n$ qui n'est pas connu à priori. Ce travail a ensuite été amélioré dans l'article de Yao et collaborateurs [194] mais souffre encore du même problème que le travail de Mestre.

Il existe donc *en principe* des méthodes nous permettant de reconstruire (partiellement) le spectre de $\mathbf{C}$ mais cela est toujours insuffisant en ce qui concerne l'estimation de la matrice $\mathbf{C}$ entière. En effet, en appliquant cette procédure, appelée *substitution*, nous supposons implicitement que les vecteurs propres de $\mathbf{E}$ sont des bons estimateurs des vrais vecteurs propres de $\mathbf{C}$. Mais est-ce vraiment le cas? Cette question soulève une des limites de l'équation de Marčenko et Pastur car elle ne donne pas d'information sur l'influence du paramètre $q$ sur les vecteurs propres de $\mathbf{E}$. Il y a relativement peu de résultats sur les vecteurs propres comparés aux valeurs propres. Ceci peut s'expliquer par le fait que ce problème est bien plus difficile que l'estimation des valeurs propres car la dimension du problème est de taille $N \times N$. Nous pouvons néanmoins citer par exemple les travaux de Jack Silverstein dans les années 1980 [159, 160] et il faudra ensuite attendre l'année 2007 pour voir deux études à ce sujet [12, 144]. En ce qui concerne les applications en statistiques, l'article de Debashis Paul est extrêmement important car il permet d'avoir des résultats explicites sur les vecteurs propres associés aux valeurs propres isolées, c'est-à-dire celles qui expliquent le plus de variance dans les données. La conclusion de son travail est la suivante: les vecteurs propres associés aux valeurs propres isolées ne sont pas de bons estimateurs dans la limite des grandes dimensions. Ce résultat est très important pour l'ACP et nous indique qu'il est important de mieux comprendre le comportement des vecteurs propres de $\mathbf{E}$ dans la limite $N \to \infty$. Depuis, les études sur les vecteurs ont été étendues à des modèles plus généraux [21, 40, 43, 113, 136] avec toujours la même conclusion. Nous pouvons donc en conclure qu'il est aussi important d'inclure ces informations sur les vecteurs propres lors de la construction d'un estimateur de $\mathbf{C}$, ce qui n'est pas le cas dans la méthode de substitution mentionnée dans le paragraphe précédent.

**1.1.3. Problème statistique.** Au vu de la section précédente, nous comprenons que l'estimateur empirique $\mathbf{E}$ n'est pas un estimateur consistant de $\mathbf{C}$ dans la limite des grandes dimensions. Ce résultat peut être exhibé aussi bien à travers les valeurs propres que les vecteurs propres grâce à la théorie des matrices aléatoires. Cependant, nous pouvons nous demander comment reconstruire – au mieux – la matrice $\mathbf{C}$ en utilisant ces observations. Pour les valeurs propres, nous avons vu qu'il existe différentes méthodes pour résoudre le problème inverse de Marčenko et Pastur et ainsi inférer le spectre de $\mathbf{C}$. Ceci est malheureusement beaucoup moins évident en ce qui concerne les vecteurs propres.

En conséquence, la classe des estimateurs *invariant par rotation* que nous allons maintenant introduire semble être un bon compromis. En effet, comme son nom l'indique, un estimateur invariant par rotation (RIE en anglais) présume qu'il n'y a aucune direction privilégiée vers laquelle les vecteurs propres de $\mathbf{C}$ doivent être dirigés. Cette hypothèse a deux conséquences importantes. Soit $\Xi(\mathbf{E})$ un estimateur de $\mathbf{C}$ qui est une fonctionnelle de $\mathbf{E}$, alors il est possible de montrer que cet estimateur possède les mêmes vecteurs propres que $\mathbf{E}$ [166]. Ensuite, et il s'agit de la différence principale, les valeurs propres de l'estimateur optimal (au sens de la norme $\mathbb{L}_2$) au sein de cette classe d'estimateurs intègre le fait que les vecteurs propres de $\mathbf{E}$ ne sont pas forcément des bons estimateurs de ceux $\mathbf{C}$. Cette classe d'estimateurs a été étudiée en détail tout d'abord dans [113] puis généralisée dans les articles [40, 43] que j'ai co-écrits avec Romain Allez, Jean-Philippe Bouchaud, Antti Knowles, et Marc Potters. L'objet de ce mémoire de thèse est donc d'expliquer la construction d'un estimateur optimal qui soit invariant par rotation.

Dans cette section, nous allons poser le problème statistique qui nous intéressera durant toute la suite. On définit par $\mathcal{M}_N \equiv \mathcal{M}$ l'ensemble des matrices réelles, symétriques, non-négatives et de dimension $N \times N$. Ensuite, on définit par $\mathcal{M}(\mathbf{E})$ l'ensemble des matrices appartenant à $\mathcal{M}$ et possédant les mêmes vecteurs propres que $\mathbf{E}$. L'estimateur optimal que nous cherchons est alors définit par

$$\tilde{\Xi} \;:=\; \underset{\Xi \in \mathcal{M}(\mathbf{E})}{\operatorname{argmin}} \big\| \Xi - \mathbf{C} \big\|^2, \tag{1.1.11}$$

où $\|\cdot\|^2$ dénote la norme $\mathbb{L}_2$. Si nous définissons par $[\mathbf{u}_i]_{i\in[\![1,N]\!]}$ les vecteurs propres de $\mathbf{E}$ , alors il est facile de montrer que la solution optimale est donnée par

$$\tilde{\Xi} = \sum_{i=1}^{N} \tilde{\xi}_i \mathbf{u}_i \mathbf{u}_i^*, \qquad \tilde{\xi}_i \;=\; \langle \mathbf{u}_i \,, \mathbf{C}\mathbf{u}_i \rangle. \tag{1.1.12}$$

D'un point de vue pratique, il est évident qu'un tel estimateur semble inutile car la solution dépend de la matrice $\mathbf{C}$ qui est précisément la matrice que nous souhaitons estimer. Dans la littérature, un tel estimateur est appelé *estimateur oracle* car il nécessite la connaissance de la quantité que nous cherchons à reconstruire et par conséquent, nous utiliserons l'exposant "ora." par la suite pour faire référence à cet estimateur. Néanmoins, son interprétation (1.1.12) est assez naturelle. En effet, si nous définissons par $[\mu_i]_{i\in[\![1,N]\!]}$ les valeurs propres de $\mathbf{C}$ et par $[\mathbf{v}_i]_{i\in[\![1,N]\!]}$ les vecteurs propres correspondants, nous pouvons alors ré-écrire l'équation (1.1.12) comme suit:

$$\tilde{\xi}_i \;\equiv\; \xi_i^{\text{ora.}} \;=\; \sum_{j=1}^{N} \mu_j \langle \mathbf{u}_i \,, \mathbf{v}_j \rangle^2. \tag{1.1.13}$$

Nous pouvons interpréter chaque valeur propre oracle $\xi^{\text{ora.}}$ comme étant la moyenne pondérée des valeurs propres de $\mathbf{C}$ où les poids sont définis par le produit scalaire (au carré) entre le $i-$ème vrai vecteur propre et le $j-$ème vecteur propre observé pour tout $j \in [\![1, N]\!]$. Ce produit scalaire peut être vu comme une probabilité de transition entre un état perturbé (les $\mathbf{u}_i$) vers un état non-perturbé (les $\mathbf{v}_i$) [150]. Ainsi, dans le cas d'une estimation parfaite des vecteurs propres, il est facile de voir que $\xi^{\text{ora.}} = \mu_i$ pour tout $i \in [\![1, N]\!]$. Mais dans le régime des grandes dimensions, nous nous attendons à obtenir une solution non triviale à condition d'être en mesure de calculer cette "probabilité de transition". Cela démontre pourquoi les statistiques des vecteurs propres sont cruciales dans ce problème et c'est pour cela qu'une grande partie de cette thèse y sera consacrée. Nous verrons en particulier que le produit scalaire $\big[\langle \mathbf{u}_i \,, \mathbf{v}_j \rangle^2\big]_{i,j=1}^{N}$ s'exprime en moyenne uniquement en fonction des valeurs propres de $\mathbf{E}$ et $\mathbf{C}$. Cela nous permettra ensuite

de montrer que l'estimateur (1.1.13) converge vers une fonction qui ne dépend que des valeurs propres observées, et c'est ce "miracle" qui intervient dans la limite des grandes dimensions qui rend cette théorie très utile pour en pratique.

## 1.2 Extension à d'autres modèles de matrices aléatoires

Jusqu'à présent, le modèle de matrice qui nous a intéressé est la matrice de covariance empirique (1.1.2) mais nous pouvons constater que le problème d'optimisation (1.1.11) est valable pour toute matrice $\mathbf{C}$ et surtout pour tout processus de mesure $\mathbf{E}$. Ainsi, il est possible de considérer le même problème mais avec des hypothèses de modélisation différentes concernant la matrice $\mathbf{E}$.

Une des hypothèses fondamentales de la matrice de covariance empirique (1.1.2) est que les $T$ observations soient i.i.d, c'est-à-dire il n'y a aucune dépendance temporelle dans les données. Cette hypothèse est rarement vérifiée en pratique et il est donc naturel de se demander s'il est possible d'étendre nos travaux à des modèles intégrant cette dépendance temporelle par exemple. Faisons l'hypothèse que $Y$ est une matrice Gaussienne avec une autocorrélation exponentielle. Alors, on peut écrire la variance sous la forme[7] [18, 47, 48]:

$$\mathbb{E}[Y_{it}Y_{jt'}] = C_{ij}\exp\big[-|t-t'|/\tau\big], \tag{1.2.1}$$

où $\tau$ contrôle la portée de la dépendance en temps. Un autre exemple classique est de mesurer les corrélations en utilisant une *moyenne mobile exponentielle* [47, 142]:[8]

$$M_{ij}(\tau, T) = (1-\alpha)\sum_{t=0}^{T}\alpha^t Y_{i,\tau-t}Y_{j,\tau-t}, \tag{1.2.2}$$

où $\tau$ est la dernière date d'estimation, $\alpha \in (0,1)$ est une constante et $T$, la taille de série temporelle. L'idée de cet estimateur est de dire que les anciennes données deviennent de progressivement obsolètes et doivent donc moins contribuer que les données récentes. Nous pouvons remarquer à partir de l'équation (10.1.2) que cet estimateur peut être ré-écrit de la façon suivante:

$$M_{ij}(\tau) \;=\; (1-\alpha)\sum_{t=0}^{T} H_{it}H_{jt}, \quad \text{with} \quad \mathbb{E}\big[H_{it}H_{it'}\big] = \delta_{tt'}(1-\alpha)\alpha^t, \tag{1.2.3}$$

où la variance des variables aléatoires a donc une dépendance temporelle explicite. Il existe bien évidemment d'autres exemples dont certains seront discutés dans le chapitre (10), mais nous pouvons constater que les modèles (10.1.1) ou (10.1.3) peuvent être regroupés au sein d'un même modèle de perturbation multiplicatif:

$$\mathbf{M} \;:=\; \mathbf{C}^{1/2}\mathbf{X}\mathbf{B}\mathbf{X}^*\mathbf{C}^{1/2}, \tag{1.2.4}$$

où $\mathbf{X} := (X_{it}) \in \mathbb{R}^{N\times T}$ est une matrice aléatoire dont les colonnes sont i.i.d de moyenne nulle et de variance $T^{-1}$ et $\mathbf{B} = (B_{tt'}) \in \mathbb{R}^{T\times T}$ est une matrice fixe et indépendante de $\mathbf{C}$. En effet, pour l'équation (10.1.1), nous avons $B_{tt'} = \exp[-|t-t'|/\tau]$ tandis que $B_{tt'} = \delta_{tt'}(1-\alpha)\alpha^t$ pour le modèle (10.1.3). Nous étudierons les modèles de la forme (1.2.4) dans les chapitres (3) et (7).

[7] Nous rappelons que nous supposons que les $T$ réalisations de $Y$ sont de moyennes nulles.

[8] Nous utilisons une lettre différente pour les estimateurs de cette section pour éviter toute confusion avec l'estimateur empirique $\mathbf{E} = \mathbf{X}\mathbf{X}^*/T$.

Au delà du modèle multiplicatif (1.2.4), on peut également considérer le cas où les observations sont corrompues par un bruit additif. Plus précisément, supposons que l'on mesure une matrice $\mathbf{M}$ de la forme

$$\mathbf{M} \;:=\; \mathbf{C} + \mathbf{B}, \tag{1.2.5}$$

où $\mathbf{B} = (B_{ij}) \in \mathbb{R}^{N\times N}$ est une matrice réelle, symétrique, invariante par rotation et indépendante de $\mathbf{C}$. L'exemple le plus connu d'un modèle de cette forme (1.2.5) est le cas où le bruit extérieur $\mathbf{B}$ est une matrice Gaussienne à entrées indépendantes (symétrie à part), qui est appelé le *Gaussian Orthogonal Ensemble* (GOE) dans la littérature. Nous pouvons encore nous demander s'il est toujours possible d'estimer l'estimateur oracle (1.1.12) par une fonction observable dans la limite des grandes dimensions. Comme pour les matrices de covariances, nous verrons que c'est toujours le cas, c'est-à-dire que les valeurs propres $\xi^{\text{ora.}}$, définies dans l'équation (1.1.13), convergent vers une valeur qui ne nécessite pas explicitement la connaissance de $\mathbf{C}$. Nous nous intéresserons ensuite au cas où la matrice $\mathbf{B}$ est une matrice invariante par rotation mais avec une distribution arbitraire. Les modèles additifs seront étudiés dans la partie II de ce mémoire et proviennent des articles [5,40] co-rédigés avec Romain Allez, Jean-Philippe Bouchaud et Marc Potters.

## 1.3 Méthodes utilisées

Dans cette section, nous allons passer en revue les différentes techniques de calculs employées durant cette thèse. Le but n'est pas de faire une présentation complète des méthodes mais plutôt de donner les intuitions et d'introduire des objets importants pour la suite de ce chapitre. Une description plus détaillée est donnée dans le chapitre 3 pour la majeure partie des méthodes excepté pour le mouvement Brownien de Dyson qui sera introduit dans le chapitre 12.

**1.3.1. Quelques définitions.** L'outil mathématique principal dans ce mémoire est la résolvante. Soit $\mathbf{M}$ une matrice de dimension $N \times N$ et on définit la résolvante par

$$\mathbf{G}_{\mathbf{M}}(z) \;:=\; \left(z\mathbf{I}_N - \mathbf{M}\right)^{-1}, \tag{1.3.1}$$

et si on dénote par $\lambda_1 \geqslant \lambda_2 \geqslant \dots \lambda_N$ les valeurs propres de $\mathbf{M}$ et par $\mathbf{u}_1, \dots, \mathbf{u}_N$ les vecteurs propres correspondants, nous pouvons déduire de l'équation (1.3.1) que

$$\mathbf{G}_{\mathbf{M}}(z) \;=\; \sum_{i=1}^{N} \frac{\mathbf{u}_i \mathbf{u}_i^*}{z - \lambda_i}, \tag{1.3.2}$$

où l'exposant $^*$ symbolise la transposée. Ainsi, nous pouvons remarquer que la résolvante possède $N$ pôles situés à chaque valeur propre $\lambda_i$ et dont le résidu est la projection sur l'espace propre associé à $\lambda_i$. En d'autres termes, chaque pôle de la résolvante nous permet donc d'avoir les informations sur les valeurs propres ainsi que les vecteurs propres. Lorsque l'on s'intéresse uniquement aux statistiques des valeurs propres, il est souvent plus simple de considérer la trace (normalisée) de la résolvante, communément appelée la transformée de Stieltjes empirique:

$$\mathfrak{g}_{\mathbf{M}}^{N}(z) \;:=\; \frac{1}{N} \operatorname{Tr} \mathbf{G}_{\mathbf{M}}(z) \;=\; \frac{1}{N} \sum_{i=1}^{N}. \tag{1.3.3}$$

Dans la limite des grandes dimensions, nous utiliserons souvent la convergence suivante:

$$\mathfrak{g}_{\mathbf{M}}^{N}(z) \;\sim\; \mathfrak{g}_{\mathbf{M}}(z) \;:=\; \int \frac{\rho(\lambda)}{z-\lambda}\mathrm{d}\lambda\,, \tag{1.3.4}$$

où $\rho(\lambda)$ est la densité des valeurs propres. La résolvante (1.3.1) et la transformée de Stieltjes (1.3.4) sont les deux quantités mathématiques fondamentales dans la théorie des matrices aléatoires. Nous allons présenter brièvement les différentes techniques nous permettant d'évaluer ces objets.

**1.3.2. Probabilités libres.** La théorie des probabilités libres, créée par Dan Voiculescu, permet d'étudier des variables aléatoires non-commutatives [181]. Une des applications les plus importantes de cette théorie concerne les matrices aléatoires, suite aux travaux du même Voiculescu [182]. En particulier, les probabilités libres offrent un cadre relativement simple pour comprendre la densité des valeurs propres résultant de l'addition ou la multiplication de grandes matrices aléatoires indépendantes. En nous remémorant les équations (1.2.4) et (1.2.5) qui nous intéressent, il est facile de comprendre pourquoi cette théorie nous est utile.

Dans le cas du modèle additif (1.2.5), nous pouvons utiliser la formule d'addition libre pour calculer la transformée de Stieltjes de $\mathbf{M}$ [182]:

$$\mathcal{R}_{\mathbf{M}}(\omega) \;=\; \mathcal{R}_{\mathbf{A}}(\omega) + \mathcal{R}_{\mathbf{B}}(\omega)\,, \qquad \mathfrak{g}_{\mathbf{M}}\left(\mathcal{R}_{\mathbf{M}}(\omega) + \frac{1}{\omega}\right) = \omega\,, \tag{1.3.5}$$

pour $\omega \in \mathbb{C}$. De la même façon, nous pouvons traiter le modèle multiplicatif (1.2.4) via la formule de multiplication libre [182]:

$$\mathcal{S}_{\mathbf{M}}(z) = \mathcal{S}_{\mathbf{A}}(z)\mathcal{S}_{\mathbf{B}}(z)\,, \qquad \mathcal{S}_{\mathbf{M}}(z) \;:=\; = \frac{z+1}{z\mathcal{T}_{\mathbf{M}}^{-1}(z)}\,, \tag{1.3.6}$$

où $\mathcal{T}_{\mathbf{M}}^{-1}$ est l'inverse fonctionnelle de $\mathcal{T}_{\mathbf{M}}(z) \;:=\; z\mathfrak{g}_{\mathbf{M}}(z) - 1$. Plus de détails sur cette méthode sont fournies dans la Section 3.1.3.

**1.3.3. Méthode des répliques.** La méthode des répliques est une technique issue de la physique statistique permettant d'étudier la valeur moyenne de systèmes complexes désordonnés en introduisant un nombre fini de répliques du système initial. Cela permet ainsi de pouvoir calculer la moyenne sur les différentes copies, ce qui simplifie souvent les calculs. Cette technique a rencontré beaucoup de succès dans différents contextes tels que la théorie des matrices aléatoires ou les systèmes désordonnés (voir [134] ou [137] pour une revue plus récente). Il est toutefois important de noter que cette technique est très puissante, mais ne repose pas sur des arguments mathématiques rigoureux. C'est pour cela qu'il est toujours conseillé de vérifier au moins numériquement les solutions obtenues par cette méthode.

Dans cette thèse, nous avons utilisé les répliques pour calculer le comportement asymptotique de la résolvante (1.3.1). Pour cela, partons de la représentation de l'inverse d'une matrice par les intégrales Gaussiennes:

$$(z\mathbf{I}_N - \mathbf{M})^{-1}_{ij} = \frac{\int \left(\prod_{k=1}^{N} d\eta_k\right) \eta_i\eta_j \exp\left\{-\frac{1}{2}\sum_{k,l=1}^{N}\eta_k(z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\}}{\int \left(\prod_{k=1}^{N} d\eta_k\right) \exp\left\{-\frac{1}{2}\sum_{k,l=1}^{N}\eta_k(z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\}}. \tag{1.3.7}$$

Dans la limite des grandes dimensions, nous pouvons supposer que l'équation (1.3.7) est auto-moyennante[9], et donc, nous avons:

$$G_{ij}(z) = \left\langle \frac{1}{\mathcal{Z}} \int \left( \prod_{k=1}^{N} d\eta_k \right) \eta_i \eta_j \exp\left\{ -\frac{1}{2} \sum_{k,l=1}^{N} \eta_k (z\delta_{kl} - \mathbf{M}_{kl}) \eta_l \right\} \right\rangle_{\mathcal{P}_{\mathbf{M}}}, \tag{1.3.8}$$

où $\mathcal{Z}$ est le dénominateur de l'équation (1.3.7). La méthode des répliques revient alors à faire la manipulation suivante:

$$\begin{aligned} G_{ij}(z) &= \lim_{n\to 0} \left\langle \mathcal{Z}^{n-1} \int \left( \prod_{k=1}^{N} d\eta_k \right) \eta_i \eta_j \exp\left\{ -\frac{1}{2} \sum_{k,l=1}^{N} \eta_k (z\delta_{kl} - \mathbf{M}_{kl}) \eta_l \right\} \right\rangle_{\mathcal{P}_{\mathbf{M}}} \\ &= \lim_{n\to 0} \int \left( \prod_{k=1}^{N} \prod_{\alpha=1}^{n} d\eta_k^\alpha \right) \eta_i^1 \eta_j^1 \left\langle \prod_{\alpha=1}^{n} \exp\left\{ -\frac{1}{2} \sum_{k,l=1}^{N} \eta_k^\alpha (z\delta_{kl} - \mathbf{M}_{kl}) \eta_l^\alpha \right\} \right\rangle_{\mathcal{P}_{\mathbf{M}}} . \end{aligned} \tag{1.3.9}$$

Ainsi, nous voyons que nous avons ré-écrit le problème initial (1.3.8) en une évaluation de $n$ répliques. Le calcul de la dernière ligne (1.3.9) est souvent plus simple, à condition d'être en mesure de calculer la valeur moyenne de chaque réplique. Nous verrons que ce calcul est en fait relié à l'étude du comportement asymptotique de l'intégrale d'Harish-Chandra–Itzykson-Zuber où l'une des matrices est de rang faible (voir Appendice A pour une étude détaillée de cette intégrale). Nous verrons que l'identité (1.3.9) sera très utile pour étudier chaque entrée de la résolvante de $\mathbf{M}$ pour les deux modèles de perturbations qui nous intéressent, à savoir le modèle additif (1.2.5) et multiplicatif (1.2.4). Le caractère non rigoureux de cette méthode est discuté en détail dans la Section 3.1.4.

**1.3.4. Mouvement Brownien de Dyson.** La dernière méthode que nous allons présenter ne s'applique – pour l'instant – que dans le cadre du modèle additif avec un bruit Gaussien. Supposons que la matrice $\mathbf{B}$ dans l'équation (1.2.5) soit une matrice du GOE de variance $\sigma^2$, alors il est possible de voir la matrice $\mathbf{M}$ comme un processus de diffusion depuis l'article fondateur de Freeman Dyson [70]. Plus précisément, il est possible de réécrire le modèle (1.2.5) comme un processus de diffusion $(\mathbf{M}(t))_{t\geq 0}$ défini dans l'espace des matrices réelles symétriques et de taille $N \times N$[10] qui démarre de la matrice déterministe $\mathbf{C}$, que nous souhaitons estimer, et qui évolue dans le temps selon la dynamique suivante:

$$\mathbf{M}(t) \;:=\; \mathbf{C} + \mathbf{B}(t) \tag{1.3.10}$$

où $(\mathbf{B}(t))_{t\geq 0}$ est un mouvement Brownien symétrique, c'est-à-dire un processus de diffusion matriciel tel que $\mathbf{B}_0 = 0$ et ayant pour entrée $\{B_{ij}(t), i \leq j\}$ définie par

$$B_{ij}(t) := \frac{1}{\sqrt{N}} W_{ij}(t) \quad \text{si} \quad i \neq j, \quad B_{ii}(t) := \frac{\sqrt{2}}{\sqrt{N}} W_{ii}(t) \tag{1.3.11}$$

avec les $W_{ij}(t), i \leqslant j$ qui sont des mouvement Browniens réels indépendants et identiquement distribués.

[9] Ceci peut être vu comme une conséquence du théorème centrale limite, voir l'appendice B.4.
[10] Il est également possible de considérer l'ensemble des matrices Hermitiennes.

Cette description dynamique est très utile car elle nous permet d'étudier l'impact du bruit aussi bien sur les valeurs propres que sur les vecteurs propres de $\mathbf{C}$. En effet, depuis les travaux de Dyson, nous savons que les valeurs propres $\lambda_1(t) \geq \lambda_2(t) \geq \cdots \geq \lambda_N(t)$ de la matrice $\mathbf{M}(t)$ évoluent selon le mouvement Brownien de Dyson [70]:

$$\mathrm{d}\lambda_i(t) = \sqrt{\frac{2}{\beta N}}\,\mathrm{d}b_i(t) + \frac{1}{N}\sum_{k\neq i}\frac{\mathrm{d}t}{\lambda_i(t)-\lambda_k(t)}\,, \quad i=1,\dots,N\,, \tag{1.3.12}$$

où les $b_i(t)$ sont des mouvements Brownien indépendants et satisfont la condition initiale

$$\lambda_i(0) = \mu_i\,, \quad i=1,\dots,N\,,$$

avec $\{\mu_i\}_i$ l'ensemble des valeurs propres de $\mathbf{C}$. Nous pouvons voir dans la dynamique (1.3.12) que les valeurs propres se repoussent mutuellement avec un potentiel logarithmique.

Conditionnellement aux trajectoires des valeurs propres, nous pouvons étudier les trajectoires des vecteurs propres $\mathbf{u}_1(t)$, $\mathbf{u}_2(t)\rangle, \dots, \mathbf{u}_N(t)$ de $\mathbf{M}(t)$. La dynamique sur les vecteurs propres de matrices aléatoires provient de l'article de Bru sur les processus de Wishart [37]. Dans le cas du modèle (1.3.11), nous avons pour tout $i=1,\dots,N$:

$$\mathrm{d}\mathbf{u}_i(t) = -\frac{1}{2N}\sum_{k\neq i}\frac{\mathrm{d}t}{(\lambda_i(t)-\lambda_k(t))^2}\mathbf{u}_i(t) + \frac{1}{\sqrt{N}}\sum_{k\neq i}\frac{\mathrm{d}w_{ik}(t)}{\lambda_i(t)-\lambda_k(t)}\mathbf{u}_k(t)\,, \tag{1.3.13}$$

$$\text{avec} \quad \mathbf{u}_i(0) = \mathbf{v}_i\,, \tag{1.3.14}$$

où la famille des mouvements Browniens indépendants (symétrie mise à part) $\{w_{ij} : i \neq j\}$ est indépendante de la trajectoire des valeurs propres, c'est-à-dire indépendant des mouvements Brownien $b_i(t)$ dans l'équation (12.1.1). Par conséquent, nous voyons que dans le cas du modèle additif Gaussien, il est possible d'étudier les valeurs propres et les vecteurs propres de $\mathbf{M}$ avec une approche dynamique.

## 1.4 Contributions principales

Après avoir présenté les différents outils que nous avons utilisés dans ce manuscrit, nous proposons dans cette section un bref résumé des résultats obtenus durant cette thèse.

**1.4.1. Matrices de covariance.** Étant donné que le sujet principal porte sur l'estimation de matrices covariance à partir de l'estimateur empirique $\mathbf{E}$, nous dédions cette première partie à ce problème. Nous discuterons des autres modèles dans les sections suivantes. La plupart des résultats qui vont suivre sont issus des articles [40–43].

Avant de présenter les résultats, nous revenons brièvement sur les hypothèses du modèle considéré. Tout d'abord, nous nous intéressons au spectre de la matrice $\mathbf{C}$: nous autorisons la présence d'un nombre fini $r \geqslant 0$ (indépendant de $N$) de valeurs propres isolées à droite du bulk. Cette hypothèse sur les valeurs propres est motivée par les faits stylisés observés dans les vraies données (par exemple en finance [110] ou en biologie [136]). Cette modélisation est d'ailleurs parfaitement en phase avec l'ACP qui exploite justement la présence de grande valeur propres isolées pour trouver les composantes principales (les vecteurs propres) permettant de maximiser la variance expliquée (les valeurs propres). Le fait que les valeurs propres isolées se situent à droite du bulk permet de simplifier le problème mais les résultats restent vrais pour

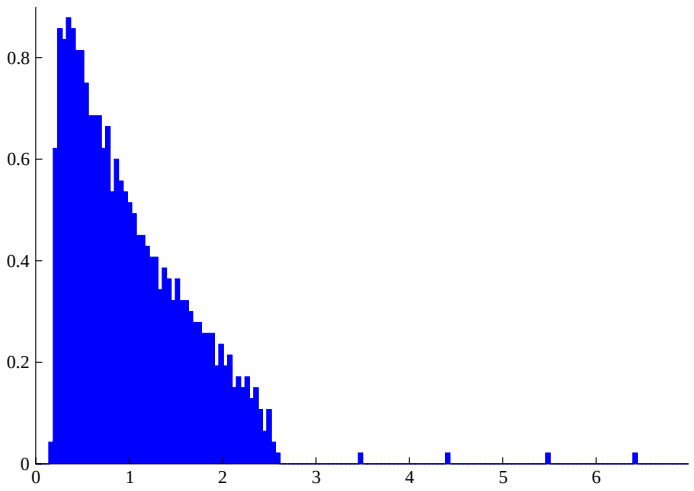



FIGURE 1.4.1. Histogramme de la densité des valeurs propres de $\mathbf{E}$ pour $N = 1000$ et $q = 0.3$. La vraie matrice est donnée par $\mathbf{C} = \mathbf{I}_N + \mathbf{B}$ où $\mathbf{B}$ est une matrice du GOE avec une variance $\sigma^2 = 0.0625$ et contenant 4 valeurs propres isolées situées à la position $\{3, 4, 5, 6\}$.

des valeurs propres isolées à gauche du bulk (suffisamment éloignées de l'origine). Par contre, nous n'effectuons aucune hypothèse sur les vecteurs propres de $\mathbf{C}$.

Ensuite, nous rappelons que l'estimateur empirique est obtenu grâce à la formule (1.1.2) où les colonnes de la matrice des données $\mathbf{Y}$ sont i.i.d. De plus, nous présumons que les 4 premiers moments de la distribution des entrées de la matrice $\mathbf{Y}$ sont bornés. Cette hypothèse technique est importante pour utiliser le résultat de Marčenko et Pastur. Nous illustrons dans la Figure 1.4.1 un exemple typique de la densité des valeurs propres observée, c'est-à-dire celles de $\mathbf{E}$, en présence de 4 valeurs propres isolées.

Nous rappelons que le problème initial est de comprendre le comportement asymptotique de l'estimateur oracle (1.1.12), qui est, comme son nom l'indique, non observable. Le résultat le plus important de cette thèse est que dans la limite des grandes dimensions, les valeurs propres de l'estimateur oracle (1.1.13) convergent vers une fonction qui ne dépend pas explicitement de la matrice $\mathbf{C}$. Ce résultat, étonnant aux premiers abords, est une conséquence directe du caractère ergodique de l'équation (1.1.13). En effet, pour $N \to \infty$, nous avons

$$\xi_i^{\text{ora.}} \sim \sum_{j=1}^{N} \mathbb{E}\big[\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2\big] \mu_j. \tag{1.4.1}$$

Au vu de ce dernier résultat, il semble qu'il y ait deux objets différents à comprendre: (i) l'espérance du produit scalaire au carré entre les vecteurs propres de $\mathbf{C}$ et de $\mathbf{E}$ (*overlap* par la suite) et (ii) les valeurs propres de $\mathbf{C}$. Néanmoins, nous allons voir que dans la limite des grandes dimensions, seule la connaissance du premier objet suffit pour obtenir le résultat annoncé. Cela montre en partie que la connaissance des valeurs propres de $\mathbf{C}$ ne semble pas être un pré-requis quand $N \to \infty$.

Depuis les travaux de Marčenko et Pastur, nous savons que la transformée de Stieltjes $\mathfrak{g}_{\mathbf{E}}(z)$ satisfait une équation au point-fixe mettant en jeu la transformée de Stieltjes de $\mathbf{C}$ [123]. De façon assez remarquable, cette relation est également valide pour la résolvante $\mathbf{G}_{\mathbf{E}}(z)$. Ce résultat, qui

est d'abord apparu dans [44] puis prouvé rigoureusement dans [109], donne:

$$\mathbf{G}_{\mathbf{E}}(z) = Z(z)\mathbf{G}_{\mathbf{C}}(Z(z))\,, \qquad Z(z) = \frac{z}{1+q+qzstj_{\mathbf{E}}(z)}\,, \tag{1.4.2}$$

et il suffit de considérer la trace normalisée de cette équation pour obtenir l'équation de Marčenko et Pastur. Durant ma thèse, j'ai pu étendre ce résultat à une classe plus large de processus stochastiques [40]. Nous reviendrons sur ce point par la suite.

Il était important de préciser ce résultat sur les résolvantes car il s'agit du point central dans le calcul des overlaps. Nous devons distinguer au moins trois cas différents dans le calcul de l'espérance des overlaps. Le premier cas concerne l'overlap entre des vecteurs propres de $\mathbf{E}$ et $\mathbf{C}$ associés à des valeurs propres du bulk. Ce cas était déjà connu suite à l'article d'Olivier Ledoit & Sandrine Péché que nous rappelons ici [113]:

$$\mathbb{E}\big[\langle \mathbf{u}_i\,, \mathbf{v}_j\rangle^2\big] \;=\; \frac{1}{N}\frac{q\lambda_i\mu_j}{\big|\lambda_i - \mu_j(1-q+q\lambda_i \lim_{\eta\to 0}\mathfrak{g}_{\mathbf{E}}(z-\mathrm{i}\eta)\big|^2}\,, \quad i\in[\![r+1,N]\!]\,, j\in[\![1,N]\!]\,, \tag{1.4.3}$$

pour $N,T\to\infty$ avec $q\geqslant 0$. Une dérivation de ce résultat à partir de la relation régissant la résolvante de $\mathbf{E}$ est donnée dans la Section 5.1.1. La remarque importante est que le vecteur propre $\mathbf{u}_i$ pour $i\in[\![1,r]\!]$ est délocalisé dans toutes les directions $\mathbf{v}_j$ pour tout $j$. Cela veut donc dire que l'information retenue par les vecteurs propres associés aux valeurs propres du bulk est arbitrairement faible. Nous pouvons ensuite nous demander s'il est possible d'étendre ce résultat pour une valeur propre isolée de $\mathbf{E}$ ($i\in[\![1,r]\!]$). Cette interrogation fut au coeur de ma collaboration avec Antti Knowles, dont l'article est en cours de rédaction. En particulier, nous avons obtenu que

$$\langle \mathbf{u}_i\,, \mathbf{v}_j\rangle^2 \;=\; \delta_{ij}\frac{\mu_i\theta(\mu_i)}{\theta'(\mu_i)} \qquad i,j\in[\![1,r]\!] \tag{1.4.4}$$

où $\theta(\mu_i)$ est une fonction d'ordre 1 définit dans l'équation (4.3.12). Cette formule généralise le résultat de Debashis Paul mentionné précédemment [144] et nous pouvons également en déduire la concentration du vecteur propre $\mathbf{u}_i$ pour $i\in[\![1,r]\!]$ autour d'un cône dans la direction du vecteur $\mathbf{v}_i$. Néanmoins, nous constatons que l'estimation n'est pas "parfaite" pour tout $q>0$, c'est-à-dire $\langle \mathbf{u}_i\,, \mathbf{v}_i\rangle^2\leqslant 1$ pour $i\in[\![1,r]\!]$, démontrant ainsi une perte d'information dûe au bruit de mesure. Maintenant, si nous considérons le cas où $j\geqslant r+1$, nous retrouvons le phénomène de délocalisation:

$$\mathbb{E}\langle \mathbf{u}_i\,, \mathbf{v}_j\rangle^2 \;=\; \frac{1}{T}\frac{\mu_j}{(1-\mu_j/\mu_i)^2}\,, \qquad i,j\in[\![1,r]\!]\,. \tag{1.4.5}$$

Ces deux résultats sont présentés plus en détails dans la section 5.1.2. Nous précisons également qu'il est possible de calculer les overlaps entre deux matrices de covariance empiriques indépendantes mais issues de la même vrai matrice $\mathbf{C}$ (voir Section 5.2 pour une présentation complète de ce problème).

Une fois que nous avons déterminé ces résultats concernant les vecteurs propres, nous sommes en mesure de conclure sur le problème principal qui est le comportement asymptotique de l'estimateur oracle (1.4.1). En effet, dans la limite $N\to\infty$, nous obtenons

$$\xi_i^{\text{ora.}} \sim \frac{\lambda_i}{\big|1-q+q\lambda_i\lim_{\eta\to 0}\mathfrak{g}_{\mathbf{E}}(\lambda_i-\mathrm{i}\eta)\big|^2}\,, \tag{1.4.6}$$

et nous constatons que la solution ne dépend plus explicitement de la matrice $\mathbf{C}$. Nous insistons sur le fait qu'il est donc possible de ré-écrire un estimateur non-observable en un estimateur

qui est dorénavant une fonction des valeurs propres de $\mathbf{E}$ dans la limite des grandes dimensions. Un cas intéressant est lorsque nous supposons que $\mathbf{C}$ est une Inverse-Wishart [9]. Alors, nous sommes en mesure de montrer en utilisant la dernière équation que l'estimateur oracle converge vers le shrinkage linéaire, ce qui est en parfaite adéquation avec les travaux de Haff [87]. Il semble qu'il y ait un lien entre ce résultat et les statistiques Bayésiennes.

Néanmoins, pour utiliser cette formule dans un cadre général, il nous faut un estimateur consistant de la transformée de Stieltjes $\mathfrak{g}_{\mathbf{E}}(z)$. Cela peut être fait en utilisant les résultats de concentration de l'article [109] pour obtenir le résultat suivant [43]:

$$\left|\hat{\xi}_i - \xi_i^{\text{ora.}}\right| \leqslant N^{-1/2+\varepsilon}, \quad \text{avec} \quad \hat{\xi}_i := \frac{\lambda_i}{\left|1 - q + qz\mathfrak{g}_{\mathbf{E}}^N(\lambda_i - \mathrm{i}N^{-1/2})\right|^2}, \tag{1.4.7}$$

pour tout $\lambda_i \geqslant c > 0$ et où nous rappelons que $\mathfrak{g}_{\mathbf{E}}^N(z)$ est la transformée de Stieltjes empirique. Un point remarquable dans le résultat (1.4.7) est que le résultat est identique pour toutes les valeurs propres de $\mathbf{E}$, même celles qui sont isolées. Nous reviendrons en détail sur ces observations dans le Chapitre 7.

Nous remarquons que le résultat (1.4.7) nécessite que la valeur propre que l'on cherche à nettoyer ne soit pas trop proche de l'origine. Or, dans de nombreux problèmes, comme celui du portefeuille de Markowitz, ce sont les petites valeurs propres qui nous intéressent. Pour remédier à ce problème, nous avons proposé dans [42] la procédure de régularisation suivante:

$$\hat{\xi}_i^{reg} = \hat{\xi}_i \times \max(1, \Gamma_i), \tag{1.4.8}$$

où $\Gamma_i$ est une fonction analytique, définit dans l'équation (9.1.4), qui permet de corriger l'erreur d'estimation pour les petites valeurs propres. Nous soulignons que cette fonction $\Gamma_i$ est indépendante de la matrice $\mathbf{C}$, ce qui signifie que nous pouvons l'utiliser en toute circonstance. De plus, sa simplicité fait que cette méthode est très simple à implémenter en pratique et redonne des résultats semblables à la méthode d'inversion de Ledoit & Wolf qui s'avère très complexe à mettre en place [117, 118].

Finalement, nous avons testé cet estimateur sur des données financières afin de le comparer avec les estimateurs classiques de la littérature. Le test considéré est la minimisation du risque réalisé d'un portefeuille de Markowitz construit avec $N = 450$ actifs, $T = 900$ jours de trading et pour trois zones géographiques distinctes: les États-Unis, l'Europe et le Japon. Les simulations de la Section 8.1.3 montrent clairement que l'estimateur (1.4.8) offre le meilleur contrôle des risques réalisés pour toutes les zones géographiques considérées, pour différentes stratégies et également pour différentes valeurs de $N$. Nous pouvons donc conclure que l'estimateur mis en place durant cette thèse atteint bien l'objectif espéré, c'est-à-dire une estimation précise des corrélations en présence d'un grand nombre de variables.

Nous pouvons résumer ce long travail sur l'estimation des grandes matrices de covariance par la Figure 2.1.2, où l'on voit la différence entre l'estimateur (1.4.8) et les estimateurs mentionnés précédemment dans l'introduction. Nous insistons sur le fait que cet estimateur est optimal dans la classe des matrices $\mathcal{M}(\mathbf{E})$ et dans la limite des grandes dimensions, mais nous pouvons nous demander s'il est possible d'améliorer les performances de l'estimateur en supposant une structure à priori sur les vecteurs propres de $\mathbf{C}$. Cette question est fondamentale en pratique et constitue une piste possible de recherche pour le futur.

**1.4.2. Le modèle additif Gaussien.** Pour montrer la grande généralité du problème, nous avons reconsidéré la même problématique pour le modèle (1.3.11). Le modèle est suffisamment simple

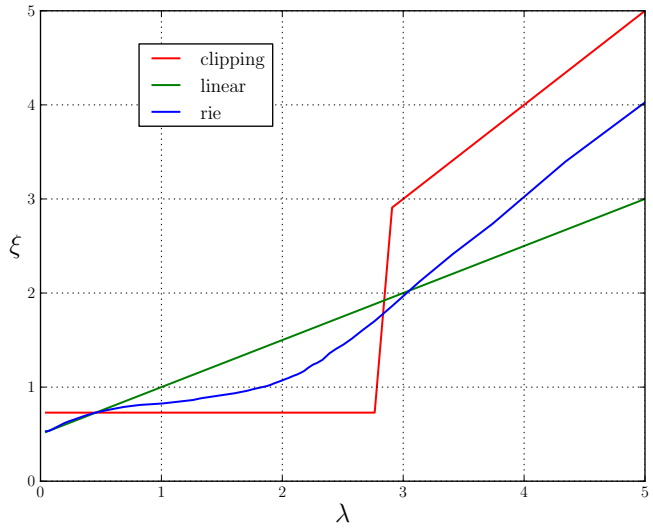



FIGURE 1.4.2. Résumé des méthodes de nettoyage des valeurs propres (axe des ordonnées) en fonction des valeurs propres observées (voir Chapitre 9 pour plus de détails). Cette image montre l'évolution des méthodes de shrinkage analytique démarrant de la méthode linéaire (vert), l'heuristique du clipping (rouge) et finalement l'estimateur optimal (1.4.8) (bleu).

pour que toutes les formules soient explicites comme pour les matrices de covariance empirique. Encore une fois, nous supposons la présence d'un nombre fini $r \geqslant 0$ de valeurs propres isolées dans le spectre de la matrice $\mathbf{C}$. En utilisant la dynamique des valeurs propres et des vecteurs propres, nous pouvons étudier les overlaps dans deux cas différents: (i) lorsque les vecteurs propres de $\mathbf{M}$ et $\mathbf{C}$ sont associés à des valeurs propres du bulk et (ii) lorsque les vecteurs propres de $\mathbf{M}$ et $\mathbf{C}$ sont associés à des valeurs propres isolées. La troisième configuration des overlaps mentionnée dans la section précédente reste un problème ouvert. Ainsi, nous ne sommes pas encore capables de déterminer l'équivalent asymptotique de l'estimateur oracle pour les valeurs propres isolées. Les résultats de cette section sont issus des articles [5, 40].

Plaçons nous dans le premier cas. En utilisant la dynamique des vecteurs propres (1.3.13), nous obtenons [5]:

$$\mathbb{E}\big[\langle \mathbf{u}_i(t), \mathbf{v}_j \rangle^2\big] \;=\; \frac{1}{N} \frac{t}{\big|\lambda_i(t) - t\mathfrak{g}_{\mathbf{M}}(z,t) - \mu_j\big|^2}\,, \qquad i \in [\![r+1, N]\!]\,, j \in [\![1, N]\!]\,, \tag{1.4.9}$$

où nous rappelons que $t$ est la variance de la matrice du GOE $\mathbf{B}(t)$. Ce résultat n'est pas nouveau [3, 150] mais l'approche par le mouvement Brownien lui donne une interprétation physique claire. Le cas des vecteurs propres associés aux valeurs propres isolées est par contre un résultat nouveau [5]:

$$\mathbb{E}\big[\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2\big] \sim \delta_{ij} \exp\left(-\frac{1}{2}\int_0^t ds \int_{\mathbb{R}} \frac{\rho_{\mathbf{M}}(\lambda, s)}{(\lambda_i(s) - \lambda)^2}\, \mathrm{d}\lambda\right)\,, \qquad i, j \in [\![1, r]\!]\,, \tag{1.4.10}$$

pour $N \to \infty$. Ces résultats sont démontrés dans le Chapitre 12.

Dans ce modèle, l'overlap entre deux réalisations bruitées de $\mathbf{C}$ est particulièrement intéressant. En effet, contrairement aux matrices de covariance, l'indépendance du bruit n'est pas nécessaire. Ainsi, supposons que nous observons deux matrices $\mathbf{M}$ et $\tilde{\mathbf{M}}$ définies par

$$\mathbf{M} = \mathbf{C} + \mathbf{B}, \qquad \tilde{\mathbf{M}} = \mathbf{C} + \tilde{\mathbf{B}}\,, \tag{1.4.11}$$

où $\mathbf{B}$, $\tilde{\mathbf{B}}$ sont des matrices du GOE corrélées (avec coefficient $\rho \in [-1, 1]$). Alors, le résultat final dans la limite des grandes dimensions est doublement surprenant. Tout d'abord, la connaissance de $\mathbf{C}$ n'est pas explicitement requise et de plus, le résultat est identique dans le cas où les bruits $\mathbf{B}$ et $\tilde{\mathbf{B}}$ sont corrélés. La seule différence intervient dans la variance du bruit. L'énoncé précis de ce résultat est donné dans le Chapitre 12 et est issu de l'article [41]. Un problème ouvert très important serait de pouvoir répéter un argument similaire dans le cas des matrices de covariance empiriques.

Pour terminer, revenons au problème de l'estimateur oracle. Posons dorénavant $\mathbf{B}$ comme une matrice du GOE avec une variance $\sigma^2$ (au lieu du paramètre $t$). En suivant la même approche que dans le Chapitre 7 mais en utilisant cette fois le résultat (1.4.9), nous pouvons montrer à nouveau que l'équation (1.1.13) converge vers une fonction limite qui ne dépend plus explicitement de $\mathbf{C}$. En effet, le résultat final est donné par:

$$\xi_i^{\text{ora.}} \sim \lambda_i - 2\sigma^2 \mathfrak{h}_{\mathbf{M}}(\lambda_i)\,, \qquad i \in [\![r+1, N]\!]\,, \tag{1.4.12}$$

où $\mathfrak{h}_{\mathbf{M}}$ est la partie réelle de $\lim_{\eta\to 0} \mathfrak{g}_{\mathbf{M}}(\lambda_i - \mathrm{i}\eta)$ dans la limite $N \to \infty$. Lorsque nous considérons $\mathbf{C}$ comme étant également une matrice du GOE de variance $\sigma^2_{\mathbf{C}}$, l'application de ce résultat nous permet réécrire (1.4.12) comme suit:

$$\xi_i^{\text{ora.}} = \lambda_i \left( \frac{\sigma^2_{\mathbf{C}}}{\sigma^2_{\mathbf{C}} + \sigma^2} \right) , , \qquad i \in [\![r+1, N]\!]\,, \tag{1.4.13}$$

où nous voyons que la solution optimale au problème d'estimation de $\mathbf{C}$ revient à nettoyer les valeurs propres de $\mathbf{M}$ par le ratio signal sur bruit. Ce résultat est un résultat connu en statistiques Bayésienne et il est intéressant de noter à nouveau un lien entre cette théorie et l'estimateur RIE. Une analyse plus poussée de ces résultats est donnée dans la Section 13 et sont issus de l'article [40].

**1.4.3. Extension aux modèles de probabilités libres.** La dernière section de cette introduction générale concerne l'extension d'une partie des résultats mentionnés précédemment dans le cadre des modèles généraux d'addition et de multiplication libres. Nous avons motivé – surtout pour le modèle multiplicatif – l'intérêt pratique d'étudier ce type de modèles. La grande majorité des résultats suivants proviennent de l'article [40]. Durant toute cette section, nous supposons que dans la limite $N \to \infty$:

$$\lim_{\eta\to 0} \mathfrak{g}_{\mathbf{M}}(\lambda - \mathrm{i}\eta) \;=\; \mathfrak{h}_{\mathbf{M}}(\lambda) + \mathrm{i}\pi\rho_{\mathbf{M}}(\lambda)\,, \tag{1.4.14}$$

soit bien définie pour toute matrice $\mathbf{M}$ caractérisée par les modèles (1.2.4) ou (1.2.5). De plus, nous supposons dans toute cette partie qu'il n'y a pas de valeur propre isolée ($r = 0$). Nous reviendrons sur cette hypothèse à la fin de cette section.

Intéressons nous d'abord au modèle d'addition libre (1.2.5). Nous rappelons que dans ce cas précis, le bruit extérieur $\mathbf{B}$ est invariant par rotation, indépendant de $\mathbf{C}$ mais possède une distribution des valeurs propres arbitraires. Le premier résultat que nous avons été en mesure de généraliser est le comportement asymptotique de la résolvante de $\mathbf{M}$:

$$\mathbf{G}_{\mathbf{M}}(z) = \mathbf{G}_{\mathbf{C}}(Z(z)), \qquad Z(z) \;:=\; z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z))\,. \tag{1.4.15}$$

Il est important de préciser que si $\mathbf{B}$ est une matrice du GOE, alors nous retrouvons le résultat attendu (voir le Chapitre 13 pour plus de détails). Encore une fois, nous voyons que la résolvante de $\mathbf{M}$ tend vers une limite déterministe, ce qui simplifie les calculs.

A partir de ce résultat, nous pouvons calculer l'espérance des overlaps entre un état perturbé et non perturbé:

$$N\mathbb{E}\big[\langle \mathbf{u}_i\,,\mathbf{v}_j\rangle^2\big] = \frac{\beta_1(\lambda)}{(\lambda-\mu-\alpha_a(\lambda))^2+\pi^2\beta_a(\lambda)^2\rho_{\mathbf{M}}(\lambda)^2}, \qquad i,j\in[\![1,N]\!] \tag{1.4.16}$$

où nous avons définis les fonctions:

$$\begin{cases} \alpha_a(\lambda) &:= \mathrm{Re}[\mathcal{R}_{\mathbf{B}}\left(\mathfrak{h}_{\mathbf{M}}(\lambda)+\mathrm{i}\pi\rho_{\mathbf{M}}(\lambda)\right)], \\ \beta_a(\lambda) &:= \dfrac{\mathrm{Im}[\mathcal{R}_{\mathbf{B}}\left(\mathfrak{h}_{\mathbf{M}}(\lambda)+\mathrm{i}\pi\rho_{\mathbf{M}}(\lambda)\right)]}{\pi\rho_{\mathbf{M}}(\lambda)}\,. \end{cases} \tag{1.4.17}$$

Cela nous permet d'en déduire une formule (formelle) pour la valeur asymptotique de l'estimateur oracle (1.4.1):

$$\xi_i^{\text{ora.}} \sim F_a(\lambda_i)\,, \qquad F_a(\lambda)=\lambda-\alpha_a(\lambda)-\beta_a(\lambda)\mathfrak{h}_{\mathbf{M}}(\lambda)\,. \tag{1.4.18}$$

Nous pouvons constater que le résultat reste toujours "observable" dans le sens où la connaissance de $\mathbf{C}$ ne semble pas être un pré-requis pour utiliser cette formule. Par contre, cela suppose que nous connaissons au moins le spectre de la matrice $\mathbf{B}$ dans la limite des grandes matrices. Il n'est pas étonnant de retrouver la transformée $\mathcal{R}$ dans ces résultats étant donné qu'elle caractérise justement l'addition d'opérateurs non-commutatifs.

La même analyse peut être menée pour le modèle de multiplication libre (1.2.4). Posons $\mathbf{M}:=\mathbf{C}^{1/2}\mathbf{\Omega}\mathbf{B}\mathbf{\Omega}^*\mathbf{C}^{1/2}$ où $\mathbf{B}$ est une matrice aléatoire symétrique, invariante par rotation et de taille $N\times N$, et $\mathbf{\Omega}$ est une matrice de rotation de taille $N\times N$ qui est distribuée selon la mesure de Haar. Pour ce modèle, la relation des résolvantes est donnée par

$$\mathbf{G}_{\mathbf{M}}(z)=Z(z)\mathbf{G}_{\mathbf{C}}(Z(z)), \qquad Z(z)\;:=\;z\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z)-1)\,, \tag{1.4.19}$$

qui est bien une généralisation de la relation (1.4.2) [40]. De façon analogue au cas additif, il n'est pas étonnant de rencontrer la transformée $\mathcal{S}$ dans ce cadre ci, étant donne qu'elle caractérise le produit d'opérateurs non-commutatifs. En utilisant ce résultat, nous pouvons en déduire l'overlap moyen:

$$N\mathbb{E}\big[\langle \mathbf{u}_i\,,\mathbf{v}_j\rangle^2\big] = \frac{\mu\beta_m(\lambda)}{(\lambda-\mu\alpha_m(\lambda))^2+\pi^2\mu^2\beta_m(\lambda)^2\rho_{\mathbf{M}}(\lambda)^2}\,, \qquad i,j\in[\![1,N]\!]\,, \tag{1.4.20}$$

où les fonctions $\alpha_m$ and $\beta_m$ sont définies comme suit:

$$\begin{cases} \alpha_m(\lambda) := \lim\limits_{z\to\lambda-i0^+}\mathrm{Re}\left[\dfrac{1}{\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z)-1)}\right] \\ \beta_m(\lambda) := \lim\limits_{z\to\lambda-i0^+}\mathrm{Im}\left[\dfrac{1}{\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z)-1)}\right]\dfrac{1}{\pi\rho_{\mathbf{M}}(\lambda)}\,. \end{cases} \tag{1.4.21}$$

Comme pour le modèle additif, l'estimateur oracle (1.4.1) converge vers une fonction qui ne nécessite pas explicitement la connaissance de la matrice $\mathbf{C}$ que l'on cherche à estimer. En effet, supposons que la transformée $\mathcal{S}$ de la matrice $\mathbf{B}$ est analytique, alors nous avons

$$\xi_i^{\text{ora.}} \sim F_2(\lambda_i); \qquad F_2(\lambda)=\lambda\gamma_{\mathbf{B}}(\lambda)+(\lambda\mathfrak{h}_{\mathbf{M}}(\lambda)-1)\omega_{\mathbf{B}}(\lambda)\,, \tag{1.4.22}$$

avec

$$\lim_{z\to\lambda-\mathrm{i}0^+} \mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z)-1) := \gamma_{\mathbf{B}}(\lambda) + \mathrm{i}\pi\rho_{\mathbf{M}}(\lambda)\omega_{\mathbf{B}}(\lambda)\,. \tag{1.4.23}$$

Les résultats du modèle multiplicatif sont expliqués dans les Chapitres 3 et 7.

En conclusion, nous sommes capables d'étudier le comportement asymptotique de l'estimateur oracle dans un cadre assez général de matrices aléatoires. Cependant, les résultats font apparaître les transformées $\mathcal{R}$ et $\mathcal{S}$, dont les structures ne sont pas simples à analyser. Il serait donc intéressant de voir s'il est possible de trouver des exemples concrets, comme ceux mentionnés dans la Section 1.2, pour lesquels nous pouvons obtenir des résultats explicites comme dans les deux sections précédentes. De plus, l'extension de ces résultats en présence de valeurs propres isolées est un problème ouvert très important aussi bien en théorie qu'en pratique.

# Part I

# Advances in large covariance matrices estimation

# Chapter 2

# Introduction

## 2.1 Motivations

This part, which is the bulk of the thesis, is dedicated to the estimation of large sample covariance matrices. Indeed, in the present era of "Big Data", new statistical methods are needed to decipher large dimensional data sets that are now routinely generated in almost all fields – physics, image analysis, genomics, epidemiology, engineering, economics and finance, to quote only a few. It is very natural to try to identify common causes (or factors) that explain the joint dynamics of $N$ quantities. These quantities might be daily returns of the different stocks of the S&P 500, temperature variations in different locations around the planet, velocities of individual grains in a packed granular medium, or different biological indicators (blood pressure, cholesterol, etc.) within a population, etc., etc. The simplest mathematical object that quantifies the similarities between these observables is an $N \times N$ correlation matrix $\mathbf{C}$. Its eigenvalues and eigenvectors can then be used to characterize the most important common dynamical "modes", i.e. linear combinations of the original variables with the largest variance. This is the well known "Principal Component Analysis" (or PCA) method. More formally, let us denote by $\mathbf{y} \in \mathbb{R}^N$ the set of demeaned and standardized[1] variables which are thought to display some degree of interdependence. Then, one possible way to quantify the underlying interaction network between these variables is through the standard, Pearson correlations:

$$\mathbf{C}_{ij} = \mathbb{E}\big[y_i y_j\big], \quad i,j \in [\![1,N]\!], \tag{2.1.1}$$

We will refer to the matrix $\mathbf{C}$ as the *population* correlation matrix throughout the following.

The major concern in practice is that the expectation value in (2.1.1) is rarely computable precisely because the underlying distribution of the vector $\mathbf{y}$ is unknown and is what one is struggling to determine. Empirically, one tries to infer the matrix $\mathbf{C}$ by collecting a large number $T$ of realizations of these $N$ variables that defines the input sample data matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T) \in \mathbb{R}^{N \times T}$. Then, in the case of a sufficiently large number of realizations $T$, one tempting solution to estimate $\mathbf{C}$ is to compute that *sample correlation matrix* estimator $\mathbf{E}$, defined as:

$$E_{ij} \;:=\; \frac{1}{T}\sum_{t=1}^{T} Y_{it}\, Y_{jt} \;\equiv\; \frac{1}{T}\left(\mathbf{Y}\mathbf{Y}^*\right)_{ij}, \tag{2.1.2}$$

[1] This apparently innocuous assumption will be discussed in Chapter 4.

where $Y_{it}$ is the realization of the $i$th observable ($i = 1, \dots, N$) at "time" $t$ ($t = 1, \dots, T$) that will be assumed in the following to be demeaned and standardized (see previous footnote).

Indeed, in the case where $N \ll T$, it is well known using result of classical multivariate statistics that $\mathbf{E}$ converges (almost surely) to $\mathbf{C}$ [179]. However, when $N$ is large, the simultaneous estimation of all $N(N-1)/2$ the elements of $\mathbf{C}$ – or in fact only of its $N$ eigenvalues – becomes problematic when the total number $T$ of observations is not very large compared to $N$ itself. In the example of stock returns, $T$ is the total number of trading days in the sampled data; but in the biological example, $T$ would be the size of the population sample, etc. Hence, in the modern framework of high-dimensional statistics, the empirical correlation matrix $\mathbf{E}$ (i.e. computed on a given realization) must be carefully distinguished from the "true" correlation matrix $\mathbf{C}$ of the underlying statistical process (that might not even be well defined). In fact, the whole point of the present part is to characterize the difference between $\mathbf{E}$ and $\mathbf{C}$, and discuss how well (or how badly) one may reconstruct $\mathbf{C}$ from the knowledge of $\mathbf{E}$ in the case where $N$ and $T$ become very large but with their ratio $q = N/T$ not vanishingly small; this is often called the large dimension limit (LDL), or else the "Kolmogorov regime".

There are numerous situations where the estimation of the high-dimensional covariance matrix is crucial. Let us give some well-known examples:

(i) Generalized least squares (GLS): Suppose we try to explain the vector $\mathbf{y}$ using a linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1.3}$$

where $X$ is a $N \times k$ design matrix ($k \geqslant 1$), $\boldsymbol{\beta}$ denotes the regression coefficients to these $k$ factors, and $\boldsymbol{\varepsilon}$ denotes the residual. Typically, one seeks to find $\boldsymbol{\beta}$ that best explains the data and this exactly the purpose of GLS. Assume that $\mathbf{E}[\boldsymbol{\varepsilon}|X] = 0$ and $\mathbb{V}[\boldsymbol{\varepsilon}|X] = \mathbf{C}$ the covariance matrix of the residuals. Then GLS estimates $\boldsymbol{\beta}$ as (see [7] for a more detailed discussion):

$$\widehat{\boldsymbol{\beta}} = (X^* \mathbf{C} X)^{-1} X^* \mathbf{C}^{-1} \mathbf{y}. \tag{2.1.4}$$

We shall investigate this estimator in Section 8.

(ii) Generalized methods of moments (GMM): Suppose one wants to calibrate the parameters $\Theta$ of a model on some data set. The idea is to compute the empirical average of a set of $k$ functions (generalized moments) of the data, which should all be zero for the correct values of the parameters, $\Theta = \Theta_0$. The distance to zero is measured using the covariance of these functions. A precise measurement of this $k \times k$ covariance matrix increases the efficiency of the GMM – see [88]. Note that GLS is a special form of GMM.

(iii) Classification (LDA) [79]: Suppose that we want to classify the variables $\mathbf{y}$ between two Gaussian populations with different mean $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, priors $\pi_1$ and $\pi_2$, but same covariance matrix $\mathbf{C}$. The LDA rule classifies $\mathbf{y}$ to class 2 if

$$\mathbf{x}^* \mathbf{C}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \frac{1}{2} (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^* \mathbf{C}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \log(\pi_2/\pi_1) \tag{2.1.5}$$

(iv) Large portfolio optimization [125]: Suppose we want to invest on a set of financial assets $\mathbf{y}$ in such a way that the overall risk of the portfolio is minimized, for a given performance target $\nu$. According to Markowitz's theory, the optimal investment strategy is a vector of weights $\mathbf{w} := (w_1, \dots, w_p)^*$ that can be obtained through a quadratic optimization program where we minimize the variance of the strategy $\langle \mathbf{w}, \mathbf{C}\mathbf{w} \rangle$ subject to a constraint

on the expectation value $\langle \mathbf{w}, \mathbf{g} \rangle \geqslant \mu$, with $\mathbf{g}$ a vector of predictors and $\mu$ fixed. (Other constraints can also be implemented). The optimal strategy reads

$$\mathbf{w} = \nu \frac{\mathbf{C}^{-1}\mathbf{g}}{\mathbf{g}^* \mathbf{C}^{-1}\mathbf{g}}. \tag{2.1.6}$$

As we shall see in Chapter 8, a common measure of the "risk" of estimation in high-dimensional problems like (i) and (iv) above is given by $\mathrm{Tr}\mathbf{E}^{-1}/\mathrm{Tr}\mathbf{C}^{-1}$, which turns out to be very close to unity $T$ is large enough for a fixed $N$, i.e. when $q = N/T \to 0$. However, when the number of observables $N$ is also large, such that the ratio $q$ is not very small, we will find below that $\mathrm{Tr}\mathbf{E}^{-1} = \mathrm{Tr}\mathbf{C}^{-1}/(1-q)$ for a wide class of processes. In other words, the out-of-sample risk $\mathrm{Tr}\mathbf{E}^{-1}$ can excess by far the true optimal risk $\mathrm{Tr}\mathbf{C}^{-1}$ when $q > 0$, and even diverge when $q \to 1$. Note that for a similar scenario when Value-at-Risk is minimized in-sample was elicited in [49] and in [53] for the Expected Shortfall. Typical number in the case of stocks is $N = 500$ and $T = 2500$, corresponding to 10 years of daily data, already quite a long strand compared to the lifetime of stocks or the expected structural evolution time of markets, but that corresponds to $q = 0.2$. For macroeconomic indicators – say inflation, 20 years of monthly data produce a meager $T = 240$, whereas the number of sectors of activity for which inflation is recorded is around $N = 30$, such that $q = 0.125$. Clearly, effects induced by a non zero value of $q$ are expected to be highly relevant in many applications.

**2.1.1. Historical survey.** The rapid growth of RMT (Random Matrix Theory) in the last two decades is due both to the increasing complexity of the data in many fields of science (the "Big Data" phenomenon) and to many new, groundbreaking mathematical results that challenge classical results of statistics. In particular, RMT has allowed a very precise study of large sample covariance matrices and also the design of estimators that are consistent in the large dimensional limit (LDL) presented above. The aim of this thesis is to provide the reader an introduction to the different RMT inspired techniques that allow one to investigate problems of high-dimensional statistics, with the estimation of large covariance matrices as the main thread.

The estimation of covariance matrices is a very old problem in multivariate statistics and one of the most influential work goes back to 1928 with John Wishart [192] who investigated the distribution of the sample covariance matrix $\mathbf{E}$ in the case of i.i.d Gaussian realizations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$. In particular, Wishart obtained the following explicit expression for the distribution of $\mathbf{E}$ given $\mathbf{C}$ [192]:

$$\mathcal{P}_W(\mathbf{E}|\mathbf{C}) = \frac{T^{NT/2}}{2^{NT/2}\Gamma_N(T/2)} \frac{\det(\mathbf{E})^{\frac{T-N-1}{2}}}{\det(\mathbf{C})^{T/2}} e^{-\frac{T}{2}\mathrm{Tr}\mathbf{C}^{-1}\mathbf{E}}, \tag{2.1.7}$$

where $\Gamma_N(\cdot)$ is the multivariate Gamma function with parameter $N$.[2] In Statistics, one says that $\mathbf{E}$ follows a Wishart$(N, T, \mathbf{C}/T)$ distribution and it is often referred to as one of the first result in RMT. Note that for a finite $N$ and $T$, the marginal probability density distribution of the eigenvalues is known [9]:

$$\rho_N(\lambda) = \frac{1}{N} \sum_{k=0}^{N-1} \frac{k!}{T-N+k} \left[L_k^{T-N}(\lambda)\right]^2 \lambda^{T-N} e^{-\lambda}, \tag{2.1.8}$$

[2] $\Gamma_N(u) = \pi^{N(N-1)/4} \prod_{j=1}^{N} \Gamma(u + (1-j)/2)$.

where we assumed that $T > N$ and $L_k^l$ are the Laguerre polynomials[3].

Even though the Wishart distribution gives us many important properties concerning $\mathbf{E}$, the behavior of the sample estimator as a function of $N$ was understood much later with the pioneering work of Charles Stein in 1956 [165]. The most important contribution of Stein can be summarized as follows: when the number of variables $N \geqslant 3$, there exist combined estimators more accurate in terms of *mean squared error* than any method that handles the variables separately (see [72] for an elementary introduction). This phenomenon is called *Stein's paradox* and establishes in particular that the sample matrix $\mathbf{E}$ becomes more and more inaccurate as the dimension of the system $N$ grows. The idea of "combined" estimators has been made precise with the James-Stein estimator [96] for the mean of a Gaussian vector that outperforms traditional methods such as maximum likelihood or least squares whenever $N \geqslant 3$. To achieve this, the authors used a *Bayesian* point of view, i.e. by assuming some *prior* probability distribution on the parameters that we aim to estimate. For sample covariance matrices, Stein's paradox also occurs for $N \geqslant 3$ as shown by using properties of the Wishart distribution and the so-called *conjugate* prior technique (see Chapter 6). This was first shown for the *precision* matrix $\mathbf{C}^{-1}$ in [71, 85] and then for the covariance matrix $\mathbf{C}$ in [87] and lead to the famous *linear shrinkage* estimator

$$\mathbf{\Xi} = \alpha_s \mathbf{E} + (1 - \alpha_s)\mathbf{I}_N, \tag{2.1.9}$$

where $\Xi$ denotes, here and henceforth, an estimator of $\mathbf{C}$ and $\alpha_s \in (0, 1)$ is the shrinkage intensity parameter. In [87], Haff proposed to estimate $\alpha_s$ using the marginal probability distribution of the observed matrix $\mathbf{Y}$ as advocated in the so-called *empirical* Bayes framework. We see that this shrinkage estimator interpolates between the empirical "raw" matrix $\mathbf{E}$ (no shrinkage, $\alpha_s = 1$) and the null hypothesis $\mathbf{I}_N$ (extreme shrinkage, $\alpha_s = 0$). This example illustrates the idea of a combined estimator, not based only on the data itself, that offers better performance when the dimension of the system grows. The improvement made by using the simple estimator (2.1.9) rather than the sample covariance matrix $\mathbf{E}$ has been precisely quantified much later in 2004 [115] in the asymptotic regime $N \to \infty$, with an explicit and observable estimator for the shrinkage intensity $\alpha_s$. To summarize, the Bayesian approach turns out to be a cornerstone in estimating high dimensional covariance matrices and will be discussed in more details in the Section 6.

Interestingly, the first result on the behavior of sample covariance matrices in the LDL did not come from the statistics community. It is due to the seminal work of Marčenko and Pastur in 1967 [123] where they obtained a self-consistent equation for the spectrum of $\mathbf{E}$ *given* $\mathbf{C}$ as $N$ goes to infinity. In particular, the influence of the quality ratio $q$ appears precisely. Indeed, it was shown in the classical limit $T \to \infty$ and $N$ fixed in 1963 by Anderson that the sample eigenvalues converge to the population eigenvalues [10], a result indeed recovered by the Marčenko-Pastur formula for $q = 0$. However, when $q = \mathcal{O}(1)$, the same formula shows that all the sample eigenvalues become noisy estimators of the "true" (population) ones no matter how large $T$ is. This is also called the *curse of dimensionality*. More precisely, the distortion of the spectrum of $\mathbf{E}$ compared to the "true" one becomes more and more substantial as $q$ becomes large (see Figure 2.1.1). The heuristic behind this phenomenon is as follows. When the sample size $T$ is very large, each individual coefficient of the covariance matrix $\mathbf{C}$ can be estimated with negligible error (provided one can assume that $\mathbf{C}$ itself does vary with time, i.e. that the observed process is stationary). But if $N$ is also large and of the order of $T$, as is often the case in many situations, the sample estimator $\mathbf{E}$ becomes "inadmissible". More specifically, the large

[3] $L_k^l(\lambda) = \frac{e^\lambda}{k!\lambda^l}\frac{d^k}{d\lambda^k}(e^{-\lambda}\lambda^{k+l})$.

number of simultaneous noisy variables creates important systematic errors in the computation of the eigenvalues of the matrix.

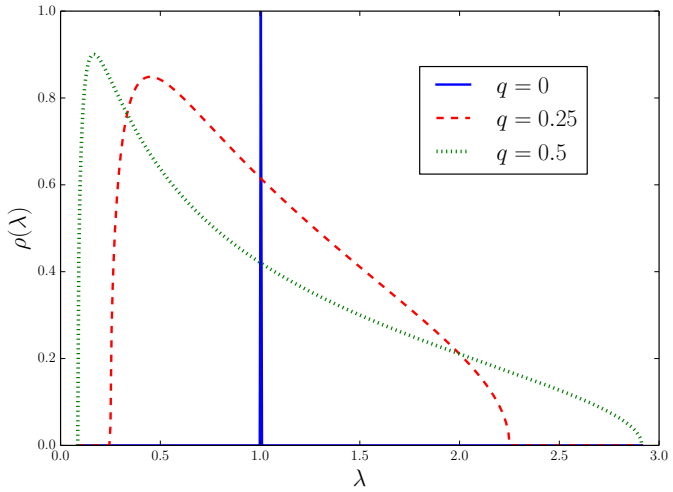



FIGURE 2.1.1. Plot of the sample eigenvalues and the corresponding sample eigenvalues density under the null hypothesis with $N = 500$. The blue line ($q = 0$) corresponds to a perfect estimation of the population eigenvalues. The larger is the observation ratio $q$, the wider is the sample density. We see that even for $T = 4N$, the deviation from the population eigenvalues is significant.

The Marčenko-Pastur result had a tremendous impact on the understanding the "curse of dimensionality". Firstly, it was understood in 1995 that this result is to a large degree *universal* when $N \to \infty$ and $q = \mathcal{O}(1)$, much as the Wigner semi-circle law is universal: the Marčenko-Pastur equation is valid for a very broad range of random measurement processes and for general population covariance matrix $\mathbf{C}$ [157, 161, 195]. This property is in fact at the core of RMT which makes this theory particularly appealing. At the same time, some empirical evidence of the relevance of these results for sample covariance matrices weres provided in [110, 151] using financial data sets, which are known to be non-Gaussian [30]. More precisely, these works suggested that most of the eigenvalues (the *bulk*) of financial correlation matrices agrees, to a first approximation, with the null hypothesis $\mathbf{C} = \mathbf{I}$, while a finite number of "spikes" (*outliers*) reside outside of the bulk. This observation is the very essence of the *spiked covariance matrix* model named after the celebrated paper of Johnstone in 2001 with many applications in *principal components analysis* (PCA) [100]. Indeed, the author showed another manifestation of universal properties of RMT, namely the Tracy-Widom distribution for the top bulk eigenvalues in the spiked covariance matrix [100, 175]. This result suggest that the edge of the bulk of eigenvalues is very *rigid* in the sense that the position of the edge has very small fluctuations of order $T^{-2/3}$. This provides a very simple recipe to distinguish meaningful eigenvalues (beyond the edge) from noisy ones (inside the bulk) [111, 151]. This method is known as "eigenvalue *clipping*": all eigenvalues in the bulk of the Marčenko-Pastur spectrum are deemed as noise and thus replaced by a constant value whereas the principal components outside of the bulk (the spikes) are left unaltered. This very simple method provides robust out-of-sample performance [29] and

emphasizes that the notion of regularization – or cleaning – is very important in high-dimension.

Even if the spiked covariance matrix model provides quite satisfactory results in many different contexts [29], one may want to work without such an assumption on the structure of $\mathbf{C}$ using the Marčenko-Pastur equation to reconstruct numerically the spectrum of $\mathbf{C}$ [163]. However, this is particularly difficult in practice since the Marčenko-Pastur equation is easy to solve in the other direction, i.e. knowing the spectrum of $\mathbf{C}$, we easily get the spectrum of $\mathbf{E}$. In that respect, many studies attempting to "invert" the Marčenko-Pastur equation appeared since 2008 [29, 104, 133, 194]. The first one consists in finding a parametric "true" spectral density that fits the data [29]. The method of [133], further improved in [194], is completely different. Under the assumption that the spectrum of $\mathbf{C}$ consists of a finite number of eigenvalues, an exact analytical estimator of each population eigenvalue is provided. However, this method requires some very strong assumptions on the structure of the spectrum of $\mathbf{C}$. The last approach can be considered as a *nonparametric* method and seems to be very appealing. Indeed, El Karoui proposed a "consistent" numerical scheme to invert the Marčenko-Pastur equation using the observed sample eigenvalues [104]. Nevertheless, while the method is very informative, it turns out that the algorithm also needs prior knowledge on the location of the true eigenvalues which makes the implementation difficult in practice.

These inversion schemes thus allow in principle to retrieve the spectrum of $\mathbf{C}$ but as far as estimating high-dimensional covariance matrices is concerned, merely substituting the sample eigenvalues by the estimated "true" ones does not give a satisfactory answer to our problem. Indeed, the Marčenko-Pastur equation only describes the spectrum of eigenvalues of large sample covariance matrices but does not yield any information about the *eigenvectors* of $\mathbf{E}$. In fact, except for some work by Jack Silverstein around 1990 [159, 160], most RMT results about sample covariance matrices were focused on the eigenvalues, as discussed above. The first fundamental result on the eigenvectors of $\mathbf{E}$ was obtained in [144] in the special case of the spiked covariance matrix model, but is somehow disappointing for inference purposes. Indeed, Paul noticed that outliers' eigenvectors obey a cone concentration phenomenon with respect to the true eigenvectors whereas all other ones retain very little information [144]. Differently said, the eigenvectors of $\mathbf{E}$ are not consistent estimators of the eigenvectors of $\mathbf{C}$ in the high-dimensional framework. A few years later, these observations were generalized to general population covariance matrices $\mathbf{C}$ [23, 40, 43, 113, 136]. When dealing with the estimation of $\mathbf{C}$, information about eigenvectors has to be taken into account somehow in the inference problem. Clearly, the above "eigenvalue substitution" method cannot be correct as it proposes to take the best estimates of the eigenvalues of $\mathbf{C}$ but in an unknown eigenvalue basis. Consequently, a different class of estimators flourished very recently that we shall refer to as *rotational invariant estimators*[4] (RIE) [40, 43, 113]. In this particular class of estimators, the main assumption is that any estimator $\Xi$ of $\mathbf{C}$ must share the same eigenvectors as $\mathbf{E}$ itself. This hypothesis has a very intuitive interpretation in practice as it amounts to posit that one has no prior insights on the structure of $\mathbf{C}$, i.e. on the particular directions in which the eigenvectors of $\mathbf{C}$ must point. It is easy to see that the linear shrinkage estimator (2.1.9) falls into this class of estimators. Compared to the aforementioned RMT-based methods, RIE explicitly uses the information on the eigenvectors of $\mathbf{E}$, in particular their average overlap with the true eigenvectors. It turns out that one can actually obtain an optimal estimator of $\mathbf{C}$ in the LDL for any general population covariance matrix $\mathbf{C}$ [43]. Note that the optimal estimator is in perfect agreement with Stein's paradox, that is to say, the optimal cleaning recipe takes into account about the information

[4] This is sometimes called rotation-equivariant estimators

of all eigenvectors and all eigenvalues of $\mathbf{E}$. The conclusion is therefore that combining all the information's about $\mathbf{E}$ always provide more accurate prediction than any method that handles the parameters separately within the modern era of "Big Data". We summarize the above long journey concerning the estimation of large sample covariance matrices in Figure 2.1.2, which can be seen as a thumbnail picture of the present thesis. Note that a very recent work [136] attempts to incorporate prior information on the true components. While it remains unclear how to use this framework for the estimation of correlation, this may allows one to construct "optimal" non-rotational invariant estimators. We shall address this issue at the end of this part.

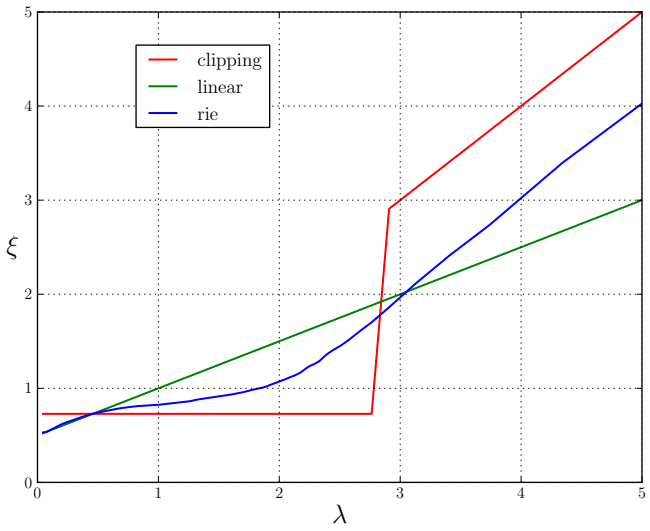



FIGURE 2.1.2. (Color online). Three shrinkage transformations: "cleaned" eigenvalues on the y-axis as a function of the sample eigenvalues (see Chapter 9 for more details). This figure is a quick summary the evolution of shrinkage estimators starting with the linear method (green), then the heuristic eigenvalues clipping method (red) to the optimal RIE (blue).

## 2.2 Outline and main contributions

Our aim is to review several Random Matrix Theory (RMT) results that take advantage of the high-dimensionality of the problem to estimate covariance matrices consistently, spanning nearly fifty years of research from the result of Marčenko and Pastur [123] to the very recent "local" optimal RIE for general population covariance matrices [43]. We emphasize that this thesis is not intended to provide detailed proofs (in the mathematical sense) but we will include references to this mathematical literature as often as possible for those who might be interested.

In Chapter 3, we begin with a detailed but still incomplete introduction to RMT and some of the analytical methods available to study the behavior of large random matrices in the asymp-

totic regime. In fact, most of the computations in Chapter 3 will be performed under very general model of random matrices and will be used throughout the following. The first method is arguably the most frequently used in the Physics literature known as the Coulomb gas analogy [36]. This is particularly useful to deal with invariant ensembles, leading to Boltzmann-like weights that allows one to recover very easily well-known results such as Wigner's semicircle law [189] or Marčenko-Pastur density [123]. This is the main purpose of Section 3.1.2. The second method is Voiculescu's *free probability theory* which was originally proposed in 1985 to understand a special class of von Neumann algebras through the concept of *freeness* [181]. Loosely speaking, two matrices $A$ and $B$ are mutually free if their eigenbasis are related to one another by a random rotation, or said differently if the eigenvectors of $A$ and $B$ are almost surely orthogonal. Voiculescu discovered in 1991 [182] that some random matrices do satisfy asymptotically the freeness relation, which considerably influenced RMT. We present in Section 3.1.3 a precise definition of the concept of freeness and then provide some applications for the computations of the spectral density of a large class of random matrices. In Section 3.1.4, we present a more formal tool known as the Replica method in statistical physics of disordered systems [134]. While being less rigorous, this method turns out to be very powerful to compute the average behavior of large complex systems (see [137] for a recent review). In our case, we shall see how this method allows us to compute the *resolvent* of a large class of random matrices which will be especially useful to deal with the statistics of eigenvectors. This section on the replica analysis lead to the article [40] with Romain Allez, Jean-Philippe Bouchaud and Marc Potters.

In Chapters 4 and 5, we study in details the different properties of large sample covariance matrices. Chapter 4 is dedicated to the statistics of the eigenvalues of $\mathbf{E}$, and in particular we propose a very simple derivation of the Marčenko-Pastur equation using tools from free probability theory. Then, we review different properties that we can learn about $\mathbf{C}$ using $\mathbf{E}$ such as the moment generating functions, or the edges of the support of the spectral density of $\mathbf{E}$. We discuss the properties of the edges of the distribution for finite $N$ and also the outliers. While most of the results are now well known in the RMT community, we provide in Sections 4.2.1, 4.2.2 and 4.2.4 some interesting properties of the spectrum of $\mathbf{E}$ that we have not seen in classical textbooks that dealt with Marčenko and Pastur equation.

In Chapter 5, we focus the recent results concerning the eigenvectors of $\mathbf{E}$ for a general $\mathbf{C}$. We distinguish two different cases. Most of these results come from different articles [40], [43] and [41] and can be distinguished in two different cases. The first one is the angle between the true and estimated eigenvectors and we shall see that the initial results of [144] hold for a general $\mathbf{C}$. The second case is the angle between two *independent* sample eigenvectors, a result that allow one to infer interesting properties about the structure of $\mathbf{C}$.

After these three relatively technical sections, we then turn on the main theme of this thesis which is the estimation of large sample covariance matrices. In Chapter 6, we formalize the Bayesian method for covariance matrices. We then present the class of conjugate prior from which we re-obtain the linear shrinkage (2.1.9) initially derived by Haff [87]. Next, we consider the class of Boltzmann-type, rotational invariant prior distributions. This Bayesian framework is actually the very origin of this thesis [39, 152] and we show here that we can relate the Bayes optimal estimator with the least squares optimal oracle estimator of $\mathbf{C}$.

The so-called oracle estimator is the main quantity of interest in the following Chapter 7. In particular, we show that this estimator converges to a limiting and – remarkably – fully observable function in the limit of large dimension using the results on eigenvectors obtained in

Chapter 5. Even if the final formula is not new [113], we extended this result to case where the spectrum of $\mathbf{E}$ contains a finite number of outliers. Moreover, we also highlight that there exists an optimal estimator of large population covariance C inside the class of RIEs that depends only on observable variables even at finite $N$. These two non trivial extensions are based on a work in preparation with Antti Knowles [43]. Hence, we shall only sketch the main arguments in this thesis. The rest of the Chapter 7 is dedicated to some theoretical and numerical applications of the optimal RIE. We also discuss about the optimal RIE of the general model of free multiplication of Chapter 3.1.3. This comes from a collaboration with Romain Allez, Jean-Philippe Bouchaud and Marc Potters [40].

Chapter 8 concerns the applications of the optimal RIE for Markowitz optimal portfolio. In particular, we characterize explicitly, under some technical assumptions, the danger of using the sample covariance matrix $\mathbf{E}$ in a large scale and out-of-sample framework. As alluded to above, we shall see that if $\mathbf{E}$ has no exact zero mode (i.e. when $q = N/T < 1$), the realized risk associated to this "naive" estimator overestimates the true risk by a factor $(1-q)^{-1}$. Also, we shall see that the best we can do in order to minimize the out-of-sample risk is actually given by the optimal RIE of the Chapter 7. We will also determine the estimated and realized risk associated to the optimal RIE in a very special case. We believe that it sheds light on the advantage of using this estimator compared to the sample estimator $\mathbf{E}$. Several alternative cleaning "recipes", proposed in previous work, are also reviewed in this Chapter.

Finally, Chapter 9 contains empirical results using real financial data sets. We give further evidence that using a correctly regularized estimator of $\mathbf{C}$ is highly recommended in real life situations. Moreover, we discuss about the implementation of the optimal RIE in the presence of finite size effects, to wit, when $N$ is large but finite. This chapter thus extends the simple cleaning recipe we proposed in [42]. Furthermore, we give some concrete applications of the two-sample test introduced in a recent paper [41] with Jean-Philippe Bouchaud and Marc Potters.

The appendices contain auxiliary results which are mentioned in this work. The first appendix copes with the so-called Harish-Chandra–Itzykson-Zuber (HCIZ) integral which routinely appears in calculations involving sums or products of free random matrices. The HCIZ is an integral over the group of orthogonal matrices for which explicit and analytical results are scarce. We give a complete derivation of the limiting behavior of this integral when one matrix has low rank and then the general case which was presented in the paper [38] written with Jean-Philippe Bouchaud, Satya Majumdar and Marc Potters. The second appendix is a reminder on some results of linear algebra which are particularly useful for the study of eigenvectors. The third appendix is another analytical tool in RMT to establish self-consistent equations for the resolvent (or the Stieltjes transform) of large random matrices. This technique is very convenient when working with independent entries and it provides a nice illustration of the Central Limit Theorem for random matrices. However, the formalism is not as synthetic as the method provided in Chapter 3 but is now standard in the RMT literature, which is why we relegate its presentation to an appendix.

# Chapter 3

# Random Matrix Theory: overview and analytical tools

## 3.1 RMT in a nutshell

**3.1.1. Large dimensional random matrices.** As announced in the introduction, the main analytical tool that we shall review in this article is Random Matrix Theory (RMT). In order to be as self-contained as possible, we recall in this section some of the basic results and techniques of RMT. The study of random matrices began with the work of Wishart in 1928, who was interested in the distribution of the so-called empirical (or sample) covariance matrices, which ultimately lead to the Marčenko-Pastur distribution in 1967. RMT was also introduced by Wigner in the 1950's as a statistical model for the energy levels of heavy nuclei, and lead to the well-known Wigner semi-circle distribution, as well as Dyson's Brownian motion (see e.g. [186], [1] for comprehensive reviews). Branching off from these early physical and statistical applications, RMT has become a vibrant research field of its own, with scores of beautiful results in the last decades – one of the most striking being the discovery of the Tracy-Widom distribution of extreme eigenvalues, which turns out to be related to a large number of topics in statistical mechanics and probability theory [62, 122]. Here, we will only consider the results of RMT that pertain to statistical inference, and leave aside many topics – see e.g. [1], [8], [171], [176], [14] or [57] for more detailed and rigorous introductions to RMT. We will also restrict to square, symmetric correlation matrices, even though the more general problem of rectangular correlation matrices (measuring the correlations between $M$ input variables and $N$ output variables) is also extremely interesting. This problem leads to the so-called Canonical Component Analysis [91] and can be dealt with the Singular Value Decomposition, for which partial results are available, see e.g. [28, 184].

We begin with a formal definition of "large" random matrices. A common assumption in RMT is that the matrix under scrutiny is of infinite size. However, this is obviously not a realistic assumption for practical problems where one rather deals with *large* but *finite* $N$ dimensional matrices. Nonetheless, we shall see that working in the $N \to \infty$ limit leads to very precise approximations of the properties of large but finite matrices. More precisely, it is well known that probability distributions describing the fluctuations of macroscopic observables often converge to limiting laws in the limit of large sizes. Hence, we expect that the statistical properties (say the distribution of eigenvalues) of a random matrix $\mathbf{M}$ of dimension $N$ shows,

to a certain extent, a deterministic or self-averaging behavior[1] when the dimension $N$ goes to infinity. These deterministic features can be used to characterize the matrix under scrutiny, provided it is large enough. This is why we consider the limit $N \to \infty$ from now on.

The limiting behavior of "large" random matrices is in fact at the heart of RMT, which predicts that infinite dimensional matrices do display *universal* features, both at the macroscopic and at the microscopic levels. To be more precise, we define a $N \times N$ random matrix[2] $\mathbf{M}$ with a certain probability measure $\mathcal{P}_\beta(\mathbf{M})$, where $\beta$ is the Dyson's threefold way index and specifies the symmetry properties of the ensemble ($\beta = 1$ for Orthogonal, $\beta = 2$ for Unitary and $\beta = 4$ for Symplectic ensembles). A property is said to be *universal* if it does not depend on the specific probability measure $\mathcal{P}_\beta(\mathbf{M})$. One well known example of universality pertains to the distribution of the distance $s$ between two successive eigenvalues (see [172] for an extended discussion).

The ensemble most relevant for our purpose is the Orthogonal one, which deals with real symmetrical matrices. In this case, the matrix $\mathbf{M}$ is said to be rotationally invariant if the probability is invariant under the transformation $\mathbf{M} \to \mathbf{\Omega M \Omega}^\dagger$ for any matrix $\mathbf{\Omega}$ belonging to the Orthogonal group $\mathbf{O}(N)$, i.e. $\mathcal{P}_\beta(\mathbf{M}) = \mathcal{P}_\beta(\mathbf{\Omega M \Omega}^\dagger), \forall \mathbf{\Omega} \in \mathbf{O}(N)$. A typical example of invariant measure in the physics literature is that $\mathcal{P}_\beta(\mathbf{M})$ is of the form of a Boltzmann distribution:

$$\mathcal{P}_\beta(\mathbf{M})\mathcal{D}\mathbf{M} \propto e^{-\frac{\beta N}{2}\mathrm{Tr}V(\mathbf{M})}\mathcal{D}\mathbf{M} \tag{3.1.1}$$

with $V$ the so called *potential* function and $\mathcal{D}\mathbf{M} = \prod_{i=1}^N d\mathbf{M}_{ii} \prod_{i<j}^N d\mathbf{M}_{ij}$ denotes the (Lebesgue) flat measure. The rotational invariant property is evident since the above parametrization only involves the trace of powers of $\mathbf{M}$. Already at this stage, it is interesting to notice that the distribution (3.1.1) can alternatively be rewritten in terms of the eigenvalues and eigenvectors of $\mathbf{M}$ as:

$$\mathcal{P}_\beta(\mathbf{M})\mathcal{D}\mathbf{M} \propto e^{-\frac{\beta N}{2}\sum_{i=1}^N V(\nu_i)} \prod_{i<j}^N |\nu_i - \nu_j|^\beta \Big(\prod_{i=1}^N \mathrm{d}\nu_i\Big)\Big(\mathrm{d}\Omega\Big), \tag{3.1.2}$$

where the Vandermonde determinant ($\prod|\nu_i - \nu_j|^\beta$)comes from the change of variables (from the $\mathbf{M}_{ij}$ to the $\nu_i$ and $\Omega_{ij}$). This representation is extremely useful, as will be illustrated below.

What kind of universal properties can be of interest in practice? Let us consider a standard problem in multivariate statistics. Suppose that we have a very large dataset with correlated variables. A common technique to deal with this large dataset is to reduce the dimension of the problem using for instance a *principal component analysis* (PCA), obtained by diagonalizing the covariance matrix of the different variables. But one can wonder whether the obtained eigenvalues $\nu_i$ and their associated eigenvectors are reliable or not (in a statistical sense). Hence, the characterization of eigenvalues (and eigenvectors) is an example of features that one would like to know a priori. In that respect, RMT provided (and continues to provide) many groundbreaking results on the eigenvalues and the eigenvectors of matrices belonging to specific invariant ensembles (Unitary, Orthogonal and Symplectic). The distribution of the eigenvalues $\{\nu_i\} : i = \{1, \ldots, N\}\}$ can be characterized through the *Empirical Spectral Distribution* (ESD) (also known as the "Eigenvalue Distribution"):

$$\rho^N_{\mathbf{M}}(x) = \frac{1}{N}\sum_{i=1}^N \delta(x - \nu_i) \tag{3.1.3}$$

[1] i.e. independent of the specific realization of the matrix itself

[2] Boldface letters will refer throughout this paper to matrices.

with $\delta$ the Dirac delta function. Note that the symmetry of the considered matrices ensures that the eigenvalues of $\mathbf{M}$ are defined on the real line (complex eigenvalues are beyond the scope of this thesis, but see [**?**, 14, 57] for more on this). One of the most important property of large random matrices is that one expects the ESD to converge (almost surely in many cases) to a unique and *deterministic* limit $\rho_{\mathbf{M}}^N \to \rho_{\mathbf{M}}$ as $N \to \infty$. Note that it is common to refer to this deterministic density function $\rho_{\mathbf{M}}$ as the *Limiting Spectral Density* (LSD), or else the "Eigenvalue Spectrum" of the matrix. An appealing feature of RMT is the predicted *self-averaging* (sometimes call *ergodicity* or *concentration*) property of the LSD: when the dimension $N$ becomes very large, a single sample of $\mathbf{M}$ spans the whole eigenvalue density function, independently of the specific realization of $\mathbf{M}$. The consequence of this self-averaging property is that we can replace the computation of the ESD (3.1.3) for a specific $\mathbf{M}$ by the average according to the probability measure of $\mathbf{M}$ (e.g. over the measure (3.1.1)):

$$\rho_{\mathbf{M}}(x) = \lim_{N\to\infty} \rho_{\mathbf{M}}^N(x), \quad \text{with} \quad \rho_{\mathbf{M}}^N(x) = \left\langle \frac{1}{N}\sum_{i=1}^{N} \delta(x-\nu_i) \right\rangle_{\mathbf{M}}. \tag{3.1.4}$$

For real life data-sets, it is often useful to distinguish the eigenvalues that lie within the spectrum of $\rho_{\mathbf{M}}$ from those that are well separated from it. We will refer to the first category as the **bulk** of the eigenvalues with a slight abuse of notation. We will call the second type of eigenvalues **outliers** or **spikes**. Throughout this work, we assume the LSD that describes the bulk of $\rho_{\mathbf{M}}$ to be a non-negative continuous function, defined on an unique compact support – denoted $\mathrm{supp}[\rho_{\mathbf{M}}]$ – meaning that $\mathrm{supp}[\rho_{\mathbf{M}}]$ consists of a single "bulk" component (often called the *one-cut* assumption). Moreover, we allow the presence of a finite number $r \ll N$ of outliers, which are of crucial importance in many fields. Throughout this chapter, we shall denote by $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_N$ the eigenvalues of $\mathbf{M}$. We furthermore define the associated eigenvectors by $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$. For $N$ that goes to infinity, it is often convenient to index the eigenvectors by their corresponding eigenvalues, i.e. $\mathbf{w}_i \equiv \mathbf{w}_{\nu_i}$ for any integer $1 \leq i \leq N$, and this is the convention that we adopt henceforth.

*Various RMT transforms.* We end this section with an overview of different transforms that appear in the RMT literature. These transforms are especially useful to study the spectral properties of random matrices in the limit of large dimension, and to deal with sums and products of random matrices.

**Resolvent and Stieltjes transform.** We start with the resolvent of $\mathbf{M}$ which is defined as[3]

$$\mathbf{G}_{\mathbf{M}}(z) := (z\mathbf{I}_N - \mathbf{M})^{-1}, \tag{3.1.5}$$

with $z := x - \mathrm{i}\eta \in \mathbb{C}^-$, where $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Im}(z) < 0\}$. We define accordingly $\mathbb{C}^+ = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$. This quantity displays several interesting properties, making it the relevant object to manipulate. First, it is a continuous function of $z$ and is easy to differentiate (compared to working directly on the ESD), providing a well-defined tool for mathematical analysis. Furthermore, it contains the complete information about the eigenvalues $\{\nu_i\}$ *and* the eigenvectors $\{\mathbf{w}_i\}$ since it can be rewritten as:

$$\mathbf{G}_{\mathbf{M}}(z) = \sum_{i=1}^{N} \frac{\mathbf{w}_i \mathbf{w}_i^*}{z - \nu_i}. \tag{3.1.6}$$

[3] Note that in the mathematical and statistical literature, the resolvent differs from ours by a minus sign.

It is easy to see that the number of singularities of the resolvent is equal to the number of eigenvalues of $\mathbf{M}$. Suppose that $z \to \nu_i$ for any $i \in [\![N]\!]$, then the residue of the pole defines a projection operator onto the eigenspace associated to the eigenvalues $\nu_i$. We will show in chapter 5 how this property can be used to study the statistics of the eigenvectors.

While the statistics of the eigenvectors is an interesting and non-trivial subject in itself, we focus for now on the statistics of the eigenvalues through the ESD (3.1.4). For this aim, we define the normalized trace of Eq. (3.1.5) as

$$\mathfrak{g}_{\mathbf{M}}^{N}(z) := \frac{1}{N}\mathrm{Tr}\left[\mathbf{G}_{\mathbf{M}}(z)\right], \tag{3.1.7}$$

We shall skip the index $_{\mathbf{M}}$ as soon as there is no confusion about the matrix we are dealing with. In the limit of large dimension, one has

$$\mathfrak{g}^{N}(z) \underset{N\to\infty}{\sim} \mathfrak{g}(z), \qquad \mathfrak{g}(z) := \int \frac{\rho(u)}{z-u}\mathrm{d}u. \tag{3.1.8}$$

which is known as the *Stieltjes* (or *Cauchy*) transform of $\rho$. The Stieltjes transform has a lot of appealing properties. For instance, if the density function $\rho$ does not contain Dirac masses, then this is the unique solution of the so-called *Riemann-Hilbert* problem, i.e :

(i) $\mathfrak{g}(z)$ is analytic in $\mathbb{C}^+$ except on its branch cut on the real axis inside $\mathrm{supp}[\rho_{\mathbf{M}}]$;

(ii) $\lim_{|z|\to\infty} z\mathfrak{g}(z) = 1$;

(iii) $\mathfrak{g}(z)$ is real for $z \in \mathbb{R}\backslash\,\mathrm{supp}[\rho_{\mathbf{M}}]$;

(iv) When near the branch cut, two different values for $\mathfrak{g}(z)$ are possible, depending on whether the cut is approached from above or from below, i.e.:

$$\lim_{\eta\to 0^+} \mathfrak{g}(x \pm \mathrm{i}\eta) = \mathfrak{h}(x) \mp i\pi\rho(x), \qquad x \in \mathrm{supp}[\rho] \text{ and } \rho(x) \in \mathbb{R}^+, \tag{3.1.9}$$

where the function $\mathfrak{h}$ denotes the *Hilbert* transform of $\rho$ defined by

$$\mathfrak{h}(x) := \fint_{\mathrm{supp}[\rho]} \frac{\rho(u)}{x-u}\mathrm{d}u \tag{3.1.10}$$

with $\fint$ denoting Cauchy's principal value.

It is now immediate to see that if one knows $\mathfrak{g}(z)$ in the complex plane, the density $\rho$ can be retrieved by inverting the last property of the Riemann-Hilbert problem:

$$\rho(x) \equiv \frac{1}{\pi}\lim_{\eta\to 0^+} \mathrm{Im}(\mathfrak{g}(x-\mathrm{i}\eta)), \qquad x \in \mathrm{supp}[\rho]. \tag{3.1.11}$$

The continuous limit of $\mathfrak{g}(z)$ in the large $N$ limit thus allows to investigate the distribution of the eigenvalues that lie in the bulk component.

Another interesting property is to study the asymptotic expansion of $\mathfrak{g}(z)$ when $z$ is large (and outside of Supp$[\rho]$). Expanding $\mathfrak{g}(z)$ in powers of $z^{-1}$ yields:

$$\mathfrak{g}(z) \underset{z\to\infty}{=} \frac{1}{z}\int \rho(u)\sum_{k=0}^{\infty}\left(\frac{u}{z}\right)^k \mathrm{d}u.$$

To leading order, we get, in agreement with property (ii) above:

$$\mathfrak{g}(z) \sim \frac{1}{z}\int \rho(u)\mathrm{d}u \equiv \frac{1}{z},$$

where the last equality comes from the fact that the ESD is normalized to unity. The other terms of the expansion are also of particular interest. Indeed, we see that

$$\mathfrak{g}(z) \underset{z\to\infty}{=} \frac{1}{z} + \frac{1}{N}\sum_{k=1}^{\infty}\frac{\mathrm{Tr}\mathbf{M}^k}{z^{k+1}} \equiv \frac{1}{z} + \sum_{k=1}^{\infty}\frac{\varphi(\mathbf{M}^k)}{z^{k+1}}, \tag{3.1.12}$$

where we defined the $k$-th moment of the ESD by $\varphi(\mathbf{M}^k) := N^{-1}\mathrm{Tr}\mathbf{M}^k$. We see that the Stieltjes transform is related to the *moment generating function* of the random matrix $\mathbf{M}$. This is another illustration of the fact that the Stieltjes transform contains the complete information about the eigenvalues density. Inversely, if one can measure the moments of the eigenvalues distribution, it is possible reconstruct a parametric eigenvalues density function that matches the empirical data. This nice property is an important feature of the Stieltjes transform for statistical inference purposes. Note that we will sometimes abbreviate $\varphi(\mathbf{M}^k) \equiv \varphi_k$ when there is no confusion about the matrix we are studying.

Last but not least, it is easy to check the following scaling property

$$\mathfrak{g}_{a\mathbf{M}}(z) = \frac{1}{a}\mathfrak{g}_{\mathbf{M}}\left(\frac{z}{a}\right), \tag{3.1.13}$$

for any $a \in \mathbb{R}\backslash\{0\}$. Moreover, suppose that $\mathbf{M}$ is invertible, then using (3.1.7) we also have

$$z\mathfrak{g}_{\mathbf{M}}(z) + \frac{1}{z}\mathfrak{g}_{\mathbf{M}^{-1}}\left(\frac{1}{z}\right) = 1, \tag{3.1.14}$$

so that we are able to compute the Stieltjes transform of $\mathbf{M}^{-1}$ given the Stieltjes transform of $\mathbf{M}$.

**Blue function and $\mathcal{R}$-transform.** There are many other useful RMT transforms, some that will turn out to be important in the next chapter. We begin with the *free cumulant* generating function which is known as the $\mathcal{R}$-transform in the literature [164, 176, 183]. To define this quantity, it is convenient to introduce the functional inverse of the Stieltjes transform, also known as the *Blue* transform [196]

$$\mathcal{B}(\mathfrak{g}(z)) = z, \tag{3.1.15}$$

and the *$\mathcal{R}$-transform* is simply defined by

$$\mathcal{R}(\omega) = \mathcal{B}(\omega) - \frac{1}{\omega}. \tag{3.1.16}$$

Note that one may deduce from (3.1.13) the following property

$$\mathcal{R}_{a\mathbf{M}}(\omega) = a\mathcal{R}_{\mathbf{M}}(a\omega), \tag{3.1.17}$$

for any $a \in \mathbb{R}$. One very nice property is that the $\mathcal{R}$-transform admits a Taylor expansion in the limit $\omega \to 0$. Indeed, by plugging $\omega = \mathfrak{g}(z)$ into Eq. (3.1.16), we obtain the formula

$$\mathcal{R}(\mathfrak{g}(z)) + \frac{1}{\mathfrak{g}(z)} = z. \tag{3.1.18}$$

Then, one can find after expanding the Stieltjes transform in powers of $z^{-1}$ that $\mathcal{R}(\omega)$ can be expanded as

$$\mathcal{R}(\omega) = \sum_{\ell=1}^{\infty} \kappa_\ell(\mathbf{M})\omega^{\ell-1} \tag{3.1.19}$$

where the sequence $\{\kappa_\ell\}_{\ell\geq 0}$ denotes the *free cumulant* of order $\ell$ which are expressed as a function of the moments of the matrix. For completeness, we give the first four free cumulants:

$$\begin{aligned}
\kappa_1 &= \varphi_1 \\
\kappa_2 &= \varphi_2 - \varphi_1^2 \\
\kappa_3 &= \varphi_3 - 3\varphi_2\varphi_1 + 2\varphi_1^3 \\
\kappa_4 &= \varphi_4 - 4\varphi_3\varphi_1 - 2\varphi_2^2 + 10\varphi_2\varphi_1^2 - 5\varphi_1^4.
\end{aligned} \tag{3.1.20}$$

Note that the first three cumulants are equivalent to the 'standard' cumulants of ordinary random variables and only differ from $\ell \geqslant 4$. Note for example that when $\varphi_1 = 0$, one finds $\kappa_4 = \varphi_4 - 2\varphi_2^2$, whereas the standard kurtosis would read $\varphi_4 - 3\varphi_2^2$. It will turn out that the free cumulants of the sum of independent – in a sense specified below – random matrices are given by the sum of the cumulants of these random matrices, i.e. $\kappa_\ell(\mathbf{M}) = \kappa_\ell(\mathbf{A}) + \kappa_\ell(\mathbf{B})$, see section 3.1.3 below.

**Moment generating function and S-transform.** The moment generating function of the LSD $\rho$ is obtained by considering

$$\mathcal{T}(z) := z\mathfrak{g}(z) - 1 = \int \frac{\mathrm{d}u\rho(u)u}{z-u}, \tag{3.1.21}$$

frequently known as the $\mathcal{T}$ (or sometimes $\eta$ [176]) transform [23]. Indeed, by taking $z \to \infty$, one readily finds

$$\mathcal{T}_{\mathbf{M}}(z) = \sum_{k=1}^{\infty} \frac{\varphi(\mathbf{M}^k)}{z^k}. \tag{3.1.22}$$

We can then introduce the so-called $\mathcal{S}$-transform as [183]:

$$\mathcal{S}(\omega) := \frac{\omega+1}{\omega\mathcal{T}^{-1}(\omega)} \tag{3.1.23}$$

where $\mathcal{T}^{-1}(\omega)$ is the functional inverse of the $\mathcal{T}$-transform. Using the series expansion of $\mathcal{T}_{\mathbf{M}}(z)$ in powers of $z^{-1}$ and Eq. (3.1.20), one finds that the $\mathcal{S}$-transform also admits a Taylor series which reads:

$$\begin{aligned}
\mathcal{S}_{\mathbf{M}}(\omega) &= \frac{1}{\varphi_1} + \frac{\omega}{\varphi_1^3}(\varphi_1^2 - \varphi_2) + \frac{\omega^2}{\varphi_1^5}(2\varphi_2^2 - \varphi_2\varphi_1^2 - \varphi_3\varphi_1) + \mathcal{O}(\omega^3) \\
&= \frac{1}{\kappa_1} - \frac{\kappa_2}{\kappa_1^3}\omega + \frac{2\kappa_2^2 - \kappa_1\kappa_3}{\kappa_1^5}\omega^2 + \mathcal{O}(\omega^3).
\end{aligned} \tag{3.1.24}$$

From this last equation, it is not hard to see that the $\mathcal{S}$-transform of a matrix $\mathbf{M}$ which has a zero trace is ill-defined. Hence, the $\mathcal{S}$-transform of a Wigner matrix does not make sense, but it will be very useful when manipulating positive definite covariance matrices (see Section 3.1.3)

Note finally that there exists a relation between the $\mathcal{R}$-transform and the $\mathcal{S}$-transform

$$\mathcal{R}(\omega) = \frac{1}{\mathcal{S}(\omega\mathcal{R}(\omega))}, \qquad \mathcal{S}(\omega) = \frac{1}{\mathcal{R}(\omega\mathcal{S}(\omega))} \tag{3.1.25}$$

which allows one to deduce $\mathcal{R}(z)$ from $\mathcal{S}(z)$ and vice versa. Other properties on the $\mathcal{R}$ and $\mathcal{S}$ transforms can be found e.g. in [46].

Let us show the second equality of (3.1.25) for the sake of completeness. The derivation of the first identity is similar and we omit details. Using (3.1.16) and (3.1.23), one obtains

$$\mathcal{R}(\omega\mathcal{S}(\omega)) = \mathcal{B}\left(\frac{\omega+1}{\mathcal{T}^{-1}(\omega)}\right) - \frac{\mathcal{T}^{-1}(\omega)}{\omega+1}. \tag{3.1.26}$$

Next, by setting $z = \mathcal{T}^{-1}(\omega)$, we can rewrite (3.1.21) as

$$\frac{\omega+1}{\mathcal{T}^{-1}(\omega)} = \mathfrak{g}\big(\mathcal{T}^{-1}(\omega)\big). \tag{3.1.27}$$

Hence, we conclude that

$$\mathcal{R}(\omega\mathcal{S}(\omega)) = \mathcal{T}^{-1}(\omega) - \frac{1}{\mathfrak{g}\big(\mathcal{T}^{-1}(\omega)\big)} = \frac{\omega}{\mathfrak{g}\big(\mathcal{T}^{-1}(\omega)\big)}. \tag{3.1.28}$$

The conclusion then follows from (3.1.27).

**3.1.2. Coulomb gas analogy.** There exists several techniques to compute the limiting value of the Stieltjes transform: (i) Coulomb gas methods, (ii) method of moments, (iii) Feynman diagrammatic expansion, (iv) Dyson's Brownian motion, (v) Replicas, (vi) Free probability, (vii) recursion formulas, (viii) supersymmetry... We devote the rest of this section to provide the reader with a brief introduction to (i), (v) and (vi). Dyson's Brownian motion (iv) and the recursion method (vii) are mentioned in appendices B.4 and 12.2.2. We refer to [8] for the moment methods (ii), to [36, 44] for Feynman diagrams (iii) or to [**?**] and references therein for summetry applied to RMT. Again, we emphasize that this presentation is not intended to be rigorous in a mathematical sense, and we refer to standard RMT textbooks such as [1, 8, 57, 171] for more details.

We begin with the *Coulomb gas analogy* that, loosely speaking, consists in considering the eigenvalues of $\mathbf{M}$ as the positions of fictitious charged particles, repelling each other via a 2-d Coulomb (logarithmic) potential (see [131] for a self-contained introduction or to e.g. [36, 62, 63] for concrete applications). We shall highlight in this section the strong link between the potential function and the Stieltjes transform $\mathfrak{g}(z)$ whenever the probability measure over the matrix ensemble is rotationally invariant, i.e. of the form Eq. (3.1.1).

*Stieltjes transform and potential function.* First, we write from (3.1.1) the *partition function* of the model as

$$\mathcal{Z} \propto \int e^{-\frac{\beta N}{2}\mathrm{Tr}V(\mathbf{M})}\mathcal{D}\mathbf{M},$$

and this can be used as a starting point to obtain the LSD – or rather its Stieltjes transform – using a saddle point method. This relation has first been obtained in the seminal paper of Brézin-Itzykson-Parisi-Zuber [36] and we repeat here the main idea of the derivation (see also [200, Section 2.1]). Let us first express the partition function in terms of the eigenvalues and eigenvectors of $\mathbf{M}$, using (3.1.2):

$$\mathcal{Z} \propto \int \Big(\prod_{i=1}^{N} \mathrm{d}\nu_i\Big) \exp\left\{-N\sum_{i=1}^{N}\left[V(\nu_i) - \frac{\beta}{2N}\sum_{i\neq j}\log|\nu_i-\nu_j|\right]\right\},$$

up to a constant factor that comes from integrating over the Haar measure $d\Omega$. It is then customary to introduce the *action* $S(\{\nu_i\}) \equiv S(\nu_1, \nu_2, \ldots, \nu_N)$ such that we can rewrite the partition function as:

$$\mathcal{Z} \propto \int \prod_{i=1}^{N} d\nu_i e^{-N^2 S(\{\nu_i\})} \quad \text{with} \quad S(\{\nu_i\}) = \frac{1}{N}\sum_{i=1}^{N} V(\nu_i) - \frac{\beta}{2N^2}\sum_{i\neq j} \log|\nu_i - \nu_j|, \tag{3.1.29}$$

Note that the action is normalized so that its large $N$ limit is of order 1. The eigenvalues can be seen as a thermal gas of one-dimensional particles in an external potential $V(z)$ and subject to a (logarithmic) "electrostatic" repulsive interaction: this is the Coulomb gas analogy. At thermal equilibrium, the eigenvalues typically gather in potential well(s), but cannot accumulate near the minimum due to the repulsive force, which keeps them at distance of order $\mathcal{O}(N^{-1})$. For instance, if we take a quadratic potential function $V(x) = x^2/2$, then all the particles tend to gather around zero as it is shown in the Fig. 3.1.1. We recall that we consider only densities which are defined on an unique compact support (*one-cut* assumption) and we thus require that the fictitious particles evolve in a confining convex potential $V(z)$. The class of potential function that we consider is such that its derivative gives a Laurent polynomial, i.e., $V'(z) = \sum_k c_k z^k$ with $k$ integers that can be negative. Since we can always rewrite $V'(z) = z^{-\ell}P(z)$, with the "order" $\ell$ the lowest (negative) power of $V'(z)$ and $P(z)$ a polynomial, we define by $d$ the "degree" of $V'(z)$ which corresponds to the degree of $P(z)$. In particular, if $V'(z)$ is a polynomial, then $\ell = 0$.

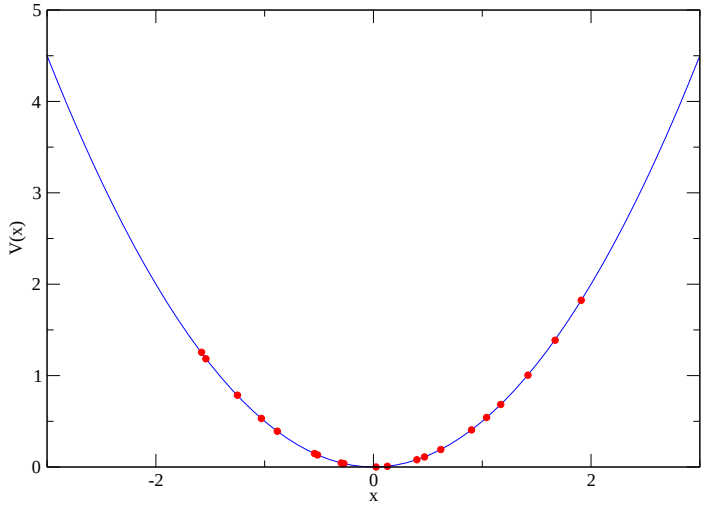



FIGURE 3.1.1. Typical configuration of a repulsive Coulomb gas with $N = 20$ particles (red dots) in the potential $V(x) = x^2/2$ as a function of x.

In the large $N$ limit, the integral over eigenvalues can be computed by the saddle-point

method which yields the following "force equilibrium" condition:[4]

$$V'(\nu_i) = \frac{\beta}{N} \sum_{j=1;j\neq i}^{N} \frac{1}{\nu_i - \nu_j}, \quad \forall\, i = 1, \dots, N. \tag{3.1.30}$$

It seems hopeless to find the eigenvalues $\{\lambda_i\}$ that solve these $N$ equations. However, we may expect to find the LSD $\rho_{\mathbf{M}}$ in the limit $N \to \infty$, corresponding to configuration of the eigenvalues that satisfies these saddle-point equations. In the case of the one-cut assumption, the result reads [36]:

$$\mathfrak{g}(z) = V'(z) - Q(z)\sqrt{(z-\nu_+)}\sqrt{(z-\nu_-)}, \tag{3.1.31}$$

where $\nu_- < \nu_+$ denote the edges of $\mathrm{supp}[\rho]$ and $Q(z)$ is also a Laurent polynomial with degree $d-1$ and order $\ell$. Therefore, we see that we have $d+1$ unknowns to determine, namely the coefficients of $Q(z)$, $\nu_-$ and $\nu_+$ which are determined using the series expansion (3.1.12). We shall give a detailed illustration of this procedure in Section 3.1.2 below[5].

We observe that as soon as we can characterize the potential function of $V(z)$ that governs the entries of $\mathbf{M}$, we are then able to find the corresponding LSD $\rho_{\mathbf{M}}$. We will show in the rest of this section that this Coulomb gas analogy allows one to retrieve some important laws in RMT.

Let us show how to obtain (3.1.31). In the following we set $\beta = 1$. First, we introduce the normailzed trace of the resolvent $\mathfrak{g}(z)$ in (3.1.30) by multiplying on both sides by $N^{-1}(z-\nu_i)^{-1}$ and summing over all $i$, which yields

$$\frac{1}{N}\sum_{i=1}^{N} \frac{V'(\nu_i)}{z-\nu_i} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1;j\neq i}^{N} \frac{1}{(z-\nu_i)(\nu_i-\nu_j)}. \tag{3.1.32}$$

Notice that this last equation is indeed an analytical function for $z \in \mathbb{C}\backslash\mathrm{Supp}[\rho_{\mathbf{M}}]$. Then, we rewrite the LHS using some algebraic manipulations that leads to

$$\frac{1}{N}\sum_{i=1}^{N} \frac{V'(\nu_i)}{z-\nu_i} = V'(z)\mathfrak{g}(z) - \frac{1}{N}\sum_{i=1}^{N} \frac{V'(z)-V'(\nu_i)}{z-\nu_i},$$

and for the RHS, we obtain

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1;j\neq i}^{N} \frac{1}{(z-\nu_i)(\nu_i-\nu_j)} \equiv \frac{1}{2}\left[\mathfrak{g}^2(z) + \frac{1}{N}\mathfrak{g}'(z)\right].$$

Regrouping these last two equations into the saddle-point equation (3.1.32) gives

$$\frac{1}{2}\left[\mathfrak{g}^2(z) + \frac{1}{N}\mathfrak{g}'(z)\right] = V'(z)\mathfrak{g}(z) - \frac{1}{N}\sum_{i=1}^{N} \frac{V'(z)-V'(\nu_i)}{z-\nu_i}.$$

Since we are interested in the limit of large $N$, we thus have to solve for $\mathfrak{g}(z)$ the following quadratic equation

$$\mathfrak{g}^2(z) - 2V'(z)\mathfrak{g}(z) + \frac{2}{N}\sum_{i=1}^{N} \frac{V'(z)-V'(\nu_i)}{z-\nu_i} = 0. \tag{3.1.33}$$

[4] The reader might wonder why a system in thermal equilibrium ends up being described by simple mechanical equilibrium, as at zero temperature. It turns out that the system is effectively at very low temperatures and that entropy effects are of order $N^{-1}$ compared to interaction effects, see e.g. [63] for a detailed discussion. Entropy effects start playing a role for extended $\beta$ ensembles where $\beta = c/N$ where $c$ is finite, see [4].

[5] In the case of positive definite covariance matrices, we can use the series (4.2.15) that corresponds to the limit $z \to 0$

The most difficult term is the last one because the sum is not explicit. For the sake of simplicity, we consider the case where $V'(z)$ is a polynomial of degree $d > 0$ as the extension to Laurent polynomial, i.e. polynomial with negative powers, is immediate. For $V'(z)$ a polynomial function in $z$, we have that

$$P(z) := \frac{1}{N}\sum_{i=1}^{N} \frac{V'(z) - V'(\nu_i)}{z - \nu_i}$$

is also a polynomial but with a degree $d-1$ whose coefficients can be determined later by the normalization constraint, or by matching some moments. Then, the solution of Eq. (3.1.33) is such that:

$$\mathfrak{g}(z) = V'(z) \pm \sqrt{V'(z)^2 - 2P(z)}.$$

The nice property in the one-cut framework (i.e., a unique compact support for $\rho$) is that the above expression can be simplified to (when $d \geqslant 1$):

$$\mathfrak{g}(z) = V'(z) \pm Q(z)\sqrt{(z-\nu_+)(z-\nu_-)}$$

where $\nu_-$ and $\nu_+$ denote the edges of $\mathrm{supp}[\rho]$ and $Q(z)$ is a polynomial with degree $d-1$ and this gives (3.1.31).

*Wigner's semicircle law.* As a warm-up exercise, we begin with Wigner's semi-circle law [189], one of the most important result in RMT. Note that this result has first been obtained in the case of Gaussian matrix with independent and identically distributed entries (while preserving the symmetry of the matrix). For real entries, we refer to this class of random matrices as the Gaussian Orthogonal Ensemble (GOE). It has been proved, see e.g. [8], that the semi-circle law can be extended to a broader class of random matrices, known as the *Wigner Ensemble* that deals with a matrix $\mathbf{M}$ with independent and identically distributed entries such that:[6]

$$\mathbb{E}\big[\mathbf{M}_{ij}\big] = 0, \quad \text{and} \quad \mathbb{E}\big[\mathbf{M}_{ij}^2\big] = \sigma^2/N. \tag{3.1.34}$$

Let us consider here the specific case of a GOE matrix. For Gaussian entries, it is not hard to see that the associated probability measure $\mathcal{P}_\beta(\mathbf{M})$ is indeed of the Boltzmann type with a potential function $V(\mathbf{M}) = \mathbf{M}^2/2\sigma^2$. From Eq. (3.1.31), we remark that the unknown polynomial $Q(z)$ is simply a constant because the derivative of the potential has degree $d = 1$. To determine this constant, we enforce the property (ii) of the Riemann-Hilbert problem which enable us to get by identification: $Q(z) = 1$, $\nu_\pm = \pm 2\sigma$. We thus finally obtain:

$$\mathfrak{g}_W(z) = \frac{z - \sqrt{z+2\sigma}\sqrt{z-2\sigma}}{2\sigma^2}, \tag{3.1.35}$$

where $\sqrt{\cdot}$ denotes throughout the following the principal square root, that is the non-negative square root of a non-negative real number. Equation (3.1.35) is indeed the Stieltjes transform of Wigner's semi-circle law. Note that it is frequent to see the above result written as

$$\mathfrak{g}_W(z) = \frac{z \pm \sqrt{z^2 - 4\sigma^2}}{2\sigma^2},$$

where the convention "$\pm$" refers to the fact that we have to chose the correct sign such that $\mathfrak{g}(z) \sim z^{-1}$ for large $|z|$ (property (ii) of the Riemann-Hilbert problem). The density function is

[6] The case where the variance of the matrix elements diverge corresponds to *Lévy matrices*, introduced in [54]. For a rigorous approach, we refer the readers to [20]. For recent developments, see [174] .

then retrieved using the inversion formula (3.1.11) that yields the celebrated *Wigner's semicircle law*:

$$\rho_W(x) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - x^2}, \qquad |x| < 2\sigma. \tag{3.1.36}$$

We plot in Fig 3.1.2 the density of the semi-circle and compared with the ESD obtained from a GOE matrix of size $N = 500$. As stated at the beginning of this section, we see that the limiting density agrees well with the ESD of the large but finite size matrix. In fact, one can rigorously estimate the expected difference between the ESD at finite $N$ and the asymptotic LSD for $N = \infty$, which vanishes as $N^{-1/4}$ as soon as the $\mathbf{M}_{ij}$'s have a finite fourth moment, and as $N^{-2/5}$ if all the moments of the $\mathbf{M}_{ij}$ are finite (see [13]).

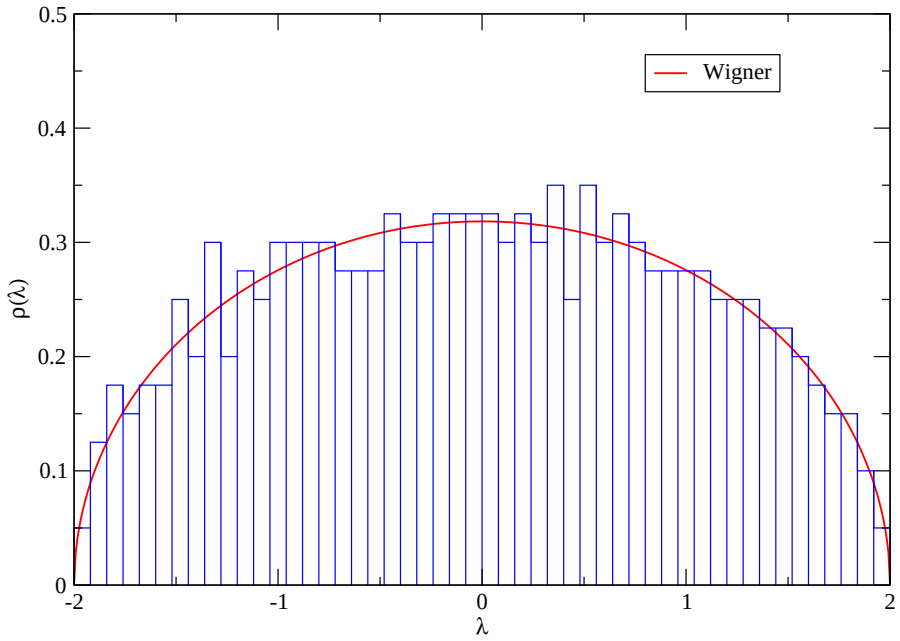



FIGURE 3.1.2. Wigner semi-circle density (3.1.36) compared with empirical results with $N = 500$ (histogram) from one sample, illustrating the convergence of the ESD at finite $N$ to the asymptotic LSD.

Due to the relative simplicity of the expression of Eq. (3.1.35), one can easily invert this expression to find the Blue transform to find that the $\mathcal{R}$-transform of the semicircle law reads

$$\mathcal{R}_W(z) = \sigma^2 z. \tag{3.1.37}$$

Since the average trace $\varphi_1$ is exactly 0, the $\mathcal{S}$-transform of a Wigner matrix is an ill-defined object.

*The Marčenko-Pastur law.* As stated in the introduction, the study of random matrices began with John Wishart [192]. More precisely, let us consider the $N \times T$ matrix $\mathbf{Y}$ consisting of $T$ independent realizations of random centered Gaussian vectors of size $N$ and covariance $\mathbf{C}$, then the Wishart matrix is defined as the $N \times N$ matrix $\mathbf{M}$ as $\mathbf{M} := T^{-1}\mathbf{Y}\mathbf{Y}^*$. In multivariate

statistics, this matrix $\mathbf{M}$ is better known as the sample covariance matrix (see Chapter 4). For *any* $N$ and $T > N$, Wishart derived the exact PDF of the entries $\mathbf{M}$ which reads:

$$\mathcal{P}_{\mathrm{w}}(\mathbf{M}|\mathbf{C}) = \frac{1}{2^{NT/2}\Gamma_N(T/2)} \frac{\det(\mathbf{M})^{\frac{T-N-1}{2}}}{\det(\mathbf{C})^{T/2}} e^{-\frac{T}{2}\mathrm{Tr}\mathbf{C}^{-1}\mathbf{M}}. \tag{3.1.38}$$

As alluded in the introduction, we say that $\mathbf{M}$ (given $\mathbf{C}$) follows a Wishart$(N, T, \mathbf{C}/T)$ distribution. In the "isotropic" case, i.e., when $\mathbf{C} = \mathbf{I}_N$, we can deduce from (3.1.38)

$$\mathcal{P}_{\mathrm{w}}(\mathbf{M}|\mathbf{I}_N) \propto \det(\mathbf{M})^{\frac{T-N-1}{2}} e^{-\frac{T}{2}\mathrm{Tr}\mathbf{M}} := e^{-\frac{T}{2}\mathrm{Tr}\mathbf{M} + \frac{T-N-1}{2}\mathrm{Tr}\log\mathbf{M}}, \tag{3.1.39}$$

which clearly belongs to the class of Boltzmann ensembles (3.1.1). Throughout the following, we shall denote by $\boldsymbol{\mathcal{W}}$ the $N \times N$ matrix whose distribution is given by (3.1.39). Ignoring sub-leading terms, the corresponding potential function is given by:

$$V(z) = \frac{1}{2q}\left[z - (1-q)\log z\right], \qquad \text{with} \qquad q := N/T. \tag{3.1.40}$$

It is easy to see that the derivative indeed gives a Laurent polynomial in $z$ as we have

$$V'(z) = \frac{1}{2qz}\left[z - (1-q)\right].$$

Following our convention, $V'(z)$ is a Laurent polynomial of degree 1 and order $\ell = -1$ so that we deduce $Q(z)$ in (3.1.32) is of the form $c/z$ with $c$ a constant to be determined using (3.1.12). We postpone the computation of the Stieltjes transform $\mathfrak{g}(z)$ to the end of this section. The final result reads:

$$\mathfrak{g}(z) = \frac{(z+q-1) - \sqrt{z-\nu_-}\sqrt{z-\nu_+}}{2qz}, \qquad \nu_\pm := (1 \pm \sqrt{q})^2, \tag{3.1.41}$$

and this is the solution found by Marčenko and Pastur in [123] in the special case $\mathbf{C} = \mathbf{I}_N$. We can now use the inversion formula (3.1.11) to find the celebrated Marčenko-Pastur (MP) law (for $q \in (0,1)$)

$$\rho_{\mathrm{MP}}(\nu) = \frac{\sqrt{4\nu q - (\nu+q-1)^2}}{2q\pi\nu}, \qquad \forall\, \nu \in \left[\nu_-, \nu_+\right]. \tag{3.1.42}$$

Note that for $q \geqslant 1$, it is plain to see that $\mathbf{M}$ has $N - T$ zero eigenvalues that contribute $(1-q)\delta_0$ to the density Eq. (3.1.42). Note that the convergence of the ESD towards the asymptotic MP law occurs, for $q < 1$, at the same speed as in the Wigner case, i.e. as $N^{-2/5}$ in the present case where the random elements of $\mathbf{Y}$ are Gaussian (for a full discussion of this issue, see [11]).

Again, the expression of $\mathfrak{g}(z)$ is simple enough to obtain a closed formula for the Blue transform, and deduce from Eq. (3.1.41) the $\mathcal{R}$-transform of the MP law:

$$\mathcal{R}_{\mathrm{MP}}(\omega) = \frac{1}{1-q\omega}. \tag{3.1.43}$$

One can compute the $\mathcal{S}$-transform of the MP law using the relation (3.1.25):

$$\mathcal{S}_{\mathrm{MP}}(\omega) = \frac{1}{1+q\omega}. \tag{3.1.44}$$

We now derive the Stieltjes transform (3.1.41) through a complete application of the BIPZ formalism introduced in Eq. (3.1.32). As alluded to above, the Stieltjes transform (3.1.32) for the isotropic Wishart matrix has the form

$$\mathfrak{g}(z) = \frac{1}{2q}\left[1 - \frac{1-q}{z}\right] - \frac{c}{z}\sqrt{z-\nu_+}\sqrt{z-\nu_-}, \tag{3.1.45}$$

and the constants that we have to determine are $c, \nu_+$ and $\nu_-$. To that end, we use (3.1.12) that tells us that when $|z| \to \infty$

$$\mathfrak{g}(z) = \frac{1}{z} + \frac{\varphi(\mathbf{M})}{z^2} + \mathcal{O}(z^{-3}). \tag{3.1.46}$$

On the other hand, one finds by taking the limit $z \to \infty$ into (3.1.45) that

$$\mathfrak{g}(z) = \frac{1}{2q}\left[1 - \frac{1-q}{z}\right] - c\left[1 - \frac{\nu_+ + \nu_-}{2z} - \frac{(\nu_+ - \nu_-)^2}{8z^2}\right] + \mathcal{O}(z^{-3}), \tag{3.1.47}$$

Then, by comparing this last equation to (3.1.46), we may fix $c$ by noticing that we have a leading order

$$\frac{1}{2q} - c = 0,$$

since $\mathfrak{g}(z)$ behave as $\mathcal{O}(z^{-1})$ for very large $z$ and therefore we have

$$c = \frac{1}{2q}. \tag{3.1.48}$$

Next, we find at order $\mathcal{O}(z^{-1})$:

$$1 = -\frac{(1-q)}{2q} + \frac{\nu_+ + \nu_-}{4q}, \tag{3.1.49}$$

that is to say

$$\nu_+ = 2(1+q) - \nu_-. \tag{3.1.50}$$

Finally, the last constant is determined with the condition at order $\mathcal{O}(z^{-2})$,

$$\varphi(\mathbf{M}) = \frac{(\nu_+ - \nu_-)^2}{16q}, \tag{3.1.51}$$

which is equivalent to

$$\nu_- = \nu_+ - 4\sqrt{q\varphi(\mathbf{M})} = (1+q) - 2\sqrt{q} = (1-\sqrt{q})^2, \tag{3.1.52}$$

where we used (3.1.50) and $\varphi(\mathbf{M}) = 1$ in the third step. Consequently, we deduce from (3.1.50) that $\nu_+ = (1+\sqrt{q})^2$ and the result (3.1.41) follows from the equations (3.1.48), (3.1.50) and (3.1.52).

*Inverse Wishart matrix.* Another very interesting case is the inverse of a Wishart matrix, simply named the "inverse Wishart" matrix. The derivation of the corresponding eigenvalue density is straightforward from the Marčenko-Pastur law (3.1.42). Indeed, one just needs to make the change of variable $u = ((1-q)\nu)^{-1}$ into Eq. (3.1.42) to obtain:[7]

$$\rho_{\text{IMP}}(u) = \frac{\kappa}{\pi u^2}\sqrt{(u_+ - u)(u - u_-)}, \qquad u_\pm := \frac{1}{\kappa}\left[\kappa + 1 \pm \sqrt{2\kappa+1}\right], \tag{3.1.53}$$

where the subscript IMP stands for "Inverse Marčenko-Pastur" and $\kappa$ is related to $q$ through

$$q = \frac{1}{2\kappa+1} \in (0,1)\,. \tag{3.1.54}$$

In particular, one notices that $u_\pm = (1-q)/\nu_\mp$ where $\nu_\mp$ is defined in Eq. (3.1.41). We plot in Fig. 3.1.3 the density of the Marčenko-Pastur (3.1.42) and of its inverse (3.1.53) both with parameter $q = 0.5$.

[7] The factor $(1-q)^{-1}$ is introduced to keep the mean at one as will be explained below.

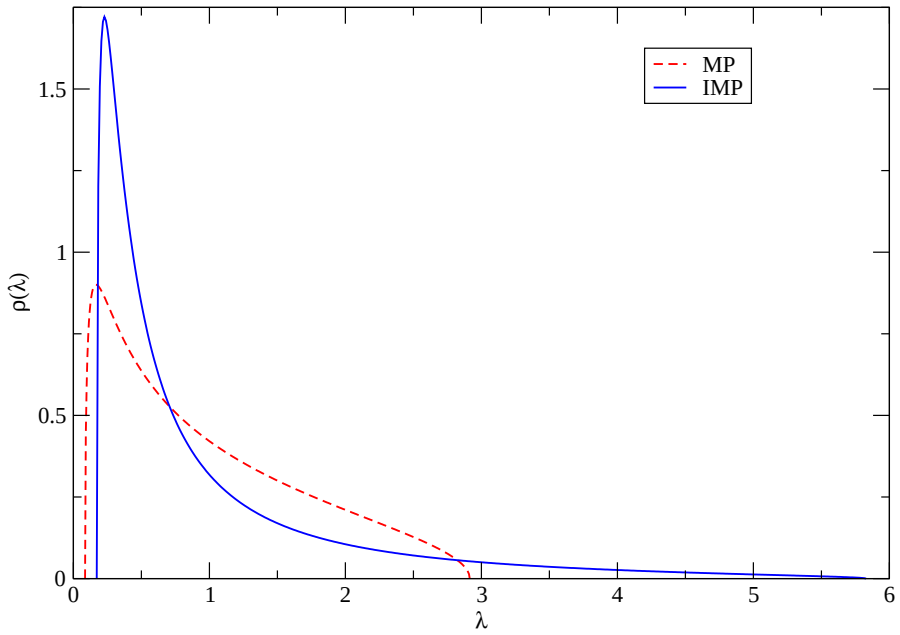



FIGURE 3.1.3. The red dotted curve corresponds to the Marčenko-Pastur density (3.1.42) with $q = 0.5$. We repeat the experiment with the Inverse Wishart matrix still with $q = 0.5$ (plain blue curve).

In addition to the eigenvalue density (3.1.53), one can also derive explicit expressions for the other transforms presented in Section 3.1.1. For the Stieltjes transform, it suffices to apply the same change of variable $u = ((1-q)z)^{-1}$ and to use the properties (3.1.13) and (3.1.14) to obtain:

$$\mathfrak{g}_{\rm iw}(u) = \frac{u(\kappa+1) - \kappa - \kappa\sqrt{u-u_-}\sqrt{u-u_+}}{u^2}\,, \tag{3.1.55}$$

where the bounds $u_\pm$ are given in Eq. (3.1.53). One can easily check with the inversion formula (3.1.9) that we indeed retrieve the density of states (3.1.53) as expected.

Using the Stieltjes transform (3.1.55), one can then compute the $\mathcal{R}$-transform of the Inverse Marčenko-Pastur density to find

$$\mathcal{R}_{\rm IMP}(\omega) = \frac{\kappa - \sqrt{\kappa(\kappa-2\omega)}}{\omega}, \qquad \kappa > 0, \tag{3.1.56}$$

and then, from (3.1.25), the $\mathcal{S}$-transform reads

$$\mathcal{S}_{\rm IMP}(\omega) = 1 - \frac{\omega}{2\kappa}. \tag{3.1.57}$$

In statistics, the derivation of the inverse Wishart distribution is slightly different. Let $\mathbf{M}$ be a $N \times N$ real symmetric matrix that we assume to be invertible and suppose that $\mathbf{M}^{-1}$ follows a Wishart$(N, T, \mathbf{C}^{-1})$ and $\mathbf{C}$ is a $N \times N$ real symmetric positive definite "reference" matrix and $T > N-1$. In that case, it turns out that the PDF of $\mathbf{M}$ is also explicit. More precisely, we say that $\mathbf{M}$ is distributed according to an Inverse-Wishart$(N, T, \mathbf{C})$ whose PDF is given by [9]:

$$\mathcal{P}_{\rm iw}(\mathbf{M}^{-1}|\mathbf{C}) \;=\; \frac{1}{2^{NT/2}\Gamma_N(T/2)} \frac{\det(\mathbf{C})^{T/2}}{\det(\mathbf{M})^{(T+N+1)/2)}} e^{-\frac{1}{2}\,\mathrm{Tr}\,\mathbf{C}\mathbf{M}^{-1}}\,. \tag{3.1.58}$$

In order to get that distribution, one should note that the Jacobian of the transformation $\mathbf{M} \to \mathbf{M}^{-1}$ is equal to $(\det \mathbf{M})^{-N-1}$, as can be derived by using the eigenvalue/eigenvector representation of the measure, see Eq. (3.1.2). A detailed derivation of this change of variable may be found e.g. in [69, Eq. (15.15)].

An important property of the Inverse-Wishart distribution is the following closed formula for the expectation value:

$$\langle \mathbf{M} \rangle_{\mathcal{P}_{\text{iw}}} = \frac{\mathbf{C}}{T - N - 1}. \tag{3.1.59}$$

The derivation of this result can be obtained using the different identities of [86].

We may now explain the factor $(1 - q)$ in the above change of variable. If we consider $\mathbf{C} = \mathbf{I}_N/T$, we deduce from (3.1.59) that

$$\langle \mathbf{M} \rangle_{\mathcal{P}_{\text{iw}}} = \frac{T}{T - N - 1}\mathbf{I}_N \underset{\text{LDL}}{\sim} \frac{1}{1 - q}\mathbf{I}_N. \tag{3.1.60}$$

In order to have a normalized spectral density, i.e. $N^{-1}\text{Tr}\mathbf{M} = 1$, we see that we need to apply $\tilde{\mathbf{M}} = (1 - q)\mathbf{M}$ so that $\langle \tilde{\mathbf{M}} \rangle = \mathbf{I}_N$. This was exactly the purpose of the change of variable $u = ((1 - q)\nu)^{-1}$ in Eq. (3.1.53).

We conclude this section by stating that one can characterize entirely the eigenvalue density function of a broad class of random matrices $\mathbf{M}$ through a potential function. This allows one to reproduce a large variety of empirical spectral densities by adequately choosing the convex confining potential.

**3.1.3. Free probability.** We saw in the previous two examples that one can derive, from the potential function, some analytical results about the ESD which can be very interesting for statistical purposes (*e.g.* the inverse Wishart density). However, the Coulomb gas method does not allow one to investigate the spectrum of a matrix that is perturbed by some noise source. This is a classical problem in Statistics where one is often interested in extracting the "true" signal from noisy observations. Standard models in statistics deal with either an additive or multiplicative noise (as will be the case for empirical correlation matrices). Unless one can write down exactly the PDF of the entries of the corrupted matrix, which is rarely the case, the Coulomb gas analogy is not directly useful.

This section is dedicated to a short introduction to free probability theory, which is an alternative method to study the asymptotic behavior of some large dimensional random matrices. More precisely, free probability provides a robust way to investigate the LSD of either sums or products of random matrices with specific symmetry properties. We will only give here the basic notions of free probability applied to symmetric real random matrices and we refer to e.g. [164] or [46] for a more exhaustive presentation.

*Freeness.* Free probability theory was initiated in 1985 by Dan Voiculescu in order to understand special classes of von Neumann algebras [181], by establishing calculus rules for non commutative operators relying on the notion of **freeness**, defined below for the special case of matrices. A few years later, Voiculescu [183] and Speicher [**?**] found that rotationally invariant random matrices asymptotically satisfy the freeness criteria, and this has had a tremendous impact on RMT.

Roughly speaking, two matrices $\mathbf{A}$ and $\mathbf{B}$ are mutually *free* if their eigenbasis are related to one another by a random rotation, i.e. when their eigenvectors are almost surely orthogonal. For random matrices, we rather use the notion of "asymptotic" freeness. The precise statement is as follows [183]: let $\mathbf{A}$ and $\mathbf{B}$ be two independent self-adjoint matrices of size $N$. If the spectral

density of each matrix converges almost surely in the large $N$ limit and if $\mathbf{B}$ is invariant under rotation, then $\mathbf{A}$ and $\mathbf{B}$ are asymptotically free. This statement can also be found in a different context in [**?**].

The notion of freeness for random matrices is the counterpart of independence for random variables. Indeed, recall that the normalized trace operator, defined as

$$\varphi(\mathbf{M}) := \frac{1}{N}\mathrm{Tr}\mathbf{M}, \tag{3.1.61}$$

is equal to the first moment of $\rho_{\mathbf{M}}$. Then, provided that $\varphi(\mathbf{A}) = \varphi(\mathbf{B}) = 0$, we say that $\mathbf{A}$ and $\mathbf{B}$ are free if the so-called *freeness* property is satisfied, to wit:

$$\varphi\left(\mathbf{A}^{n_1}\mathbf{B}^{m_1}\mathbf{A}^{n_2}\mathbf{B}^{m_2}\dots\mathbf{A}^{n_k}\mathbf{B}^{m_k}\right) = \varphi(\mathbf{A}^{n_1})\varphi(\mathbf{B}^{m_1})\varphi(\mathbf{A}^{n_2})\varphi(\mathbf{B}^{m_2})\dots\varphi(\mathbf{A}^{n_k})\varphi(\mathbf{B}^{m_k}), \tag{3.1.62}$$

for any integers $n_1, \dots, n_k$ and $m_1, \dots, m_k$ with $k \in \mathbb{N}^+$. Note that if $\varphi(\mathbf{A}) \neq 0$ and $\varphi(\mathbf{B}) \neq 0$, then it suffices to consider the centered matrices $\mathbf{A} - \varphi(\mathbf{A})\mathbf{I}_N$ and $\mathbf{B} - \varphi(\mathbf{B})\mathbf{I}_N$.

Let us explore (3.1.62) in the simplest case. For any free matrices $\mathbf{A}$ and $\mathbf{B}$ defined as above, one has

$$\varphi\left((\mathbf{A} - \varphi(\mathbf{A}))(\mathbf{B} - \varphi(\mathbf{B}))\right) = 0, \tag{3.1.63}$$

from which we deduce $\varphi(\mathbf{AB}) = \varphi(\mathbf{A})\varphi(\mathbf{B})$. Hence, if one thinks of the trace operator (3.1.61) as the analogue of the expectation value for non commutative random variables, the freeness property is the analogue of the moment factorization property. More generally, freeness allows the computation of mixed moments of products of matrices from the knowledge of the moments of $\mathbf{A}$ and $\mathbf{B}$, similar to classical independence in probability theory. For example, from

$$\varphi\left((\mathbf{A} - \varphi(\mathbf{A}))(\mathbf{B} - \varphi(\mathbf{B}))(\mathbf{A} - \varphi(\mathbf{A}))\right) = 0, \tag{3.1.64}$$

we can deduce that

$$\varphi(\mathbf{ABA}) = \varphi(\mathbf{A}^2\mathbf{B}) = \varphi(\mathbf{A}^2)\varphi(\mathbf{B}). \tag{3.1.65}$$

One typical example of free pairs of matrices is when $\mathbf{A}$ is a fixed matrix and when $\mathbf{B}$ is a random matrix belonging to a rotationally invariant ensemble, i.e. $\mathbf{B} = \Omega\mathbf{B}_{\mathrm{diag}}\Omega^*$, where $\mathbf{B}_{\mathrm{diag}}$ is diagonal and $\Omega$ distributed according to the Haar (flat) measure over the orthogonal group, in the limit where $N$ is infinitely large. This concept of asymptotic freeness is also related to the notion of vanishing non-planar diagrams [90]. As we shall see in Chapter 8, the computation of mixed moments will be used to derive some useful relations for estimating over-fitting for statistical estimation problems.

*Sums of free matrices.* In addition to the computation of mixed moments such as Eq. (3.1.64), free probability theory allows us to compute the LSD of sums and products of invariant random matrices, as we discuss now.

Let us look at the additive case first. Suppose that we observe a matrix $\mathbf{M}$ which is built from the addition of a fixed "signal" matrix $\mathbf{A}$ and a noisy (or random) matrix $\mathbf{B}$ that we assume to be invariant under rotation, i.e.,

$$\mathbf{M} = \mathbf{A} + \mathbf{\Omega}\mathbf{B}\mathbf{\Omega}^*,$$

for any $N \times N$ matrix $\mathbf{\Omega}$ that belongs to the orthogonal group $\mathbf{O}(N)$. A typical question is to evaluate the LSD of $\mathbf{M}$ and estimate the effect of the noise on the signal in terms of the modification of its eigenvalues. Assuming that the ESD of $\mathbf{A}$ and $\mathbf{B}$ converge to a well defined

limit, the spectral density of $\mathbf{M}$ can be computed using the law of addition for non commutative operators, namely Voiculescu's *free addition*

$$\mathcal{R}_{\mathbf{M}}(\omega) = \mathcal{R}_{\mathbf{A}}(\omega) + \mathcal{R}_{\mathbf{B}}(\omega). \tag{3.1.66}$$

Hence, we can interpret the $\mathcal{R}$-transform (3.1.16) as the analogue in RMT of the logarithm of the Fourier transform for standard additive convolution. It is possible to rewrite Eq. (3.1.66) as a function of the Stieltjes transform of $\mathbf{M}$ that contains all the information about the spectral density of $\mathbf{M}$. Equation (3.1.66) is equivalent to

$$\mathcal{B}_{\mathbf{M}}(\omega) = \mathcal{B}_{\mathbf{A}}(\omega) + \mathcal{R}_{\mathbf{B}}(\omega).$$

Next, we introduce $\omega = \mathfrak{g}_{\mathbf{M}}(z)$ that yields

$$\mathcal{B}_{\mathbf{A}}(\mathfrak{g}_{\mathbf{M}}(z)) = z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z)).$$

It now suffices to apply the function $\mathfrak{g}_{\mathbf{A}}$ on both sides to obtain

$$\mathfrak{g}_{\mathbf{M}}(z) = \mathfrak{g}_{\mathbf{A}}(z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z))). \tag{3.1.67}$$

This last relation establishes the influence of the additive noise coming from the matrix $\mathbf{B}$ on the "signal" (or true) eigenvalues of $\mathbf{A}$.

To gain more insight on this result, let us assume that the noise matrix $\mathbf{B}$ is a simple GOE matrix with centered elements of variance $\sigma^2/\mathrm{N}$. We know from Eq. (3.1.37) that $\mathcal{R}_{\mathbf{B}}(z) = \sigma_{\mathbf{B}}^2 z$. Hence, the spectrum of the sample matrix $\mathbf{M}$ is characterized by the following fixed-point equation:[8]

$$\mathfrak{g}_{\mathbf{M}}(z) = \mathfrak{g}_{\mathbf{A}}(z - \sigma_{\mathbf{B}}^2 \mathfrak{g}_{\mathbf{M}}(z)). \tag{3.1.68}$$

This is the Stieltjes transform of the deformed GOE matrix[9] which is a well-known model in statistical physics of disordered systems. Indeed, this model can be seen as a Hamiltonian that consists of a fixed source subject to an external additive perturbation $\mathbf{B}$ [35]. Taking $\mathbf{A}$ to be a GOE as well, we find that $\mathbf{M}$ is a GOE with variance $\sigma_{\mathbf{A}}^2 + \sigma_{\mathbf{B}}^2$, as expected. In a inference theory context, this model might be useful to describe general linear model where the signal we try to infer is corrupted by an additive noise.

Another interesting application is when the matrix $\mathbf{B}$ has low rank, frequently named a *factor model*. In the example of stocks market, this model can be translated into the fact that there exist few common factors to all stocks such as global news about the economy for instance. For the sake of simplicity, we consider the rank-1 case but the following argument can be easily generalized to a finite rank $r \ll N$. Let us denote the unique nontrivial eigenvalue of $\mathbf{B}$ as $\beta > 0$ and ask ourselves how adding a (randomly oriented) rank-1 matrix affects the spectrum of $\mathbf{M}$. This problem can be solved explicitly using free matrix tools in the LDL. Indeed, as we show below, the largest eigenvalue pops out of the spectrum of $\mathbf{A}$ whenever there exists $z \in \mathbb{R} \backslash \operatorname{supp}[\rho_{\mathbf{A}}]$ such that

$$\mathfrak{g}_{\mathbf{A}}(z) = \frac{1}{\beta}. \tag{3.1.69}$$

[8] This equation can also be interpreted as the solution of a Burgers equation, that appears within the Dyson Brownian motion interpretation of the same problem – see Appendix 11 for more about this.

[9] This result can be generalized to the class of deformed Wigner matrices, i.e. where the noise is given by (3.1.34) but not necessarily Gaussian, see e.g. [107].

For instance, if $\mathbf{A}$ is a Wigner matrix with variance $\sigma^2 > 0$, one can easily check from (3.1.69) and (3.1.37) that the largest eigenvalue $\nu_1$ of $\mathbf{M}$ is given by

$$\nu_1 = \begin{cases} \beta + \sigma^2/\beta & \text{if } \beta > \sigma \\ 2\sigma & \text{otherwise}\,. \end{cases} \tag{3.1.70}$$

When $\beta > \sigma$, we say that $\nu_1$ is an *outlier*, i.e. it lies outside the spectrum of $\rho_{\mathbf{A}}$. Hence, we see that free probability allows one to find a simple criterion for the possible presence of outliers.

Let us now derive the criterion (3.1.69). First we need to compute the $\mathcal{R}$-transform of the rank one matrix $\mathbf{B}$ in order to use (3.1.66). From (3.1.8), we easily find that

$$\mathfrak{g}_{\mathbf{B}}(u) = \frac{1}{N}\frac{1}{u-\beta} + \left(1 - \frac{1}{N}\right)\frac{1}{u} = \frac{1}{u}\left[1 + \frac{1}{N}\frac{\beta}{1-u^{-1}\beta}\right]. \tag{3.1.71}$$

Using perturbation theory, we can invert this last equation to find the Blue transform, and this yields at leading order,

$$\mathcal{B}_{\mathbf{B}}(\omega) = \frac{1}{\omega} + \frac{\beta}{N(1-\omega\beta)} + \mathcal{O}(N^{-2}). \tag{3.1.72}$$

We may therefore conclude from (3.1.16) that

$$\mathcal{R}_{\mathbf{B}}(\omega) = \frac{\beta}{N(1-\beta\omega)} + \mathcal{O}(N^{-2}). \tag{3.1.73}$$

Hence, we obtain by applying (3.1.66) and (3.1.16) that

$$\mathcal{B}_{\mathbf{M}}(\omega) = \mathcal{B}_{\mathbf{A}}(\omega) + \frac{\beta}{N(1-\beta\omega)} + \mathcal{O}(N^{-2}). \tag{3.1.74}$$

Next, we set $\omega = \mathfrak{g}_{\mathbf{M}}(z)$ so that this latter equation becomes

$$z = \mathcal{B}_{\mathbf{A}}(\mathfrak{g}_{\mathbf{M}}(z)) + \frac{\beta}{N(1-\beta\mathfrak{g}_{\mathbf{M}}(z))} + \mathcal{O}(N^{-2}). \tag{3.1.75}$$

From this equation, we expect the Stieltjes transform of $\rho_{\mathbf{M}}$ to be of the form

$$\mathfrak{g}_{\mathbf{M}}(z) = \mathfrak{g}_0(z) + \frac{\mathfrak{g}_1(z)}{N} + \mathcal{O}(N^{-2}). \tag{3.1.76}$$

By plugging this ansatz into (3.1.75), we see that $\mathfrak{g}_0(z)$ and $\mathfrak{g}_1(z)$ satisfies

$$\begin{aligned} z &= \mathcal{B}_{\mathbf{A}}(\mathfrak{g}_0(z)) \\ \mathfrak{g}_1(z) &= -\frac{\beta}{\mathcal{B}'_{\mathbf{A}}(\mathfrak{g}_0(z))(1-\mathfrak{g}_0(z)\beta)}\,. \end{aligned} \tag{3.1.77}$$

It is easy to find that $\mathfrak{g}_0(z) = \mathfrak{g}_{\mathbf{A}}(z)$ as expected. We now focus on the $1/N$ correction term and using that $\mathcal{B}'_{\mathbf{A}}(\mathfrak{g}_{\mathbf{A}}(z)) = 1/\mathfrak{g}_{\mathbf{A}}(z)$, we conclude that

$$\mathfrak{g}_1(z) = -\frac{\beta\mathfrak{g}'_{\mathbf{A}}(z)}{1-\mathfrak{g}_{\mathbf{A}}(z)\beta}\,. \tag{3.1.78}$$

Finally, we obtained that

$$\mathfrak{g}_{\mathbf{M}}(z) \approx \mathfrak{g}_{\mathbf{A}}(z) - \frac{1}{N}\frac{\beta\mathfrak{g}'_{\mathbf{A}}(z)}{1-\mathfrak{g}_{\mathbf{A}}(z)\beta}, \tag{3.1.79}$$

and we see that the correction term only survive in the large $N$ limit if $\mathfrak{g}_{\mathbf{A}}(z) = \beta^{-1}$ has a non trivial solution. Differently said, $z$ is an eigenvalue of $\mathbf{M}$ and not of $\mathbf{A}$ if there exists $z \in \mathbb{R}\backslash\,\mathrm{supp}[\rho_{\mathbf{A}}]$ such that $\mathfrak{g}_{\mathbf{A}}(z) = \beta^{-1}$ and this leads to the criterion (3.1.69).

*Products of free matrices.* Similar results are available for free multiplicative convolution. Before showing how to obtain the LSD of the product of free matrices, we first emphasize that one has to carefully define the product of free matrices. Indeed, the naive analogue of the free addition would be to define $\mathbf{M} = \mathbf{AB}$. However the product $\mathbf{AB}$ is in general not self-adjoint when $\mathbf{A}$ and $\mathbf{B}$ are self-adjoint but not commuting. In the case where $\mathbf{A}$ is positive definite, we can see that the product $\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}$ makes sense and share the same moments than the product $\mathbf{AB}$. Therefore, we define the product of free matrices by

$$\mathbf{M} := \sqrt{\mathbf{A}}\mathbf{B}\sqrt{\mathbf{A}}. \tag{3.1.80}$$

Note that in this case, $\mathbf{B}$ need not be necessarily positive definite but must have a trace different from zero (see the Taylor expansion below). For technical reason, we need the LSD of $\mathbf{B}$ to be well-defined. Under this assumption, the free multiplicative convolution rule for random matrices is given by

$$\mathcal{S}_{\mathbf{M}}(\omega) \;=\; \mathcal{S}_{\mathbf{A}}(\omega)\mathcal{S}_{\mathbf{B}}(\omega). \tag{3.1.81}$$

This is the so-called *free multiplication*, which has been first obtained by Voiculescu [183] and then by [198] in a physics formalism.

Again, if one is interested in the limiting spectral density of $\mathbf{M}$, one would like to write (3.1.81) in terms of its Stieltjes transform. Using the very definition of the $\mathcal{S}$-transform, we rewrite (3.1.81) as

$$\frac{1}{\mathcal{T}_{\mathbf{M}}^{-1}(\omega)} = \frac{\mathcal{S}_{\mathbf{B}}(\omega)}{\mathcal{T}_{\mathbf{A}}^{-1}(\omega)}.$$

The trick is the same as above so we therefore set $\omega = \mathcal{T}_{\mathbf{M}}(z)$ to find

$$\mathcal{T}_{\mathbf{A}}^{-1}(\mathcal{T}_{\mathbf{M}}(z)) = z\mathcal{S}_{\mathbf{B}}(\mathcal{T}_{\mathbf{M}}(z)). \tag{3.1.82}$$

It is now immediate to get the analogue of (3.1.67) for the multiplicative case

$$\mathcal{T}_{\mathbf{M}}(z) = \mathcal{T}_{\mathbf{A}}\left(z\mathcal{S}_{\mathbf{B}}(\mathcal{T}_{\mathbf{M}}(z))\right), \tag{3.1.83}$$

that gives in terms of the Stieltjes transform

$$z\mathfrak{g}_{\mathbf{M}}(z) = Z(z)\mathfrak{g}_{\mathbf{A}}\left(Z(z)\right), \qquad Z(z) \;:=\; z\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1). \tag{3.1.84}$$

This is certainly one the most important results of RMT for statistical inference. It allows one to generalize the Marčenko-Pastur law for sample covariance matrices to arbitrary population covariance matrices $\mathbf{C}$ (see next section), and obtain results on the eigenvectors as well. We emphasize that the literature on free products can be adapted to non Hermitian matrices, see [46] or [45] for a recent review on the multiplication of random matrices.

### 3.1.4. Replica analysis.

*Resolvent and the Replica trick.* As we noticed above (Eq. 3.1.6), information about the eigenvectors can be studied through the resolvent. However, both the Coulomb gas analogy and free probability tools are blind to the structure of eigenvectors since these only give information about the normalized trace of the resolvent. In order to study the resolvent matrix, we need to introduce other tools, for example one borrowed from statistical physics named the *Replica* method. To make it short, the Replica method allows one to rewrite the expectation value of a logarithm in terms of moments, expressed as expectation values of many copies, named the *replicas*, of the

initial system. This method has been extremely successful in various contexts, including RMT and disordered systems, see e.g. [?, 134], or [137] for a more recent review. We stress that even if this method turns out to be a very powerful heuristic, it is not rigorous mathematically speaking (see below). Therefore, it is essential to verify the result obtain from the Replica method using other methods, for example numerical simulations. Note that a rigorous but more difficult way to deal with resolvent is the recursion technique that uses linear algebra results, as explained in Appendix B.4. Other available techniques include Feynman diagrams [44, 48].

As a warm-up exercise, we present briefly the approach for the Stieltjes transform and then explain how to extend it to the study of full resolvent. We notice that any Stieltjes transform can be expressed as

$$\mathfrak{g}(z) = \sum_{i=1}^{N} \frac{1}{z-\nu_i} = \frac{\partial}{\partial z} \log \prod_{i=1}^{N} (z-\nu_i) = \frac{\partial}{\partial z} \log \det(zI - \mathbf{M}). \tag{3.1.85}$$

Then, using the Gaussian representation of $\det(zI-\mathbf{M})^{-1/2}$, we have that

$$\mathcal{Z}(z) \equiv (\det(zI-\mathbf{M}))^{-1/2} = \int \exp\left[-\frac{1}{2}\sum_{i,j=1}^{N} \eta_i (zI-\mathbf{M})_{ij}\eta_j\right] \prod_{j=1}^{N}\left(\frac{\mathrm{d}\eta_j}{\sqrt{2\pi}}\right). \tag{3.1.86}$$

Plugging this last equation into (3.1.85) and assuming that the Stieltjes transform is self-averaging, we see that we need to compute the average of the logarithm of $\mathcal{Z}(z)$:

$$\mathfrak{g}(z) = -2\frac{\partial}{\partial z}\mathbb{E}\log \mathcal{Z}(z), \tag{3.1.87}$$

where the average is taken over the probability distribution $\mathcal{P}_{\mathbf{M}}$. However, it would be easier to compute the moments $\mathbb{E}\mathcal{Z}^n(z)$ instead of $\mathbb{E}\log\mathcal{Z}(z)$ and this is precisely the purpose of the *Replica trick* which was initially formulated as the following identity

$$\log \mathcal{Z} = \lim_{n\to 0} \frac{\mathcal{Z}^n - 1}{n}, \tag{3.1.88}$$

so that one formally has

$$\mathfrak{g}(z) = \lim_{n\to 0} \frac{\partial}{\partial z}\frac{\mathbb{E}\mathcal{Z}^n - 1}{n}. \tag{3.1.89}$$

We have thus transformed the problem (3.1.87) into the computation of $n$ replicas of the system involved in $\mathcal{Z}^n(z)$. The non-rigorous part of this method is quite obvious at this stage. While the integer moments of $\mathcal{Z}$ can indeed be expressed as an average of the replicated system, the identity (3.1.88) requires vanishingly small, *real* values of $n$. Typically, one works with integer $n$'s and then perform an analytical continuation of the result to real values of $n$ before taking the limit $n \to 0$ (after, as it turns out, sending the size of the matrix $N$ to infinity!). Therefore, the main concern of this method is that we assume that the analytical continuation poses no problem, which is not necessarily the case. It is precisely this last step that could lead to uncontrolled approximations in some cases [143], which is why numerical (or other) checks are mandatory. Nonetheless, the Replica trick gives a simple heuristic to compute the Stieltjes transform $\mathfrak{g}(z)$ which, as shown below, is exact for the quantities considered in this thesis.

For our purposes, we need to extend the above Replica formalism for the entire resolvent and not only its normalized trace. In that case, we will need a slightly different *Replica identity*,

extending (3.1.88), that we shall now present. The starting point is to rewrite the entries of the resolvent matrix $\mathbf{G}(z)$ using the Gaussian integral representation of an inverse matrix

$$(z\mathbf{I}_N - \mathbf{M})^{-1}_{ij} = \frac{\int \left(\prod_{k=1}^N d\eta_k\right) \eta_i \eta_j \exp\left\{-\frac{1}{2}\sum_{k,l=1}^N \eta_k (z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\}}{\int \left(\prod_{k=1}^N d\eta_k\right) \exp\left\{-\frac{1}{2}\sum_{k,l=1}^N \eta_k (z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\}}. \tag{3.1.90}$$

As explained in Appendix B.4, we expect that (3.1.90) is self-averaging in the LDL thanks to the Central Limit Theorem, so that:

$$(z\mathbf{I}_N - \mathbf{M})^{-1}_{ij} = \left\langle \frac{1}{\mathcal{Z}} \int \left(\prod_{k=1}^N d\eta_k\right) \eta_i \eta_j \exp\left\{-\frac{1}{2}\sum_{k,l=1}^N \eta_k (z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\} \right\rangle_{\mathcal{P}_\mathbf{M}}, \tag{3.1.91}$$

where $\mathcal{Z}$ is as above the partition function, i.e. the denominator in Eq. (3.1.90). The replica identity for resolvent is given by

$$\begin{aligned} G_{ij}(z) &= \lim_{n\to 0} \left\langle \mathcal{Z}^{n-1} \int \left(\prod_{k=1}^N d\eta_k\right) \eta_i \eta_j \exp\left\{-\frac{1}{2}\sum_{k,l=1}^N \eta_k (z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\} \right\rangle_{\mathcal{P}_\mathbf{M}} \\ &= \lim_{n\to 0} \int \left(\prod_{k=1}^N \prod_{\alpha=1}^n d\eta_k^\alpha\right) \eta_i^1 \eta_j^1 \left\langle \prod_{\alpha=1}^n \exp\left\{-\frac{1}{2}\sum_{k,l=1}^N \eta_k^\alpha (z\delta_{kl} - \mathbf{M}_{kl})\eta_l^\alpha\right\} \right\rangle_{\mathcal{P}_\mathbf{M}} \end{aligned} \tag{3.1.92}$$

Again, we managed to rewrite the initial problem (3.1.91) as the computation of $n$ replicas. We emphasize that (3.1.92) is valid for any random matrix $\mathbf{M}$, and is useful provided that we are able to compute the average over the probability density $\mathcal{P}_\mathbf{M}$. The identity (3.1.92) is the central tool of this section. In particular, it allows one to study the asymptotic behavior of the resolvent entry-wise, which contains more information about the spectral decomposition of $\mathbf{M}$ than just the normalized trace [40]. As will become apparent below, we consider a model of random matrices inspired by Free Probability theory, i.e. $\mathbf{M} = \mathbf{A} + \mathbf{\Omega B \Omega}^*$ and $\mathbf{M} = \mathbf{A}^{1/2}\mathbf{\Omega B \Omega}^*\mathbf{A}^{1/2}$ (see Section 3.1.3 above for a more details). We shall focus on the model of free multiplication since the arguments below may be repeated almost verbatim for the free additive case (see Appendix 11).

*Matrix multiplication using replicas.* We reconsider the model (3.1.80) and assume without loss of generality that $\mathbf{A}$ is diagonal. In that case, we see that $\mathcal{P}_\mathbf{M}$ is simply the Haar measure over the orthogonal group $\mathbf{O}(N)$. We specialize the replica identity (3.1.92) to $\mathbf{M} = \mathbf{A}^{1/2}\mathbf{\Omega B \Omega}^*\mathbf{A}^{1/2}$ so that we get

$$G_{ij}(z) = \lim_{n\to 0} \int \left(\prod_{k=1}^N \prod_{\alpha=1}^n d\eta_k^\alpha\right) \eta_i^1 \eta_j^1 e^{-\frac{z}{2}\sum_{\alpha=1}^n \sum_{k=1}^N (\eta_k^\alpha)^2} \mathcal{I}_1\left(\sum_{\alpha=1}^n \left(\eta^\alpha \mathbf{A}^{1/2}\right)\left(\eta^\alpha \mathbf{A}^{1/2}\right)^*, \mathbf{B}\right), \tag{3.1.93}$$

where

$$\mathcal{I}_\beta(\mathbf{A}', \mathbf{B}) := \int \exp\left[-\frac{\beta N}{2} \operatorname{Tr} \mathbf{A}' \Omega \mathbf{B} \Omega^*\right] \mathcal{D}\Omega, \tag{3.1.94}$$

is the so-called *Harish-Chandra–Itzykson-Zuber* integral [89, 95]. Explicit results for this integral are known for Hermitian matrices ($\beta = 2$) for any integer dimension $N$, but not for real orthogonal matrices. Even the study of (3.1.94) in the limit $N \to \infty$ is highly non trivial (see

Appendix A). Nevertheless, in the case where $\mathbf{A}'$ is of finite rank, the leading contribution for $N \to \infty$ is known for any symmetry group. Fortunately, we see that $\mathbf{A}'$ in our case is of rank $n$ and the result is obtained from Eq. (A.1.7) in Appendix A:[10]

$$\mathcal{I}_1\Big(\sum_{\alpha=1}^{n}(\eta^\alpha \mathbf{A}^{1/2})(\eta^\alpha \mathbf{A}^{1/2})^*, \mathbf{B}\Big) \underset{N\to\infty}{\sim} \exp\left[\frac{N}{2}\sum_{\alpha=1}^{n}\mathcal{W}_{\mathbf{B}}\left(\frac{1}{N}\sum_{i=1}^{N}(\eta_i^\alpha)^2 a_i\right)\right], \tag{3.1.95}$$

with

$$\mathcal{W}'_{\mathbf{B}}(.) = \mathcal{R}_{\mathbf{B}}(.)\,, \tag{3.1.96}$$

and where we assume that the vectors $[\eta^\alpha]_{\alpha=1}^n$ are orthogonal to each other, which is generically true provided $n \ll N$. We then plug this result into (3.1.93) and introduce an auxiliary variable $p^\alpha = \frac{1}{N}\sum_{i=1}^{N}(\eta_i^\alpha)^2 a_i$ that we enforce using the exponential representation of a Dirac delta function

$$\delta\Big(p^\alpha - \frac{1}{N}\sum_{i=1}^{N}(\eta_i^\alpha)^2 a_i\Big) = \int \frac{1}{2\pi}\exp\left[\mathrm{i}\zeta^\alpha\Big(p^\alpha - \frac{1}{N}\sum_{i=1}^{N}(\eta_i^\alpha)^2 a_i\Big)\right]\mathrm{d}\zeta^\alpha, \tag{3.1.97}$$

for each $\alpha = 1, \ldots, n$. This allows to retrieve a Gaussian integral on $\eta^\alpha$. Renaming $\zeta^\alpha = -2\mathrm{i}\zeta^\alpha/N$ yields the result

$$G_{ij}(z) \propto \int\int\left(\prod_{\alpha=1}^{n}\mathrm{d}p^\alpha \mathrm{d}\zeta^\alpha\right)\frac{\delta_{ij}}{z-\zeta^1 a_i}\exp\left[-\frac{Nn}{2}F_0(p^\alpha,\zeta^\alpha)\right] \tag{3.1.98}$$

where $F_0$ is the free energy given by

$$F_0(p^\alpha,\zeta^\alpha) = \frac{1}{n}\sum_{\alpha=1}^{n}\left[\frac{1}{N}\sum_{k=1}^{N}\log(z-\zeta^\alpha a_k) + \zeta^\alpha p^\alpha - \mathcal{W}_{\mathbf{B}}(p^\alpha)\right]. \tag{3.1.99}$$

Now, one sees that the integral over $\mathrm{d}p^\alpha \mathrm{d}\zeta^\alpha$ involves the exponential of $Nn/2$ times the free energy, which is of order unity. Provided that $n$ is non-zero, one can estimate this integral via a saddle point method (but of course $n$ will be sent to zero eventually...). We assume a *replica symmetric* ansatz for the saddle point, i.e. $p^\alpha = p^*$ and $\zeta^\alpha = \zeta^*$, $\forall \alpha = 1, \ldots, n$. This is natural since $F_0$ is invariant under the permutation group $P_n$. Note however that the replica symmetric ansatz can lead to erroneous results and this phenomenon is known as *replica symmetry breaking*, see e.g. [134, 143] or [167] and references therein for a mathematical formalism. The rest of the calculation relies on a saddle-point analysis whose details we postpone below, and we finally obtain a so-called "global law" for the resolvent of $\mathbf{M}$:[11]

$$z\mathbf{G}_{\mathbf{M}}(z) \underset{N\to\infty}{\sim} Z(z)\mathbf{G}_{\mathbf{A}}(Z(z)), \qquad Z(z) := z\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z)-1), \tag{3.1.100}$$

which is often referred to as a *subordination* relation between the resolvent of $\mathbf{M}$ and $\mathbf{A}$. Note that the average resolvent $\mathbf{G}_{\mathbf{M}}(z)$ is diagonal in the eigenbasis of $\mathbf{A}$, as expected by symmetry. Taking the trace of both sides of the above equation, one notices that (3.1.100) is a generalization of the formula (3.1.84) as a matrix. We should emphasize that Eq. (3.1.100) is self-averaging

[10] Recall that we work with $n$ as an integer throughout the intermediate steps of the computation.

[11] The term "global" assumes that the imaginary part of $z$ is much larger than $N^{-1}$, in contrast to many different studies of the resolvent at a "local" scale (see [22] for a detail presentation of this concept for Wigner matrices).

element by element for the matrix $\mathbf{G}_{\mathbf{M}}(z)$, i.e. $G_{ij}(z) = \langle G_{ij}(z)\rangle + \mathcal{O}(N^{-1/2})$. The matrix $\mathbf{G}_{\mathbf{M}}(z)$ taken as a whole cannot be considered deterministic, for example $\langle \mathbf{G}_{\mathbf{M}}(z)\rangle^2$ is in general different from $\langle \mathbf{G}^2_{\mathbf{M}}(z)\rangle$. Nevertheless in what follows we will write deterministic equations for $\mathbf{G}_{\mathbf{M}}(z)$ which should be interpreted as element by element self-averaging equations.

We can redo the exact same calculations for the free addition model $\mathbf{M} = \mathbf{A} + \Omega\mathbf{B}\Omega^*$, still with $\mathbf{A} = \mathrm{diag}(a_1, a_2, \ldots, a_N)$ (see Appendix 11). Starting from the replica identity (3.1.92) and then applying (A.1.7), we obtain the following expression [40]:

$$G_{ij}(z) \propto \int\int \left(\prod_{\alpha=1}^{n} dp^\alpha d\zeta^\alpha\right) \frac{\delta_{ij}}{z - \zeta^1 - a_i} \exp\left\{-\frac{Nn}{2} F_0^a(p^\alpha, \zeta^\alpha)\right\}, \tag{3.1.101}$$

where the 'free energy' $F_0^a$ is given by

$$F_0^a(p, \zeta) := \frac{1}{Nn} \sum_{\alpha=1}^{n} \left[\sum_{k=1}^{N} \log(z - \zeta^\alpha - a_k) - \mathcal{W}_{\mathbf{B}}(p^\alpha) + p^\alpha \zeta^\alpha\right]. \tag{3.1.102}$$

Invoking once again the replica symmetric ansatz, the subordination for the resolvent under the free addition model follows from a saddle-point analysis [40]

$$\mathbf{G}_{\mathbf{M}}(z) \underset{N\to\infty}{\sim} \mathbf{G}_{\mathbf{A}}(Z_a(z)), \qquad Z_a(z) := z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z)), \tag{3.1.103}$$

which is exactly the result obtained in [103] in a mathematical formalism. Again taking the trace of both sides of this equation allows one to recover the relation (3.1.67) between Stieltjes transforms.

*Free multiplication: replica saddle-point analysis.*

We now present the derivation of (3.1.100) from (3.1.98). We shall that it actually provides an elementary derivation of the free multiplication formula (3.1.81). Under the replica symmetric ansatz, the free energy becomes

$$F_0(p^\alpha, \zeta^\alpha) \equiv F_0(p, \zeta) = \frac{1}{N} \sum_{k=1}^{N} \log(z - \zeta a_k) + \zeta p - \mathcal{W}_{\mathbf{B}}(p),$$

which needs to be extremized. We first consider the first order condition with respect to $p$ which leads to

$$\zeta^* = \mathcal{R}_{\mathbf{B}}(p^*). \tag{3.1.104}$$

The other derivative with respect to $\zeta$ gives:

$$p^* = \frac{1}{\zeta^* N} \sum_{k=1}^{N} \frac{a_k}{z/\zeta^* - a_k} = \frac{\mathcal{T}_{\mathbf{A}}\left(\frac{z}{\mathcal{R}_{\mathbf{B}}(p^*)}\right)}{\mathcal{R}_{\mathbf{B}}(p^*)}. \tag{3.1.105}$$

Hence, plugging (3.1.104) and (3.1.105) into (3.1.98), we get in the large $N$ limit and then the limit $n \to 0$ by

$$G_{ij}(z)_{ij} = \frac{\delta_{ij}}{z - \mathcal{R}_{\mathbf{B}}(p^*) c_i}. \tag{3.1.106}$$

We can find a genuine simplification of the last expression using the connection with the free multiplication convolution. By taking the normalized trace of $\mathbf{G}_{\mathbf{M}}(z)$, we see that we have

$$z\mathfrak{g}_{\mathbf{M}}(z) = Z\mathfrak{g}_{\mathbf{A}}(Z), \quad \text{with} \quad Z \equiv Z(z) = \frac{z}{\mathcal{R}_{\mathbf{B}}(p^*)}, \tag{3.1.107}$$

which can rewrite as

$$\mathcal{T}_{\mathbf{M}}(z) = \mathcal{T}_{\mathbf{A}}(Z).$$

Let us define

$$\omega = \mathcal{T}_{\mathbf{M}}(z) = \mathcal{T}_{\mathbf{A}}(Z). \tag{3.1.108}$$

Using Eq. (3.1.105), this latter equation implies $p^* = \omega/\mathcal{R}_{\mathbf{B}}(p^*)$. Let us now show how to retrieve the free multiplicative convolution (3.1.81) from (3.1.107) in the large $N$ limit. Indeed, let us rewrite (3.1.108) as

$$z\mathcal{T}_{\mathbf{M}}(z) = Z\mathcal{T}_{\mathbf{A}}(Z)\mathcal{R}_{\mathbf{B}}(p^*), \tag{3.1.109}$$

and it is trivial to see that using (3.1.108) that this last expression can be rewritten as $\omega\mathcal{T}_{\mathbf{M}}^{-1}(\omega) = \omega\mathcal{T}_{\mathbf{A}}^{-1}(\omega)\mathcal{R}_{\mathbf{B}}(p^*)$. Finally, using the definition of the $\mathcal{S}$-transform (3.1.23), this yields

$$\mathcal{S}_{\mathbf{M}}(\omega) = \mathcal{S}_{\mathbf{A}}(\omega)\frac{1}{\mathcal{R}_{\mathbf{B}}(p^*)}. \tag{3.1.110}$$

Using (3.1.25), we also have

$$\frac{1}{\mathcal{R}_{\mathbf{B}}(p^*)} = \mathcal{S}_{\mathbf{B}}(p^*\mathcal{R}_{\mathbf{B}}(p^*)), \tag{3.1.111}$$

But recalling that $p^* = \omega/\mathcal{R}_{\mathbf{B}}(p^*)$, we conclude from (3.1.104), (3.1.108) and (3.1.111) that

$$\frac{1}{\zeta^*} = \mathcal{R}_{\mathbf{B}}(p^*) = \mathcal{S}_{\mathbf{B}}(\mathcal{T}_{\mathbf{M}}(z)). \tag{3.1.112}$$

Going back to (3.1.110), we see that the spectral density of $\mathbf{M}$ is given by Voiculescu's free multiplication formula

$$\mathcal{S}_{\mathbf{M}}(\omega) = \mathcal{S}_{\mathbf{A}}(\omega)\mathcal{S}_{\mathbf{B}}(\omega), \tag{3.1.113}$$

confirming that the replica symmetry ansatz is indeed valid in this case. Finally, by plugging (3.1.112) into (3.1.106), we get the result (3.1.100).

# Chapter 4

# Spectrum of large empirical covariance matrices

## 4.1 Sample covariance matrices

**4.1.1. Setting the stage.** After a general introduction to RMT and to some of the many different analytical tools, we are now ready to handle the main issue of this thesis, which is the statistics of sample covariance matrices. As a preliminary remark, note that we assume that the variance of each variable can be estimated independently with great accuracy given that we have $T \gg 1$ observations for each of them. Consequently, all variables will be considered to have unit variance in the following and we will not distinguish further covariances and correlations henceforth.

As stated in the introduction, the study of correlation matrices has a long history in statistics. Suppose we consider a (random) vector $\mathbf{y} = (y_1, y_2, \ldots, y_N)$. One standard way to characterize the underlying interaction network between these variables is through their correlations. Hence, the goal is to measure as precisely as possible the *true* (or *population*) covariance matrix, defined as

$$\mathbf{C}_{ij} = \mathbb{E}\big[y_i y_j\big], \quad i, j \in [\![1, N]\!] \tag{4.1.1}$$

where we assumed that the $\{y_i\}_{i\in[\![1,N]\!]}$ have zero mean without loss of generality (see below). It is obvious from the definition of $\mathbf{C}$ that the covariance matrix is symmetric. Throughout the following, we shall define the spectral decomposition of $\mathbf{C}$ as

$$\mathbf{C} \;=\; \sum_{i=1}^{N} \mu_i \mathbf{v}_i \mathbf{v}_i^*, \tag{4.1.2}$$

with $\mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_N$ the real eigenvalues and $\mathbf{v}_1, \ldots, \mathbf{v}_N$ the corresponding eigenvectors.

As illustrated in the introduction, the concept of covariances is of crucial importance in a wide range of applications. For instance, let us consider an example that stems from financial applications. The probability of large losses of a diversified portfolio is dominated by the correlated moves of its different constituents (see section 8.1 for more details). In fact, the very notion of diversification depends on the correlations between the assets in the portfolio. Hence, the estimation of the correlations between the price movements of these assets is at the core of risk management policies.

The major concern in practice is that the *true* covariance matrix $\mathbf{C}$ is in fact unknown. To bypass this problem, one often relies on a large number $T$ *independent* measurements, namely

the "samples" $\mathbf{y}_1, \ldots, \mathbf{y}_T$, to construct empirical estimates of $\mathbf{C}$. We thus define the $N \times T$ matrix $\mathbf{Y}_{it} \in \mathbb{R}^{N \times T}$, whose elements are the $t$-th measurement of the variable $y_i$. Within our example from finance, the random variable $Y_{it}$ would be the return of the asset $i$ at time $t$. Eq. (4.1.1) is then approximated by an average value over the whole sample data of size $T$, leading to the *sample* (or *empirical*) covariance matrix estimator:

$$\mathbf{E}_{ij} = \frac{1}{T}(\mathbf{Y}\mathbf{Y}^*)_{ij} = \frac{1}{T}\sum_{\tau=1}^{T} \mathbf{Y}_{it}\mathbf{Y}_{jt}. \tag{4.1.3}$$

In the statistical literature, this estimator is known as *Pearson* estimator and in the RMT community, the resulting matrix sometimes referred to as defining the Wishart Ensemble. Whereas the Wigner Ensemble has been the subject of a huge amount of studies in physics [1], results on the Wishart Ensemble mostly come from mathematics & statistics [14,123,145], telecommunication [57] or the financial/econophysics literature [29,44,151], although some work in the physics literature also exists [121,149,180,191] to cite a few.

In what we call the "classical" statistical limit, i.e. $T \to \infty$ with $N$ fixed, the law of large numbers tells us that $\mathbf{E}$ converges to the true covariance $\mathbf{C}$. However, as recalled in the introduction, in the present "Big Data" era where scientists are confronted with large datasets such that the sample size $T$ *and* the number of variables $N$ are both very large, but with an observation ratio $q = N/T$ of order unity, specific issues arise. This setting is known in the literature as the high-dimensional limit or Kolmogorov regime (or more commonly named the Big Data regime). This regime clearly differs from the traditional large $T$, fixed $N$ situation (i.e. $q \to 0$), where classical results of multivariate statistics apply. The setting $q \sim O(1)$ is exactly where tools from RMT can be helpful to make precise statements on the empirical covariance matrix (4.1.3).

A typical question would be to study the ESD of $\mathbf{E}$ in order to quantify its 'deviation' from the eigenvalue distribution of the true covariance matrix $\mathbf{C}$. More precisely, does the ESD converges to an explicit LSD? If it does, can we get a tractable expression for this LSD? In the case where the samples $\{\mathbf{y}_t\}_{t=1}^T$ are given by a multivariate Gaussian distribution with zero mean and covariance $\mathbf{C}$, the distribution of the matrix $\mathbf{E}$ is exactly known since Wishart [192], and is given by Eq. (3.1.38) above, with $\mathbf{M} \to \mathbf{E}$. In the case where $\mathbf{C} = T^{-1}\mathbf{I}_N$, we retrieve the isotropic Wishart matrix above that we fully characterized in the previous chapter. The aim is now to provide the LSD of $\mathbf{E}$ for an *arbitrary* true covariance matrix $\mathbf{C}$. More specifically, we shall look at linear models where the data matrix $\mathbf{Y}$ can be decomposed as

$$\mathbf{Y} = \sqrt{\mathbf{C}}\mathbf{X}, \tag{4.1.4}$$

where $\mathbf{X}$ is a $N \times T$ random matrix with uncorrelated entries satisfying

$$\mathbb{E}[X_{it}] = 0, \qquad \mathbb{E}[X_{it}^2] = \frac{1}{T}. \tag{4.1.5}$$

The above decomposition is always possible for multivariate Gaussian variables. Otherwise, the above framework assumes that our correlated random variables $y_i$ are obtained as linear combinations of uncorrelated random variables. In addition, we also require that the random variables $\sqrt{T}X_{i,t}$ have a bounded 4-th moment, in other words that the distribution cannot be extremely fat-tailed.

Next, we introduce the spectral decomposition of $\mathbf{E}$ as

$$\mathbf{E} = \sum_{i=1}^{N} \lambda_i \mathbf{u}_i \mathbf{u}_i^*, \tag{4.1.6}$$

with $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_N$ the eigenvalues and $\mathbf{u}_1, \ldots, \mathbf{u}_N$ the corresponding eigenvectors. Let us now give the main assumptions on the spectrum of $\mathbf{E}$ that we shall suppose to hold for the rest of this thesis:

(i) The support of $\rho_{\mathbf{E}}$ consists of $r+1$ (connected) components with $r \geqslant 0$. We call the $r$ largest components the *outliers* and the smallest component the bulk. The boundary points of the bulk components are labelled $\lambda_-$ and $\lambda_+$ (with $\lambda_- \leqslant \lambda_+$).

(ii) We suppose that the outliers are separated from each other and from the bulk (non-degeneracy).

(iii) We suppose that the bulk is *regular* in the sense that the density of $\rho_{\mathbf{E}}$ vanishes as a square root at the boundary points $\lambda_-, \lambda_+$.

In this chapter, we will look at the eigenvalues statistics of this model and the following section will be devoted to the eigenvectors.

We end this short introduction of the sample covariance matrix with two different remarks. The first one comments the zero-mean assumption made above, while the second one is concerned with the possible fat-tailed nature of the random variables under scrutiny.

**4.1.2. Zero-mean assumption.** In real datasets, that sample vectors $\mathbf{y}_t$ usually have a non-zero mean (even if the true underlying distribution is of zero mean). One can choose therefore choose to shift the sample vectors in such a way that the empirical mean is exactly zero. This leads to the following definition of the empirical correlation matrix, often found in the literature:

$$\check{E}_{ij} \;=\; \frac{1}{T-1} \sum_{t=1}^{T} \Big( Y_{it} - \overline{Y_i} \Big) \Big( Y_{jt} - \overline{Y_j} \Big), \qquad \overline{Y_i} = \frac{1}{T} \sum_{\tau=1}^{T} Y_{i\tau}. \tag{4.1.7}$$

which is clearly unbiased as for $T \to \infty$ with $N$ fixed. This can be rewritten as:

$$\check{\mathbf{E}} = \frac{1}{T-1} \mathbf{Y} \left( \mathbf{I}_T - \mathbf{e}\mathbf{e}^* \right) \mathbf{Y}^*, \qquad \mathbf{e} := (1, 1, \ldots, 1)^* / \sqrt{T} \in \mathbb{R}^T.$$

Still, the asymptotic properties of the eigenvalues (and eigenvectors) of $\mathbf{E}$ and of $\check{\mathbf{E}}$ are identical, up to a possible extra outlier eigenvalue located at zero when $q > 1$. The simplest way to understand that the outlier has no influence on the asymptotic behavior of the spectrum is when $\mathbf{Y}$ is a Gaussian matrix. In this case, we know that a Gaussian matrix is statistically invariant under rotation so one can always rotate the vector $\mathbf{e}$ in the $T$ dimensional space such that it becomes, say, $(1, 0, \ldots, 0)$. Then one has:

$$\check{E}_{ij} \sim \frac{1}{T-1} \sum_{t=2}^{T} \mathbf{Y}_{it} \mathbf{Y}_{jt}$$

which means that $\check{\mathbf{E}}$ and $\mathbf{E}$ share identical statistically properties when $N, T \to \infty$ up to a rank one perturbation of eigenvalue $\sim T^{-1} \to 0$ (see Section 3.1.3 for a related discussion). For $q < 1$,

this has no influence at all on the spectrum since the corresponding eigenvalue is reabsorbed in the bulk. The possible spike associated to the rank-one perturbation only survives when $N \geqslant T$, and it leads to an extra zero eigenvalue from the last equation. But in the case where $q \geqslant 1$, we know that there are $(N - T)$ additional zero eigenvalues, meaning that the extra spike at the origin is harmless. The case where $\mathbf{Y}$ is not rotationally invariant is harder to tackle and needs more sophisticated arguments for which we refer the reader to [27, Section 9] for more details. As a consequence, all the results concerning the statistics of the eigenvalues of $\mathbf{E}$ that we shall review below hold for $\check{\mathbf{E}}$ as well. From a practical point of view, it is indifferent to consider raw data or demeaned data. We will henceforth assume that the samples data $(\mathbf{y}_1, \ldots, \mathbf{y}_T)$ has exactly zero mean and will work only with $\mathbf{E}$ in the next sections.

**4.1.3. Distribution of the data entries.** The second remark deals with the distribution of the entries of the matrix $\mathbf{Y}$ given in Eq. (4.1.5). It is well-known for instance that financial returns are strongly non-Gaussian, with power-law tails [30], and hence, the condition of a sufficient number of bounded moments can be seen as restrictive. What can be said in the case of entries that possess extremely fat tails? This is the main purposes of the theory *robust* estimators [93, 127] where the RMT regime $N \asymp T$ has been subject to a lot studies in the past few years, especially in the case of elliptical distributions [26, 51, 57, 74]. In particular, the so-called *Maronna* robust $M$-estimator of $\mathbf{C}$ is the (unique) solution of the fixed point equation

$$\mathbf{M} := \frac{1}{T} \sum_{t=1}^{T} U\Big( \frac{1}{N} \mathbf{y}_t^* \mathbf{M}^{-1} \mathbf{y}_t \Big) \mathbf{y}_t \mathbf{y}_t^*, \tag{4.1.8}$$

where $U$ is a non-increasing function. It was shown recently [59] that the matrix $\mathbf{M}$ converges to a matrix of the form encountered in Eq. (3.1.80) and thus different from $\mathbf{E}$. However, tractable formula are scarce except for the multivariate Student distribution where $U(x) = x^{-1}$ [26, 74, 178, 197]. In that case, we have from [58] that $\mathbf{M}$ converges (almost surely) to $\mathbf{E}$ as $N \to \infty$. Therefore, all the results that we will present below holds for the robust estimator of $\mathbf{C}$ under a multivariate Student framework (see also [26]). We postpone discussions about other class of distributions to Chapter 10.

## 4.2 Bulk statistics

**4.2.1. Marčenko-Pastur equation.** As we alluded in the introduction, the fundamental tool to analyze the spectrum of large sample covariance matrices is the Marčenko-Pastur equation [123]. We actually have already encountered a special case of this equation in Section 3.1.2 where we consider the LSD of $\mathbf{E}$ under the null hypothesis, isotropic case $\mathbf{C} = \mathbf{I}_N$. In this section, we allow the population correlation matrix $\mathbf{C}$ to be *anistropic*, that is to say not proportional to the identity matrix. As we shall see, the final result is not as simple as Eq. (3.1.41) but many properties can be inferred from it.

The Marčenko-Pastur (MP) equation dates back to their seminal paper [123] which gives an exact relation between the limiting Stieltjes transforms of $\mathbf{E}$ and $\mathbf{C}$. This result is at the heart of many advances in statistical inference in high dimension (see Chapter 8 for some examples or [145] and references therein). There are several ways to obtain this result, using e.g. recursion techniques [162], Feynman diagram expansion [44], replicas (see [157] or Section 3.1.4 above for a generalization) or free probability. We will present this last approach, which is perhaps the simplest way to derive the MP equation.

The key observation is that, for linear models, we can always rewrite $\mathbf{E}$ using Eq. (4.1.4) as

$$\mathbf{E} = \sqrt{\mathbf{C}}\boldsymbol{\mathcal{W}}\sqrt{\mathbf{C}}, \qquad \boldsymbol{\mathcal{W}} := \mathbf{X}\mathbf{X}^*,$$

where the matrix $\mathbf{X}$ satisfies Eq. (4.1.5) and is independent from $\mathbf{C}$. After some contemplations, it becomes obvious that the model falls down into the model of free multiplication encountered in Section 3.1.3 since $\mathbf{E}$ is the free multiplicative convolution of $\mathbf{C}$ with a white Wishart kernel for $N \to \infty$ [135]. Therefore, the Stieltjes transform of $\mathbf{E}$ is exactly given by Eq. (3.1.84) that we specialize to

$$z\mathfrak{g}_{\mathbf{E}}(z) \;=\; Z(z)\mathfrak{g}_{\mathbf{C}}\left(Z(z)\right), \qquad \text{with} \qquad Z(z) \;:=\; z\mathcal{S}_{\boldsymbol{\mathcal{W}}}(z\mathfrak{g}_{\mathbf{E}}(z)-1). \tag{4.2.1}$$

Moreover, the $\mathcal{S}$-transform of $\boldsymbol{\mathcal{W}}$ was obtained in Eq. (3.1.44), i.e. $\mathcal{S}_{\boldsymbol{\mathcal{W}}}(z) = (1+qz)^{-1}$ for any $q > 0$. Thus, we can reexpress $Z(z)$ as:

$$Z(z) \;=\; \frac{z}{1-q+qz\mathfrak{g}_{\mathbf{E}}(z)}, \tag{4.2.2}$$

which is exactly the Marčenko-Pastur self-consistent equation which relates the Stieltjes transforms of $\mathbf{E}$ and $\mathbf{C}$. The remarkable thing is that the RHS of Eq. (4.2.1) is "deterministic" as $\mathbf{C}$ is fixed in this framework. Note that this equation is often written in the mathematical and statistical literature in an equivalent way as:

$$\mathfrak{g}_{\mathbf{E}}(z) = \int \frac{\rho_{\mathbf{C}}(\mu)\mathrm{d}\mu}{z-\mu(1-q+qz\mathfrak{g}_{\mathbf{E}}(z))}. \tag{4.2.3}$$

There are two ways to interpret the above Marčenko-Pastur equation:

1. the 'direct' problem: we know $\mathbf{C}$ and we want to compute the expected eigenvalues density $\rho_{\mathbf{E}}$ of the empirical correlation matrix;

2. the 'inverse' problem: we observe $\mathbf{E}$ and try to infer the true $\mathbf{C}$ that satisfies equation (4.2.1).

Obviously, the inverse problem is the one of interest for many statistical applications, but is much more difficult to solve than the direct one as the mapping between $\mathfrak{g}_{\mathbf{C}}$ from $\mathfrak{g}_{\mathbf{E}}$ is numerically unstable. Still, the work of El-Karoui [104] and, more recently, of Ledoit & Wolf [116] allows one to make progress in this direction with a numerical scheme that solves a discretized version of the inverse problem Eq. (4.2.3). On the other hand, the direct problem leads to a self-consistent equation, which can be exactly solved numerically and sometimes analytically for some special forms of $\mathfrak{g}_{\mathbf{C}}$ (see next section).

Let us finally make a remark that we have not seen in the literature before. Enhancing $Z(z)$ to $Z(z,q)$ to emphasize its dependence on $q$, one can check that this object obeys the following simple PDE:

$$\frac{\partial Z(z,q)}{\partial q} = (Z(z,q)-z)\frac{\partial Z(z,q)}{\partial z}, \tag{4.2.4}$$

with initial condition $Z(z, q \to 0) = z + q(1 - z\mathfrak{g}_{\mathbf{C}}(z))$. Whether this representation has a direct interpretation and is useful numerically or analytically remains to be seen.

**4.2.2. Spectral statistics of the sample covariance matrix.** For statistical purposes, the Marčenko-Pastur equation provides an extremely powerful framework to understand the behavior of large dimensional sample covariance matrices, despite the fact that the inverse problem is not numerically stable. As we shall see in this section, one can infer many properties of the spectrum of $\mathbf{E}$ knowing that of $\mathbf{C}$, using the moment generating function. Recall the definition of the $\mathcal{T}$-transform in Eq. (3.1.21), it is easy to see that we can rewrite Eq. (4.2.1) as

$$\mathcal{T}_{\mathbf{E}}(z) \;=\; \mathcal{T}_{\mathbf{C}}(Z(z)), \qquad Z(z) \;=\; \frac{z}{1+q\mathcal{T}_{\mathbf{E}}(z)}. \tag{4.2.5}$$

We know from Eq. (3.1.22) that the $\mathcal{T}$-transform can be expressed as power series for $z \to \infty$, hence we have

$$\mathcal{T}_{\mathbf{E}}(z) \underset{z\to\infty}{=} \sum_{k=1}^{\infty} \varphi(\mathbf{E}^k) z^{-k}, \tag{4.2.6}$$

where $\varphi(.) = N^{-1}\mathrm{Tr}(.)$ is the normalized trace operator. We thus deduce that

$$Z(z) \underset{z\to\infty}{=} \frac{z}{1+q\sum_{k=1}^{\infty}\varphi(\mathbf{E}^k)z^{-k}}.$$

Therefore we have for $z \to \infty$

$$\mathcal{T}_{\mathbf{C}}(Z(z)) \underset{z\to\infty}{=} \sum_{k=1}^{\infty} \frac{\varphi(\mathbf{C}^k)}{z^k}\left(1+q\sum_{\ell=1}^{\infty}\varphi(\mathbf{E}^\ell)z^{-\ell}\right)^k. \tag{4.2.7}$$

All in all, one can thus relate the moments of $\rho_{\mathbf{E}}$ with the moments of $\rho_{\mathbf{C}}$ by taking $z \to \infty$ in Eq. (4.2.5) which yields

$$\sum_{k=1}^{\infty}\frac{\varphi(\mathbf{E}^k)}{z^k} = \sum_{k=1}^{\infty}\frac{\varphi(\mathbf{C}^k)}{z^k}\left(1+q\sum_{\ell=1}^{\infty}\varphi(\mathbf{E}^\ell)z^{-\ell}\right)^k, \tag{4.2.8}$$

which was first obtained in [44]. In particular, we infer from Eq. (4.2.8) that the first three moments of $\rho_{\mathbf{E}}$ satisfy

$$\begin{aligned}
\varphi(\mathbf{E}) &= \varphi(\mathbf{C}) &&= 1\\
\varphi(\mathbf{E}^2) &= \varphi(\mathbf{C}^2) &&+ q\\
\varphi(\mathbf{E}^3) &= \varphi(\mathbf{C}^3) &&+ 3q\varphi(\mathbf{C}^2) + q^2.
\end{aligned} \tag{4.2.9}$$

We thus see that the variance of the LSD of $\mathbf{E}$ is equal to that of $\mathbf{C}$ plus $q$, i.e. the spectrum of the sample covariance matrix $\mathbf{E}$ is always be wider (for $q > 0$) than the spectrum of the population covariance matrix $\mathbf{C}$. This an alternative way to convince ourself that $\mathbf{E}$ is a noisy estimator of $\mathbf{C}$ in the high-dimensional regime.

Note that we can also express the Marčenko-Pastur equation in terms of a cumulant expansion. Indeed, we can rewrite Eq. (4.2.1) in terms of the $\mathcal{R}$-transform (see below for a derivation)

$$\omega\mathcal{R}_{\mathbf{E}}(\omega) = \zeta(\omega)\mathcal{R}_{\mathbf{C}}(\zeta(\omega)), \qquad \zeta(\omega) = \omega\big(1+q\omega\mathcal{R}_{\mathbf{E}}(\omega)\big). \tag{4.2.10}$$

Using the cumulants expansion of the $\mathcal{R}$-transform, given in Eq. (3.1.19), we obtain for $\omega \to 0$

$$\omega\mathcal{R}_{\mathbf{E}}(\omega) = \sum_{\ell=1}^{\infty}\kappa_\ell(\mathbf{E})\omega^\ell, \tag{4.2.11}$$

and

$$\zeta(\omega)\mathcal{R}_{\mathbf{C}}(\zeta(\omega)) = \sum_{\ell=1}^{\infty} \kappa_\ell(\mathbf{C})\omega^\ell \left(1 + q\sum_{m=1}^{\infty} \kappa_m(\mathbf{E})\omega^m\right)^\ell. \tag{4.2.12}$$

By regrouping these last two equations into Eq. (4.2.10), the analogue of Eq. (4.2.8) in terms of free cumulants reads:

$$\sum_{\ell=1}^{\infty} \kappa_\ell(\mathbf{E})\omega^\ell = \sum_{\ell=1}^{\infty} \kappa_\ell(\mathbf{C})\omega^\ell \left(1 + q\sum_{m=1}^{\infty} \kappa_m(\mathbf{E})\omega^m\right)^\ell, \tag{4.2.13}$$

which would allow one to express the cumulants of $\mathbf{E}$ in terms of the cumulants of $\mathbf{C}$.

Another interesting expansion is the case where $q < 1$, meaning that $\mathbf{E}$ is invertible. Hence $\mathfrak{g}(z)$ for $z \to 0$ is analytic and one can readily find

$$\mathfrak{g}(z) \underset{z\to 0}{=} -\sum_{k=1}^{\infty} \varphi(\mathbf{E}^{-k}) z^{k-1}. \tag{4.2.14}$$

This allows one to study the moment of the LSD of $\mathbf{E}^{-1}$ and this turns out to be an important quantity many applications (see Chapter 8). Using Eq. (4.2.1), we can actually relate the moments of the spectrum $\mathbf{E}^{-1}$ to those of $\mathbf{C}^{-1}$ as one has, for $z \to 0$:

$$Z(z) = \frac{z}{1 - q - q\sum_{k=1}^{\infty} \varphi(\mathbf{E}^{-k}) z^k}.$$

Hence, we obtain the following expansion for Eq. (4.2.1) at $z \to 0$ and $q \in (0, 1)$:

$$\sum_{k=1}^{\infty} \varphi(\mathbf{E}^{-k}) z^k = \sum_{k=1}^{\infty} \varphi(\mathbf{C}^{-k}) \left(\frac{z}{1-q}\right)^k \left(\frac{1}{1 - \frac{q}{1-q}\sum_{\ell=1}^{\infty} \varphi(\mathbf{E}^{-\ell}) z^\ell}\right)^k, \tag{4.2.15}$$

that is a little bit more cumbersome than the moment generating expansion Eq. (4.2.8) or the cumulant expansion (4.2.13). Still, we get at leading order that

$$\varphi(\mathbf{E}^{-1}) = \frac{\varphi(\mathbf{C}^{-1})}{1-q}, \qquad \varphi(\mathbf{E}^{-2}) = \frac{\varphi(\mathbf{C}^{-2})}{(1-q)^2} + \frac{q\varphi(\mathbf{C}^{-1})^2}{(1-q)^3}. \tag{4.2.16}$$

We will see in Section 8.1 that the first relation has direct consequences for the out-of-sample risk of optimized portfolios.

Let us now give a formal derivation of Eq. (4.2.10). Let us define

$$\omega = \mathfrak{g}_{\mathbf{E}}(z), \qquad \zeta = \mathfrak{g}_{\mathbf{C}}(Z), \tag{4.2.17}$$

which allows us to rewrite Eq. (4.2.1) as

$$\omega\mathcal{B}_{\mathbf{E}}(\omega) = \zeta\mathcal{B}_{\mathbf{C}}(\zeta), \qquad Z \equiv \mathcal{B}_{\mathbf{C}}(\zeta) = \frac{\mathcal{B}_{\mathbf{E}}(\omega)}{1 - q + q\omega\mathcal{B}_{\mathbf{E}}(\omega)}. \tag{4.2.18}$$

Then, using the definiton (3.1.16) of the $\mathcal{R}$-transform, we can rewrite this last equation as

$$\omega\mathcal{R}_{\mathbf{E}}(\omega) = \zeta\mathcal{R}_{\mathbf{C}}(\zeta), \qquad \mathcal{R}_{\mathbf{C}}(\zeta) + \frac{1}{\zeta} = \frac{\mathcal{R}_{\mathbf{E}}(\omega) + 1/\omega}{1 + q\omega\mathcal{R}_{\mathbf{E}}(\omega)}. \tag{4.2.19}$$

We deduce that

$$\mathcal{R}_{\mathbf{C}}(\zeta) = \frac{\mathcal{R}_{\mathbf{E}}(\omega) + 1/\omega}{1 + q\omega\mathcal{R}_{\mathbf{E}}(\omega)} - \frac{1}{\zeta}, \tag{4.2.20}$$

and it is not hard to see that this yields

$$\omega\mathcal{R}_{\mathbf{E}}(\omega) = \zeta\left(\frac{\mathcal{R}_{\mathbf{E}}(\omega) + 1/\omega}{1 + q\omega\mathcal{R}_{\mathbf{E}}(\omega)} - \frac{1}{\zeta}\right). \tag{4.2.21}$$

By re-arranging the terms in this last equation, we obtain

$$\omega\mathcal{R}_{\mathbf{E}}(\omega) + 1 = \frac{\zeta}{\omega}\left(\frac{\omega\mathcal{R}_{\mathbf{E}}(\omega) + 1}{1 + q\omega\mathcal{R}_{\mathbf{E}}(\omega)}\right), \tag{4.2.22}$$

that is to say

$$\zeta \equiv \zeta(\omega) = \omega\big(1 + q\omega\mathcal{R}_{\mathbf{E}}(\omega)\big), \tag{4.2.23}$$

and Eq. (4.2.10) immediately follows by plugging this last equation into Eq. (4.2.20).

**4.2.3. Dual representation and edges of the spectrum.** Although a lot of informations about the spectrum of $\mathbf{E}$ can be gathered from the Marčenko-Pastur equation (4.2.1), the equation itself is not easy to solve analytically. In particular, what can be said about the edges of the spectrum of $\mathbf{E}$? We shall see that we are able to answer some of these questions by using a dual representation of Eq. (4.2.1).

The "dual" representation that we are speaking about comes from studying the $T \times T$ matrix $\mathbf{S}$:

$$\mathbf{S} := \frac{1}{T}\mathbf{Y}^*\mathbf{Y} \equiv \mathbf{X}^*\mathbf{C}\mathbf{X}, \tag{4.2.24}$$

where we used Eq. (4.1.4) in the last equation. The dual matrix $\mathbf{S}$ can also be interpreted as a correlation matrix. In a financial context, $\mathbf{E}$ tells us how similar is the movement of two stocks over time, while $\mathbf{S}$ tells us how similar are two dates in terms of the overall movements of the stocks on these two particular dates. Using a singular value decomposition, it is not difficult to show that $\mathbf{S}$ and $\mathbf{E}$ share the same non-zero eigenvalues – hence the "duality". In the case where $T > N$, the matrix $\mathbf{S}$ has a zero eigenvalue with multiplicity $T - N$ in addition to the eigenvalues $\{\lambda_i\}_{i\in[\![1,N]\!]}$ of $\mathbf{E}$. Therefore, it is easy to deduce the Stieltjes transform of $\mathbf{S}$ (for $q > 1$):

$$\mathfrak{g}_{\mathbf{S}}(z) = \frac{1}{T}\left[\frac{T-N}{z} + N\mathfrak{g}_{\mathbf{E}}(z)\right] = \frac{1-q}{z} + q\mathfrak{g}_{\mathbf{E}}(z) = \frac{1}{Z(z)}. \tag{4.2.25}$$

The introduction of this dual representation of the empirical matrix allows one to get the following expression from Eq. (4.2.3):

$$\mathfrak{g}_{\mathbf{S}}(z) = \frac{1}{z}\left(1 - q + q\int \frac{\rho_{\mathbf{C}}(\mu)d\mu}{1 - \mu\mathfrak{g}_{\mathbf{S}}(z)}\right).$$

After some manipulations, we can rewrite this last equation as

$$z = \frac{1}{\mathfrak{g}_{\mathbf{S}}(z)} + q\int \frac{\rho_{\mathbf{C}}(\mu)d\mu}{\mu^{-1} - \mathfrak{g}_{\mathbf{S}}(z)}. \tag{4.2.26}$$

Writing $z = \mathcal{B}_{\mathbf{S}}(\mathfrak{g}_{\mathbf{S}}(z))$ in the above equation, we obtain a characterization of the functional inverse of $\mathfrak{g}_{\mathbf{S}}$ as

$$\mathcal{B}_{\mathbf{S}}(\omega) := \frac{1}{\omega} + q\int \frac{\rho_{\mathbf{C}}(\mu)d\mu}{\mu^{-1} - \omega}, \tag{4.2.27}$$

and this is the dual representation of the Marčenko-Pastur equation (4.2.1). The analytic behavior of this last equation has been the subject of several studies, especially in [163]. In particular, it was proved that there exists a *unique* $\omega \in \mathbb{C}_+$ that solves the equation (4.2.27). This yields the Stieltjes transform of $\mathbf{S}$ from which we re-obtain the Stieltjes transform of $\mathbf{E}$ using Eq. (4.2.25). We will see in the next section that the dual representation (4.2.27) of the Marčenko-Pastur equation is particularly useful when we will try to solve the direct problem.

In addition, the position of the edges of the LSD of $\mathbf{E}$ can be inferred from Eq. (4.2.27). Within a one cut-assumption, the edges of the support of $\rho_{\mathbf{E}}$ are given by:

$$\lambda_{\pm}^{\mathbf{E}} = \mathcal{B}_{\mathbf{S}}(\omega_{\pm}) \quad \text{where } \omega_{\pm} \in \mathbb{R}^+ \text{ is such that } \mathcal{B}'_{\mathbf{S}}(\omega_{\pm}) = 0. \tag{4.2.28}$$

Indeed, knowing the spectral density of $\mathbf{S}$ allows us to get the spectral density of $\mathbf{E}$ since from Eq. (4.2.25) one gets:

$$\rho_{\mathbf{S}}(\lambda) = q\rho_{\mathbf{E}}(\lambda) + (1-q)^+\delta_0, \tag{4.2.29}$$

for any $\lambda \in \operatorname{supp}\rho_{\mathbf{S}}$. Next, and one easily obtains

$$\mathfrak{g}'_{\mathbf{S}}(z) = -\int \frac{\rho_{\mathbf{S}}(x)\mathrm{d}x}{(z-x)^2} < 0, \tag{4.2.30}$$

for any $z \notin \operatorname{supp}\rho_{\mathbf{S}}$, meaning that it is strictly decreasing outside of the support. We saw in Section 3.1.1 that the Stieltjes transform $\mathfrak{g}(z)$ is analytical and positive for any $z \in \mathbb{R}$ outside of the support. Moreover, for $z \to \infty$, we have $\mathfrak{g}_{\mathbf{S}}(z) \sim z^{-1} + \mathcal{O}(z^{-2})$ so that we deduce $\mathfrak{g}_{\mathbf{S}}(z)$ is a bijective decreasing function. Its inverse function $\mathcal{B}_{\mathbf{S}}$ therefore also decreases in those same intervals. Consequently, the union of intervals where $\mathcal{B}_{\mathbf{S}}(x)$ is decreasing will lead to the complement of the support and the edges of the support of $\rho_{\mathbf{S}}$ are thus given by the critical points of $\mathcal{B}_{\mathbf{S}}$, as in Eq. (4.2.28) above. If one assumes that there are a finite number $r$ of (non-denegerate) spikes, we can readily generalize the above arguments and find that there will be $2(r+1)$ critical points (see Figure 4.2.1 for an illustration with two non-degenerate spikes).

**4.2.4. Solving Marčenko-Pastur equation.** In this section, we investigate the direct problem of solving the Marčenko-Pastur equation Eq. (4.2.1) for $\mathfrak{g}_{\mathbf{E}}$ given $\mathfrak{g}_{\mathbf{C}}$. We will discuss briefly the inverse problem at the end of this section.

**Exactly solvable cases.** As far as we know, there are only a few cases where we can find an explicit expression for the LSD of $\mathbf{E}$. The first one is trivial: it is when one considers the "classical" limit in statistics where $T \to \infty$ for a fixed value of $N$. In this case $q = 0$ in (4.2.3), and obviously $\mathfrak{g}_{\mathbf{E}}(z) = \mathfrak{g}_{\mathbf{C}}(z)$ in this case, as expected.

However, for any finite observation ratio $q > 0$, we anticipate from the discussion of Section 4.2.2 above that the LSD of $\mathbf{E}$ will be significantly different from that of $\mathbf{C}$. The influence of $q$ can be well understood in the simple case where $\mathbf{C} = \mathbf{I}_N$. We know from Section 3.1.2 that this case is exactly solvable and the LSD of $\mathbf{E}$ is the well-known Marčenko-Pastur law (3.1.42), that we recall here:

$$\mathfrak{g}_{\mathbf{E}}(z) = \frac{z+1-q-\sqrt{z-\lambda_-^{\mathrm{mp}}}\sqrt{z-\lambda_+^{\mathrm{mp}}}}{2qz}, \qquad \lambda_{\pm}^{\mathrm{mp}} = (1\pm\sqrt{q})^2 \tag{4.2.31}$$

In words, the sample eigenvalues spans the interval $[(1-\sqrt{q})^2, (1+\sqrt{q})^2]$ while the population eigenvalues are all equal to unity. We therefore deduce that the rms of the sample eigenvalue distribution is order $\sqrt{q}$, highlighting the systematic bias in the estimation of the eigenvalues

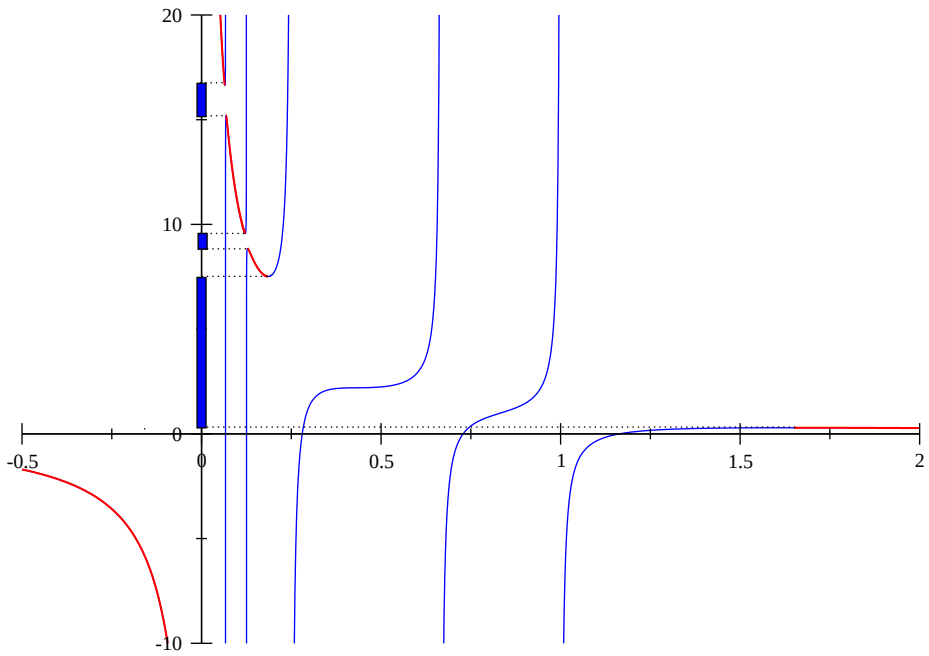



FIGURE 4.2.1. The function $B_{\underline{\mathbf{E}}}(x)$ with population eigenvalues density given by $0.002\,\delta_{15} + 0.002\,\delta_8 + 0.396\,\delta_3 + 0.3\,\delta_{1.5} + 0.3\,\delta_1$. Here $T = 1000$, $N = 500$ and we have 3 connected components. The vertical asymptotes are located at each $-x^{-1}$ for $x \in \{1, 1.5, 3, 8, 15\}$. The support of $\rho_{\mathbf{S}}$ is indicated with thick blue lines on the vertical axis. The inverse of $\mathfrak{g}_{\mathbf{S}}|_{\mathbb{R}\setminus\operatorname{supp}\varrho_{\mathbf{S}}}$ is drawn in red.

using $\mathbf{E}$ when $q = \mathcal{O}(1)$. This effect can be visualized using the quantile representation of the spectral distribution. Indeed, it is known since [27, 109] that the bulk eigenvalues $[\lambda_i]_{i\in[\![r+1,N]\!]}$ converge in the high-dimensional regime to their "average positions" $[\gamma_i]_{i\in[\![r+1,N]\!]}$. More precisely, this reads:

$$\lambda_i \approx \gamma_i, \quad \text{where} \quad \frac{i}{N} = \int^{\gamma_i} \rho_{\mathbf{E}}(\lambda)\mathrm{d}\lambda, \qquad i \geqslant r+1\,. \tag{4.2.32}$$

We plot the $\gamma_i$'s of the Marčenko-Pastur law in Fig. 4.2.2 for $q = 1/4$ and $q = 1/2$, and observe systematic and significant deviations from the "classical" positions $\gamma_i^{q=0} \equiv 1$. This again illustrates that $\mathbf{E}$ is an untrustworthy estimator when the sample size is of the same order of magnitude than the number of variables.

Now that the qualitative impact of the observation ratio $q$ is well understood, a natural extension would be to examine the Marčenko-Pastur equation for a non trivial correlation matrix $\mathbf{C}$. To this aim, we now consider another interesting solvable case, especially for statistical inference, which is the case and of an (isotropic) inverse Wishart matrix with hyperparameter $\kappa > 0$. From Section 3.1.2, we recall that

$$\mathcal{S}_{\mathbf{C}}(\omega) = 1 - \frac{\omega}{2\kappa},$$

for $\kappa > 0$. Then, using the free multiplication formula (3.1.81), we have $\mathcal{S}_{\mathbf{E}}(\omega) = \mathcal{S}_{\mathbf{C}}(\omega)\mathcal{S}_{\mathbf{W}}(\omega)$ where $\mathcal{S}_{\mathbf{W}}(\omega)$ is given in (3.1.44) which yields a quadratic equation in $\mathcal{T}_{\mathbf{E}}(z)$. This means that

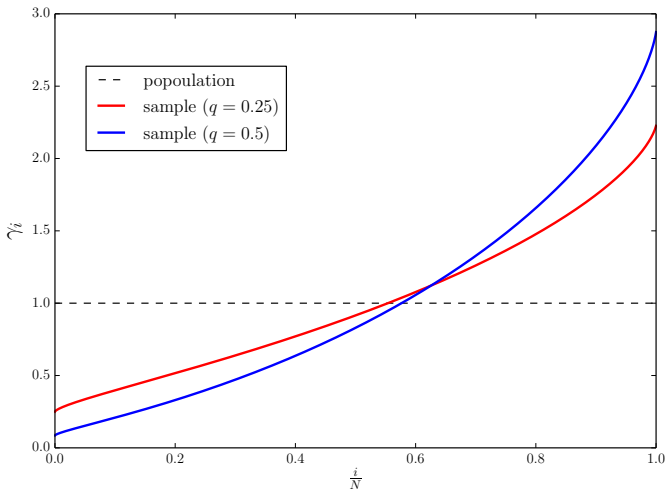



FIGURE 4.2.2. Typical position of the sample eigenvalues under the Marčenko-Pastur law (3.1.42) with a finite observation ratio $q = 0.25$ (red line) and $q = 0.5$ (blue line). The dotted line corresponds to the locations of the population eigenvalues and we see a significant deviation.

$\mathfrak{g}_{\mathbf{E}}$ is explicit and reads:

$$\mathfrak{g}_{\mathbf{E}}(z) = \frac{z(1+\kappa) - \kappa(1-q) \pm \sqrt{(\kappa(1-q) - z(1+\kappa))^2 - z(z+2q\kappa)(2\kappa+1)}}{z(z+2q\kappa)}, \tag{4.2.33}$$

from which we can retrieve the edges of the support:

$$\lambda_{\pm}^{\text{iw}} = \frac{1}{\kappa}\left[(1+q)\kappa + 1 \pm \sqrt{(2\kappa+1)(2q\kappa+1)}\right]. \tag{4.2.34}$$

One can check that the limit $\kappa \to \infty$ recovers the null hypothesis case $\mathbf{C} = \mathbf{I}_N$; the lower $\kappa$, the wider the spectrum of $\mathbf{C}$. We plot in Figure 4.2.3 the spectral density $\rho_{\mathbf{C}}$ and $\rho_{\mathbf{E}}$ for $q = 0.25$ and $q = 0.5$ as a function of the eigenvalues. Again, we see that the spectral density of $\mathbf{E}$ puts significant weights on regions of the real axis which are outside the support of $\rho_{\mathbf{C}}$, due to the measurement noise. From an inference theoretic viewpoint, the interest of the Inverse-Wishart ensemble is to provide a parametric prior distribution for $\mathbf{C}$ where everything can be computed analytically (see Chapter 6 below for some applications).

There exist several other examples where the Marčenko-Pastur equation is exactly solvable even though the Stieltjes transform is not explicit. For instance, if we consider $\mathbf{C}$ to be a Wishart matrix of parameter $q_0$ independent from $\mathbf{W}$, then we have from (3.1.81) that

$$\mathcal{S}_{\mathbf{E}}(\omega) = \frac{1}{(1+q_0\omega)(1+q\omega)}.$$

It is then easy to see from the definition (3.1.23) that $\mathcal{T}_{\mathbf{E}}(z) \equiv \omega(z)$ is solution of the cubic equation,

$$z(1+\omega(z))(1+q_0\omega(z))(1+q\omega(z)) - \omega(z) = 0, \tag{4.2.35}$$

from which we obtain $\mathfrak{g}_{\mathbf{E}}(z)$ thanks to (3.1.21) and by choosing the unique solution of the latter equation in $\mathbb{C}^+$ (see the following section for details on this point). Another toy example that uses the Marčenko-Pastur with the $\mathcal{R}$-transform formalism is when $\mathbf{C}$ is a GOE centered around the identity matrix. In this case we have

$$\mathcal{R}_{\mathbf{C}}(\omega) = 1 + \sigma^2\omega, \tag{4.2.36}$$

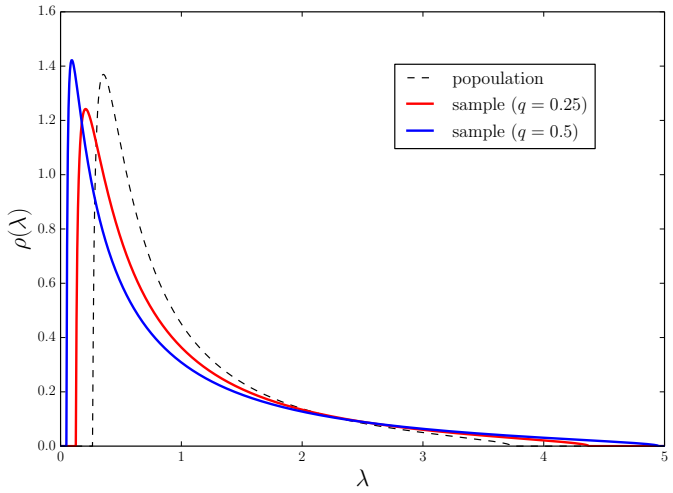



FIGURE 4.2.3. Solution of the Marčenko-Pastur equation for the eigenvalue distribution of $\mathbf{E}$ when $\mathbf{C}$ is an inverse Wishart matrix with parameter $\kappa = 1.0$ for $q = 0.25$ (red line) and $q = 0.5$ (blue line). The black dotted line corresponds to the LSD $\rho_{\mathbf{C}}$.

where we add the constraint $\sigma \leqslant 0.5$ such that $\mathbf{C}$ remains a positive semi-definite matrix. Then, by plugging this formula into (4.2.10), we find that $\mathfrak{g}_{\mathbf{E}}(z) = \omega$ is the solution of quartic equation:

$$\sigma^2\omega^2(1 + q\omega\mathcal{R}_{\mathbf{E}}(\omega))^2 + \omega(1 + q\omega\mathcal{R}_{\mathbf{E}}(\omega)) - \omega\mathcal{R}_{\mathbf{E}}(\omega) = 0, \tag{4.2.37}$$

and as above, we take the unique solution in $\mathbb{C}^+$ in order to get the right Stieltjes transform.

**The general case: numerical method** Apart from the very specific cases discussed above, finding an explicit expression for $\mathfrak{g}_{\mathbf{E}}(z)$ is very difficult. This means that we have to resort to numerical schemes in order to solve the Marčenko-Pastur equation. In that respect, the dual representation (4.2.27) of Eq. (4.2.1) comes to be particularly useful. To solve the MP equation for a given $z$, we seek a $\mathfrak{g} \equiv \mathfrak{g}_{\mathbf{S}}$ such that[1]

$$z = \mathcal{B}_{\mathbf{S}}(\mathfrak{g}), \qquad \mathfrak{g} \in \mathbb{C}_+, \tag{4.2.38}$$

where the expression of $\mathcal{B}_{\mathbf{S}}$ in terms of $\rho_{\mathbf{C}}$ is explicit and given in Eq. (4.2.27). Numerically, the above equation is easily solved using a simple gradient descent algorithm, i.e. find $\mathfrak{g} \in \mathbb{C}_+$ such that

$$\begin{cases} \mathrm{Re}(z) = \mathrm{Re}\big[\mathcal{B}_{\mathbf{S}}(\mathfrak{g})\big] \\ \mathrm{Im}(z) = \mathrm{Im}\big[\mathcal{B}_{\mathbf{S}}(\mathfrak{g})\big], \end{cases} \tag{4.2.39}$$

It then suffices to use Eq. (4.2.25) in order to get $\mathfrak{g}_{\mathbf{E}}(z)$ for any $z \in \mathbf{C}_-$. Hence, if one wants to retrieve the eigenvalues density $\rho_{\mathbf{E}}$ at any point on the real line, we simply have to set $z = \lambda - \mathrm{i}\varepsilon$ with $\lambda \in \mathrm{Supp}(\mathbf{E})$ and $\varepsilon$ an arbitrary small real positive number into Eq. (4.2.39). Note that in the case where $\mathfrak{g}_{\mathbf{C}}$ is known, one can rewrite equation (4.2.27) as

$$\mathcal{B}_{\mathbf{S}}(x) = \frac{1}{x}\left[1 - q + \frac{q}{x}\mathfrak{g}_{\mathbf{C}}\left(\frac{1}{x}\right)\right], \tag{4.2.40}$$

[1] Recall that $\mathbf{S}$ is the $T \times T$ equivalent of $\mathbf{E}$ defined in Eq. (4.2.24).

which is obviously more efficient since we avoid to compute the integral over eigenvalues.

In order to illustrate this numerical scheme, let us consider a covariance matrix whose LSD has a heavy right tail. One possible parameterization is to assume a power-law distribution of the form [29]:

$$\rho_{\mathbf{C}}(\lambda) = \frac{sA}{(\lambda+\lambda_0)^{1+s}}\Theta(\lambda-\lambda_{\min}), \tag{4.2.41}$$

where $\Theta(x) = x^+$ is the Heaviside step function, $s$ is an exponent that we choose to be $s = 2$ [29], and $\lambda_{\min}$ the lower edge of the spectrum below which there are no eigenvalues of $\mathbf{C}$. $A, \lambda_{\min}$ are then determined by the two normalization constraints $\int \rho_{\mathbf{C}}(x)dx = 1$ and $\int x\rho_{\mathbf{C}}(x)dx = 1$. This leads to: $\lambda_{\min} = (1-\lambda_0)/2$ and $A = (1-\lambda_{\min})^2$. We see that $\lambda_{\min}$ may become negative for $\lambda_0 > 1$ and that $\rho_{\mathbf{C}}$ becomes singular for $\lambda_0 = 1$. From the density Eq. (4.2.41), one can perform the Stieltjes transform straightaway to find

$$\mathfrak{g}_{\mathbf{C}}(z) = \frac{1}{z+1-2\lambda_0} + \frac{2(1-\lambda_0)}{(z+1-2\lambda_0)^2} + \frac{2(1-\lambda_0)^2}{(z+1-2\lambda_0)^3}\left[\log\left(\frac{\lambda_0-z}{1-\lambda_0}\right)\right], \tag{4.2.42}$$

which allows one to solve Eq. (4.2.40) for $\mathfrak{g}_{\mathbf{E}}(z)$ with only a few iterations. As we observe in Fig. 4.2.4, the theoretical value obtained from the numerical scheme (4.2.39) agrees perfectly with the empirical results, obtained by diagonalizing matrices of size $N = 500$ matrices obtained as $\sqrt{\mathbf{C}}\boldsymbol{\mathcal{W}}\sqrt{\mathbf{C}}$, where $\boldsymbol{\mathcal{W}}$ is a Wishart matrix. This illustrates the robustness of the above numerical scheme, even when the spectrum of $\mathbf{C}$ is fat-tailed. In addition, we can notice that the more we add structure in the true covariance $\mathbf{C}$, the wider is the empirical distribution as in this "degenerate" case, the spectrum of $\mathbf{E}$ embraces nearly all the positive real number line.

## 4.3 Edges and outliers statistics

As we alluded to several times, the practical usefulness of the above predictions for the eigenvalue spectra of random matrices is (i) their universality with respect to the distribution of the underlying random variables and (ii) the appearance of sharp edges in the spectrum, meaning that the existence of eigenvalues lying outside the allowed region is a possible indication against simple "null hypothesis" benchmarks. Illustrating the last point, Fig. 4.3.1 shows the empirical spectral density of the correlation matrix corresponding to $N = 406$ and $T = 1300$ so that $q \approx 0.31$, compared to the simplest Marčenko-Pastur spectrum in the null hypothesis case $\mathbf{C} = \mathbf{I}_N$. While the bulk of the distribution is roughly accounted for (but see Section 8.2 for a much better attempt), there seems to exist a finite number of eigenvalues lying outside the Marčenko-Pastur sea, which may be called outliers or spikes. However, if there are no such spikes in the spectrum of $\mathbf{C}$, one expects to see, for finite $N$ some eigenvalues beyond the Marčenko-Pastur upper edge. The next two sections are devoted first to a discussion of these finite size effects, and then to a model with "true" outliers that survive in the large $N$ limit.

**4.3.1. The Tracy-Widom region.** This existence of sharp edges delimiting a region where one expects to see a non zero density of eigenvalues from a region where there should be none is only true in the asymptotic $N, T \to \infty$, and in the absence of "fat-tails" in the distribution of matrix elements (see [20, 25]). For large but finite $N$, on the other hand, one expects that the probability to find an eigenvalue beyond the Marčenko-Pastur "sea" is very small but finite. The width of the transition region, and the tail of the density of states was investigated already a while ago [33], culminating in the beautiful results by Tracy & Widom on the distribution

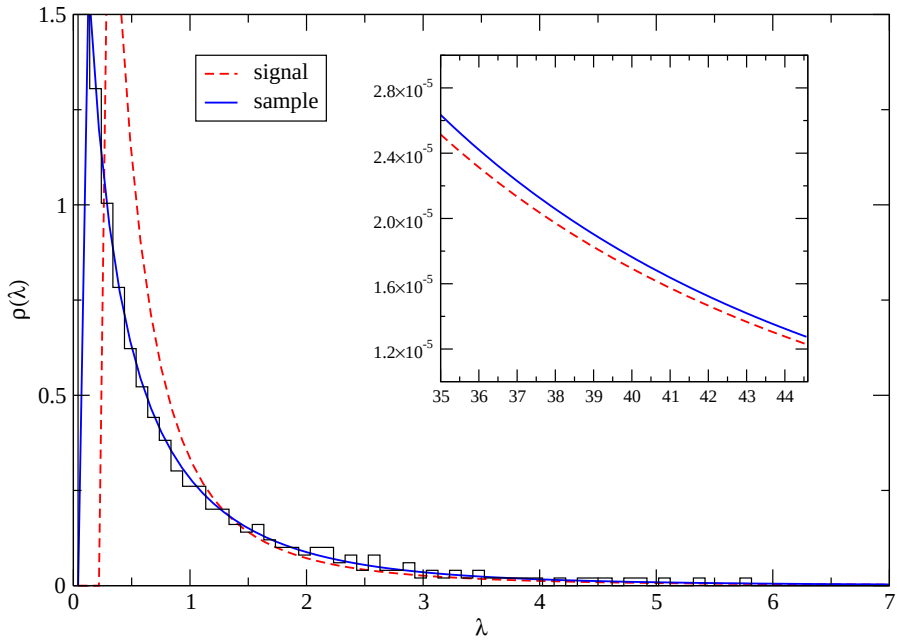



FIGURE 4.2.4. Effect of the Marčenko-Pastur equation when $\rho_{\mathbf{C}}$ is given a power law density with parameter $\lambda_0 = 0.3$ and a finite observation ratio $q = 0.5$ and $N = 500$. The dotted line corresponds to the LSD of $\mathbf{C}$ while the plain line corresponds to the LSD of $\mathbf{E}$. The histogram is the ESD when we compute $\mathbf{E}$ from the definition (4.1.3). The main figure covers the bulk of the eigenvalues while the inset zoom in the region of very large eigenvalues.

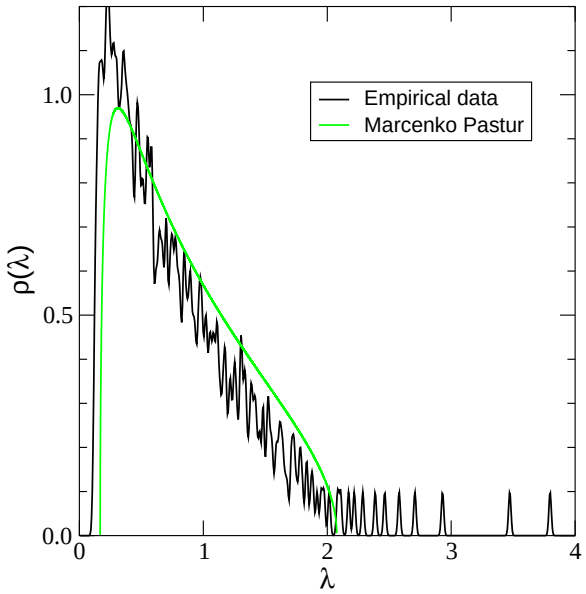



FIGURE 4.3.1. Test of the null hypothesis on the empirical correlation matrix $\mathbf{E}$ using US stocks' data with $N = 406$ and $T = 1300$.

of the *largest* eigenvalue of a random matrix [175]. The Tracy-Widom result is actually a very nice manifestation of the universality phenomenon that describes the fluctuations of macroscopic observables in many large dimensional systems (see the recent paper [120] on this topic). The derivation of the Tracy-Widom distribution mainly relies on Orthogonal polynomials that we will not discuss in this thesis (see e.g. [138,175]) but there exist also alternative approach [153]. The link between this limiting law and the largest eigenvalues of large sample covariance matrices has been subject to a large amount of studies that we will not attempt to cover here (see e.g. [15,62,98,100,122,148] for details and references).

The Tracy-Widom result characterizes precisely the distance between the largest eigenvalue $\lambda_1$ of $\mathbf{E}$ and the upper edge of the spectrum that we denoted by $\lambda_+$. This result can be (formally) stated as follows: the rescaled distribution of $\lambda_1 - \lambda_+$ converges towards the Tracy-Widom distribution, usually noted $F_1$,

$$\mathcal{P}\left(\lambda_{\max} \leqslant \lambda_+ + \gamma N^{-2/3} u\right) = F_1(u), \tag{4.3.1}$$

where $\gamma$ is a constant that depends on the problem. For the isotropic Marčenko-Pastur problem, $\lambda_+ = (1+\sqrt{q})^2$ and $\gamma = \sqrt{q}\lambda_+^{2/3}$, whereas for the Wigner problem, $\lambda_+ = 2$ and $\gamma = 1$. We stress that this result holds for a large class of $N \times N$ matrices (e.g. symmetric random matrices with IID elements with a finite fourth moment, see [20,25]).

Everything is known about the Tracy-Widom density $f_1(u) = F_1'(u)$, in particular its left and right far tails:

$$\ln f_1(u) \propto -u^{3/2}, \quad (u \to +\infty); \qquad \ln f_1(u) \propto -|u|^3, \quad (u \to -\infty); \tag{4.3.2}$$

One notices that the left tail is much thinner: pushing the largest eigenvalue inside the allowed band implies compressing the whole Coulomb gas of repulsive charges, which is difficult. Using this analogy, the large deviation regime of the Tracy-Widom problem (i.e. for $\lambda_{\max} - \lambda_+ = \mathcal{O}(1)$) can also be obtained [62].

Note that the distribution of the smallest eigenvalue $\lambda_{\min}$ around the lower edge $\lambda_-$ is also Tracy-Widom, except in the particular case of Marčenko-Pastur matrices with $q = 1$. In this case, $\lambda_- = 0$ which is a 'hard' edge since all eigenvalues of the empirical matrix must be non-negative. This special case is treated in, e.g. [147].

**4.3.2. Outlier statistics.** Now, there are cases where a finite number of eigenvalues genuinely reside outside the Marčenko-Pastur sea (or more generally outside of the "bulk" region) even when $N \to \infty$. For example, the empirical data shown in Fig. Fig. 4.3.1 indeed suggests the presence of true outliers, that have a real financial interpretation in terms of economic sectors of activity. Therefore, we need a framework to describe correlation matrices that contain both a bulk region and a finite number of "spikes". The purpose of this section is to study the statistics of these eigenvalues from an RMT point of view.

The standard way to treat outliers is to "dilate" a finite number of eigenvalues of a given (spikeless) correlation matrix $\underline{\mathbf{C}}$, that we construct as:

$$\underline{\mathbf{C}} = \sum_{i=1}^{N} \underline{\mu}_i \mathbf{v}_i \mathbf{v}_i^*, \quad \text{where} \quad \underline{\mu}_i = \begin{cases} \mu_0 & \text{if } i \leqslant r \\ \mu_i & \text{if } i \geqslant r+1\,. \end{cases} \tag{4.3.3}$$

We choose the eigenvalue $\mu_0$ within the spectrum of $\underline{\mathbf{C}}$ such that there is no outliers initially. Here we fix $\mu_0 = \mu_{r+1}$ for simplicity, but any other choice in the set $[\mu_i]_{i \geqslant r+1}$ would do equally

well. Then with this prescription, we may rewrite $\mathbf{C}$ as a small rank perturbation of $\underline{\mathbf{C}}$. Indeed, since each outlier $[\mu_i]_{i\leqslant r}$ are well separated from the bulk by assumption, we may parametrize each spike $\mu_i$ by a positive real number $d_i$ for any $i \leqslant r$ as follows:

$$\mu_i = \mu_0(1+d_i) \equiv \mu_{r+1}(1+d_i), \qquad d_i > 0\ ,\ i \leqslant r. \tag{4.3.4}$$

Hence, the population covariance matrix $\mathbf{C}$ is given by:

$$\mathbf{C} = \sum_{i=1}^{N} \mu_i \mathbf{v}_i \mathbf{v}_i^*, \quad \text{where} \quad \mu_i = \begin{cases} \mu_0(1+d_i) & \text{if } i \leqslant r \\ \mu_i & \text{if } i \geqslant r+1\,. \end{cases} \tag{4.3.5}$$

More synthetically, one can write $\mathbf{C}$ as:

$$\mathbf{C} = \underline{\mathbf{C}}\big(\mathbf{I}_N + \mathbf{V}^{(r)}\mathbf{D}\mathbf{V}^{(r)*}\big), \tag{4.3.6}$$

where $\mathbf{V}^{(r)} := [\mathbf{v}_1, \dots, \mathbf{v}_r] \in \mathbb{R}^{N\times r}$ and $\mathbf{D} := \operatorname{diag}(d_1, \dots, d_r)$ is a diagonal matrix that characterizes the spikes. We also define a fictitious spikeless sample covariance matrix as $\underline{\mathbf{E}} = \underline{\mathbf{C}}^{1/2}\mathbf{X}\mathbf{X}^*\underline{\mathbf{C}}^{1/2}$ and denote by $\underline{\mathbf{S}} = \mathbf{X}^*\underline{\mathbf{C}}\mathbf{X}$ the $T \times T$ its "dual" matrix. As noticed in [43], the statistics of the outliers of $\mathbf{E}$ can be investigated through that of $\underline{\mathbf{E}}$. Let us consider the rank-one $r = 1$ case for the sake of simplicity (see [43] for the general case). Then, we have

$$\det(z\mathbf{I}_N - \mathbf{E}) = \det(z\mathbf{I}_N - \mathbf{X}^*\underline{\mathbf{C}}(\mathbf{I}_N + d_1\mathbf{v}_1\mathbf{v}_1^*)\mathbf{X}) = \det(z\mathbf{I}_N - \mathbf{X}\mathbf{X}^*\underline{\mathbf{C}}(\mathbf{I}_N + d_1\mathbf{v}_1\mathbf{v}_1^*)).$$

which can be transformed into:

$$\det(z\mathbf{I}_N - \mathbf{E}) = \det(z\mathbf{I}_N - \underline{\mathbf{E}})\det(\mathbf{I}_N - d_1(z\mathbf{I}_N - \underline{\mathbf{E}})^{-1}\mathbf{v}_1\mathbf{v}_1^*\mathbf{E}) \tag{4.3.7}$$

We can conclude that $\lambda_1$ in an eigenvalue of $\mathbf{E}$ and not of $\underline{\mathbf{E}}$ if and only if the second determinant vanishes, i.e. if $d_1(\lambda_1\mathbf{I}_N - \underline{\mathbf{E}})^{-1}\mathbf{v}_1\mathbf{v}_1^*\underline{\mathbf{E}}$ has an eigenvalue equals to unity. To find $\lambda_1$, we remark that this second determinant is simply a rank-one update, meaning that it has only one non-trivial eigenvalue given by the equation:

$$d_1\left[\lambda_1\langle\mathbf{v}_1, \mathbf{G}_{\underline{\mathbf{E}}}(\lambda_1)\mathbf{v}_1\rangle - 1\right] = 1, \tag{4.3.8}$$

where $\mathbf{G}_{\underline{\mathbf{E}}}$ is the resolvent of $\underline{\mathbf{E}}$. The difficult part of (4.3.8) is to find an (asymptotic) expression for the scalar product $\langle\mathbf{v}_1, \mathbf{G}_{\underline{\mathbf{E}}}\mathbf{v}_1\rangle$. Let us assume without loss of generality[2] that $\mathbf{C}$ is Gaussian, which allows us to set $\mathbf{v}_1 = (1, 0, \dots, 0)$. Then the equation we try to solve is:

$$\lambda_1\mathbf{G}_{\underline{\mathbf{E}}}(\lambda_1)_{11} = d_1^{-1} + 1. \tag{4.3.9}$$

As we shall see in the next section, the entries of $\mathbf{G}_{\underline{\mathbf{E}}}$ actually converges to a deterministic quantity for $N \to \infty$ and one obtains using Eq. (5.1.2) (see (B.4.19) for an alternative derivation). The result reads

$$\mathbf{G}_{\underline{\mathbf{E}}}(z)_{11} \approx \frac{1}{z - \underline{\mu}_1(1 - q + qz\mathfrak{g}_{\underline{\mathbf{E}}}(z))} = \frac{1}{z(1 - \mu_{r+1}\mathfrak{g}_{\underline{\mathbf{S}}}(z))},$$

[2]The extension to non-Gaussian entries can be done using standard comparison techniques, see e.g. [109] for details.

where we used the identity (4.2.25) and that $\underline{\mu}_1 \equiv \mu_{r+1}$ by construction of (4.3.6) in the last step. If $\lambda_1$ is not an eigenvalue of $\underline{\mathbf{E}}$, we find that Eq. (4.3.9) becomes in the LDL

$$\frac{1}{1-\mu_{r+1}\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1)} = d_1^{-1} + 1, \tag{4.3.10}$$

which is equivalent to:

$$\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1) = \frac{1}{\mu_{r+1}(1+d_1)} \equiv \frac{1}{\mu_1}, \tag{4.3.11}$$

where we used (4.3.4) in the last step. Hence, we see that $\lambda_1$ is an outlier if it satisfies for large $N$:

$$\lambda_1 = \theta(\mu_1) := \mathcal{B}_{\underline{\mathbf{S}}}\left(\frac{1}{\mu_1}\right), \tag{4.3.12}$$

This result is very general and can be extended for any outlier $\lambda_i$ with $i \in [\![1, r]\!]$. Moreover, we see that for $N \to \infty$, the (random) outlier $\lambda_1$ converges to a deterministic function of $\mu_1$. Hence, the function (4.3.12) depicts the "classical location" at which an outlier sticks and can therefore be interpreted as the analog of (4.2.32) for outliers. Note however that (4.3.12) requires the knowledge of the spikeless matrix $\underline{\mathbf{S}}$ (or $\underline{\mathbf{E}}$), which requires in practice to make some assumptions on which empirical eigenvalue should be considered as a spike.

The result (4.3.12) generalizes the result of Baik-Ben Arous-Péché for the spiked covariance matrix model [15]. Indeed, let us assume that the eigenvalues of the true covariance matrix $\mathbf{C}$ is composed of one outlier and $N-1$ eigenvalues at unity. Then, one trivially deduces that $\underline{\mu}_i = 1$ for all $i = 1, \dots, N$ which implies that the spectrum of $\underline{\mathbf{E}}$ is governed by the Marčenko-Pastur law (3.1.42). In fact, in the limit $N \to \infty$, the spectrum of $\underline{\mathbf{E}}$ and $\mathbf{E}$ are equivalent since the perturbation is of finite rank. Therefore, we can readily compute the Blue transform of the dual matrix $\underline{\mathbf{S}}$ from (4.2.27) to find

$$\mathcal{B}_{\underline{\mathbf{S}}}(x) = \frac{1}{x} + \frac{q}{1-x}, \tag{4.3.13}$$

and applying this formula into Equation (4.3.12) leads to the so-called BBP phase transition

$$\begin{cases} \lambda_1 = \mu_1 + q\frac{\mu_1}{\mu_1 - 1} & \text{if } \mu_1 > 1 + \sqrt{q}; \\ \lambda_1 = \lambda_+ = (1+\sqrt{q})^2 & \text{if } \mu_1 \leqslant 1 + \sqrt{q}, \end{cases} \tag{4.3.14}$$

where $\mu_1 = \mu_0(1+d_1)$ is the largest eigenvalue of $\mathbf{C}$, which is assumed to be a spike. Note that in the limit $\mu_1 \to \infty$, we get $\lambda_1 \approx \mu_1 + q + \mathcal{O}(\mu_1^{-1})$. For rank $r$ perturbation, all eigenvalues such that $\mu_k > 1 + \sqrt{q}$, $1 \leqslant k \leqslant r$ will end up isolated above the Marčenko-Pastur sea, all others disappear below $\lambda_+$. All these isolated eigenvalues have Gaussian fluctuations of order $T^{-1/2}$ [15]. The typical fluctuation of order $T^{-1/2}$ is also true for an arbitrary $\mathbf{C}$ [43], and is much smaller than the uncertainty in the bulk of the distribution, of order $\sqrt{q}$. Note that a naive application of Eq. (4.2.1) to outliers would lead to a "mini-Wishart" distribution around the top eigenvalue, which incorrect (the distribution is Gaussian) except if the top eigenvalue has a degeneracy proportional to $N$.

# Chapter 5

# Statistics of the eigenvectors

We saw in the previous chapter that tools from RMT allows one to infer many properties of its (asymptotic) spectrum, be it for the bulk or more localized regions of the spectrum (edges and outliers). These results allow us to characterize with great detail the statistics of the eigenvalues of large sample covariance matrices. In particular, it is clear that in the high-dimensional limit, the use of sample covariance matrices is certainly not recommended as each sample eigenvalue $[\lambda_i]_i$ converges to a non-deterministic value, but this value is different from the corresponding "true" population eigenvalue $[\mu_i]_i$. Note that the results presented above only cover a small part of the extremely vast literature on this topic, including the study microscopic/local statistics (down to the $N^{-1}$ scale) [84, 109, 149, 191].

On the other hand, results concerning the eigenvectors are comparatively quite scarce. One reason is that most studies in RMT focus on rotationally invariant ensembles, such that the statistics of eigenvectors is featureless by definition. Notwithstanding, this question turns out to be very important for sample covariance matrices since in this case the direction of the eigenvectors of the "population" matrix must somehow leave a trace. There are, at least, two natural questions about the eigenvectors of the sample matrix $\mathbf{E}$:

(i) How similar are sample eigenvectors $[\mathbf{u}_i]_{i\in[\![N]\!]}$ and the true ones $[\mathbf{v}_i]_{i\in[\![N]\!]}$?

(ii) What information can we learn about the population covariance matrix by observing two independent realizations – say $\mathbf{E} = \sqrt{\mathbf{C}}\boldsymbol{\mathcal{W}}\sqrt{\mathbf{C}}$ and $\mathbf{E}' = \sqrt{\mathbf{C}}\boldsymbol{\mathcal{W}}'\sqrt{\mathbf{C}}$ – that remain correlated through $\mathbf{C}$?

The aim of this chapter is to present some of the most recent results about the eigenvectors of large sample covariance matrices that will allow us to answer these two questions. More precisely, we will show how the tools developed in Section 3 can help us extract the statistical features of the eigenvectors $[\mathbf{u}_i]_{i\in[\![1,N]\!]}$. Note that we will discuss these issues for a multiplicative noise model (see (3.1.80) above), but the same questions can be investigated for additive noise as well, see [3, 5, 23, 41, 109] and Appendix 11.

A natural quantity to characterize the similarity between two arbitrary vectors – say $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ – is to consider the scalar product of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$. More formally, we define the "overlap" $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle$. Since the eigenvectors of real symmetric matrices are only defined up to a sign, we shall in fact consider the squared overlaps $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle^2$. In the first problem alluded to above, we want to understand the relation between the eigenvectors of the population matrix $[\mathbf{v}_i]_i$ and those of the sample matrix $[\mathbf{u}_i]_i$. The matrix of squared overlaps is defined as $\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2$, which actually forms

a so-called bistochastic matrix (positive elements with the sums over both rows and columns all equal to unity).

In order to study these overlaps, the central tool of this chapter is again the resolvent (and not its normalized trace as in the previous section). Indeed, that if we choose the $\mathbf{v}$'s to be our reference basis, we find from (3.1.6):

$$\langle \mathbf{v}\,, \mathbf{G}_{\mathbf{E}}(z)\mathbf{v}\rangle = \sum_{i=1}^{N} \frac{\langle \mathbf{v}\,, \mathbf{u}_i\rangle^2}{z-\lambda_i}, \tag{5.0.1}$$

for $\mathbf{v}$ a deterministic vector in $\mathbb{R}^N$ of unit norm. Note that we can extend the formalism to more general entries of $\mathbf{G}_{\mathbf{E}}(z)$ of the form:

$$\langle \mathbf{v}\,, \mathbf{G}_{\mathbf{E}}(z)\mathbf{v}'\rangle = \sum_{i=1}^{N} \frac{\langle \mathbf{v}\,, \mathbf{u}_i\rangle\langle \mathbf{u}_i\,, \mathbf{v}'\rangle}{z-\lambda_i}, \tag{5.0.2}$$

for $\mathbf{v}$ and $\mathbf{v}'$ two unit norm deterministic vectors in $\mathbb{R}^N$.

We see from Eqs. (5.0.1) and (5.0.2) that each pole of the resolvent defines a projection onto the corresponding sample eigenvectors. This suggests that the techniques we need to apply are very similar to the ones used above to study of the density of states. However, one should immediately stress that contrarily to eigenvalues, each eigenvector $\mathbf{u}_i$ for any given $i$ continues to fluctuate when $N \to \infty$,[1] and never reaches a deterministic limit. As a consequence, we will need to introduce some averaging procedure to obtain a well defined result. We will thus consider the following quantity[2]

$$\Phi(\lambda_i, \mu_j) := N\mathbb{E}[\langle \mathbf{u}_i, \mathbf{v}_j\rangle^2], \tag{5.0.3}$$

where the expectation $\mathbb{E}$ can be interpreted either as an average over different realizations of the randomness or, perhaps more meaningfully for applications, as an average *for a fixed sample* over small intervals of eigenvalues of width $\mathrm{d}\lambda = \eta$ that we choose in the range $1 \gg \eta \gg N^{-1}$ (say $\eta = N^{-1/2}$) such that there are many eigenvalues in the interval $\mathrm{d}\lambda$, while keeping $\mathrm{d}\lambda$ sufficiently small for the spectral density to be constant. Interestingly, the two procedures lead to the same result for large matrices, i.e. the locally "smoothed" quantity $\Phi(\lambda, \mu)$ is self averaging. Note also the factor $N$ in the definition above, indicating that we expect typical square overlaps to be of order $1/N$, see below.

For the second question, the main quantity of interest is, similarly, the (mean squared) overlap between two independent noisy eigenvectors

$$\Phi(\lambda_i, \tilde{\lambda}_j) \;:=\; N\mathbb{E}[\langle \mathbf{u}_i, \tilde{\mathbf{u}}_j\rangle^2], \tag{5.0.4}$$

where $[\tilde{\lambda}_i]_i$ and $[\tilde{\mathbf{u}}_i]_i$ are the eigenvalues and eigenvectors of $\tilde{\mathbf{E}}$, i.e. another sample matrix that is independent from $\mathbf{E}$ (but with the same underlying population matrix $\mathbf{C}$).

[1] Recall that we have indexed the eigenvectors by their associated eigenvalue.

[2] We emphasize that we consider the population eigenvectors to be deterministic throughout this section. Only the sample eigenvectors are random.

## 5.1 Asymptotic eigenvectors deformation in the presence of noise

We consider in this section the first question, that is: can we characterize the effect of the noise on the eigenvectors? Differently said, how do the sample eigenvectors deviate from the population ones? In order to answer to this question, Eq. (5.0.3) seems to be a good starting point since it allows one to extract exactly the projection of the sample eigenvectors onto the population ones. We shall show now that Eq. (5.0.3) converges to a deterministic quantity in the large $N$ limit; more precisely, we can summarize the main results of this section as follows:

(i) Any bulk sample eigenvectors is *delocalized* in the population basis, i.e. $\Phi(\lambda_i, \mu_j) \sim \mathcal{O}(1)$ for any $i \in [\![r+1, N]\!]$ and $j \in [\![N]\!]$ (and not $\mathcal{O}(N)$);

(ii) For any outlier (i.e. $i \leqslant r$), $\mathbf{u}_i$ is concentrated within a cone with its axis parallel to $\mathbf{v}_i$ but is completely delocalized in any direction orthogonal to the spike direction $\mathbf{v}_i$.

Therefore, these results look quite disappointing for a inference standpoint. Indeed, for the bulk eigenvectors, we discover that projection the estimated eigenvectors and their corresponding "true" directions converges almost surely to zero for large $N$; i.e. sample eigenvectors appear to contain very little information about the the true eigenvectors (on this point, see however [136]). Still, as we will see below, the square-overlaps are not all equal to $1/N$ but some interesting modulations appear, that we compute below by extending the Marčenko-Pastur equation to the full resolvent. For the outliers, on the other hand, the global picture is quite different. In particular, the phase transition phenomenon alluded in section 4 above also holds for the projection of the sample spike eigenvector onto its parent population spike: as soon as an eigenvalue pops out from the bulk, the square overlap becomes of order 1, as noticed in e.g. [23, 92, 144]. In fact, the angle between the sample spike eigenvectors with the parent spike can be computed exactly, see below.

**5.1.1. The bulk.** Let us focus on the bulk eigenvectors first, i.e. eigenvectors associated to eigenvalues lying in the bulk of spectral density when the dimension of the empirical correlation matrix grows to infinity. This question has been investigated very recently in [40, 113] and we repeat the different arguments here. The first step is to characterize the asymptotic behaviour of the resolvent of sample covariance matrices. This can be done by specializing Eq. (3.1.92) for the resolvent of the product of free matrices to the case where $\mathbf{A} = \mathbf{C}$ and $\mathbf{B} = \mathbf{X}\mathbf{X}^*$. In words, $\mathbf{A}$ is the population matrix while $\mathbf{B}$ is a white Wishart matrix that plays the role of the noisy multiplicative perturbations. Using (3.1.44), we know the $\mathcal{S}$-transform of white Wishart matrices explicitly so that one finds from Eq. (3.1.44), for $N \to \infty$:

$$z\mathbf{G}_{\mathbf{E}}(z) = Z(z)\mathbf{G}_{\mathbf{C}}(Z(z)), \quad \text{with} \quad Z = \frac{z}{1 - q + qz\mathfrak{g}_{\mathbf{E}}(z)}. \tag{5.1.1}$$

In the literature, such a limiting result is referred to as a "deterministic equivalent", as the RHS depends only to deterministic quantities[3], and this is another evidence of the self-averaging property for large random matrices.

One should notices that (5.1.1) is a relation between resolvent matrices that generalizes the scalar Marčenko-Pastur equation (4.2.1) (which can be recovered by taking the trace on both sides of the equation). This relation first appeared in [44], obtained using a planar diagram expansion valid for Gaussian entries. A few years later, that result was proven rigorously in

[3] Recall that $\mathfrak{g}_{\mathbf{E}}(z)$ is the *limiting* Stieltjes transform.

Ref. [109] in a much more general framework, highlighting again the *universal* nature of the resolvent of random matrices, down to the local scale.[4] Choosing to work in the eigenbasis of $\mathbf{C}$ is diagonal, Eq. (5.1.1) reduces to:

$$\mathbf{G}_{\mathbf{E}}(z)_{ij} = \frac{\delta_{ij}}{z - \mu_i(1 - q + qz\mathfrak{g}_{\mathbf{E}}(z))}, \tag{5.1.2}$$

and it was shown by Knowles and Yin that this deterministic equivalent holds with fluctuations or order $(\eta N)^{-1/2}$ [109] when $\mathrm{Re}[z]$ lies in the spectrum of $\mathbf{E}$. More interestingly, an explicit upper bound for the error term is actually provided in [109]. In particular, the authors showed that Eq. (5.1.1) holds at a local scale $\eta = \widehat{\eta}N^{-1}$ with $\widehat{\eta} \gg 1$, with an error term bounded from above by:

$$\Psi(z) \;:=\; \sqrt{q\frac{\mathrm{Im}\,\mathfrak{g}_{\mathbf{S}}(z)}{\widehat{\eta}}} + \frac{q}{\widehat{\eta}}, \tag{5.1.3}$$

provided that $N$ is large enough. We give an illustration of this ergodic behavior in Figure 5.1.1, and we see the agreement is excellent. Note that when $\mathrm{Re}[z] \notin \mathrm{supp}[\rho_{\mathbf{E}}]$, the error term is bounded from above by $T^{-1/2+\varepsilon}$ with high probability and for $\varepsilon > 0$ a small constant [109].

How can we compute the mean squared overlap using (5.1.1)? The idea is to derive an inversion formula similar to (3.1.11) for the full resolvent. More specifically, we start from (3.1.6) for a given $\mathbf{v} = \mathbf{v}_j$ and notice that the true eigenvectors are deterministic. Therefore, the sum on the RHS of the latter equation is expected to converge in the large $N$ limit provided $z$ is outside of the support of the spectrum of $\mathbf{E}$. Moreover, the eigenvalues in the bulk converge to their classical position (4.2.32) so that we obtain for $N \to \infty$ that

$$\langle \mathbf{v}_j\,, \mathbf{G}_{\mathbf{E}}(z)\mathbf{v}_j\rangle \underset{N\to\infty}{\sim} \int \frac{\Phi(\lambda,\mu_j)\rho_{\mathbf{E}}(\lambda)}{\lambda_i - \lambda - \mathrm{i}\eta}\mathrm{d}\lambda. \tag{5.1.4}$$

where we have set $z = \lambda_i - \mathrm{i}\eta$, $\eta \gg N^{-1}$ and $\Phi(\lambda,\mu_j)$ is the smoothed squared overlap, averaged over a small interval of width $\eta$ around $\lambda$. Therefore, the final inversion formula is obtained using the Sokhotski-Plemelj identity as:

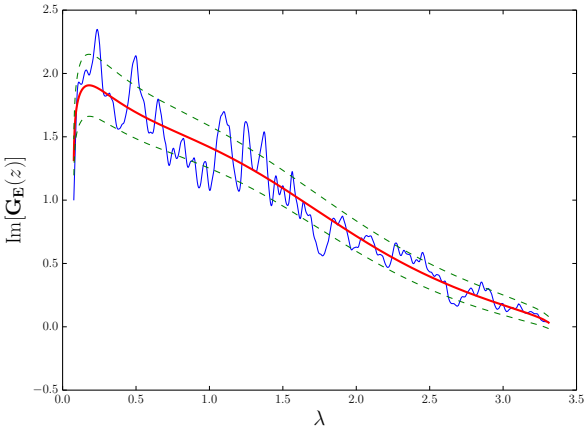
$$\Phi(\lambda_i,\mu_j) = \frac{1}{\pi\rho_{\mathbf{E}}(\lambda_i)}\lim_{\eta\to 0}\mathrm{Im}\langle \mathbf{v}_j\,, \mathbf{G}_{\mathbf{E}}(\lambda_i - \mathrm{i}\eta)\mathbf{v}_j\rangle, \tag{5.1.5}$$

(note the assumption that $\lambda_i$ lies in the bulk of the spectrum is crucial here). This last identity thus allows us to compute the squared overlap $\Phi(\lambda_i,\mu_j)$ from the full resolvent $\mathbf{G}_{\mathbf{E}}$, for any in the bulk ($i \geqslant r+1$) and a fixed $j \in [\![1,N]\!]$. Specializing to the explicit form of $\mathbf{G}_{\mathbf{E}}(z)$ given in Eq. (5.1.2), we finally obtain a beautiful explicit result for the (rescaled) average squared overlap:

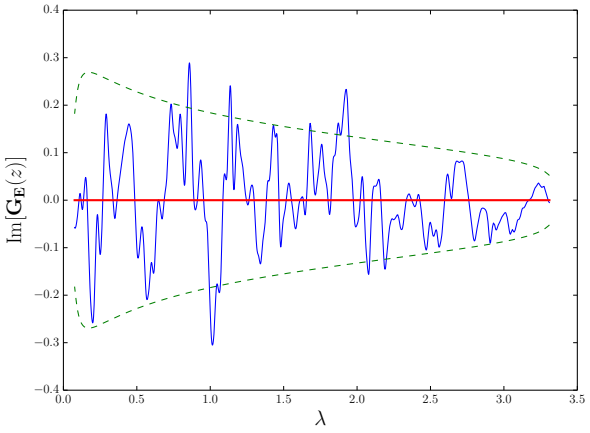
$$\Phi(\lambda_i,\mu_j) = \frac{q\mu_j\lambda_i}{(\mu_j(1-q) - \lambda_i + q\mu_j\lambda_i\mathfrak{h}_{\mathbf{E}}(\lambda_i))^2 + q^2\mu_j^2\lambda_i^2\pi^2\rho_{\mathbf{E}}^2(\lambda_i)}, \tag{5.1.6}$$

with $i \in [\![r+1,N]\!], j \in [\![1,N]\!]$ and $\mathfrak{h}_{\mathbf{E}}(\lambda_i)$ denotes the real part of the Stieltjes transform $\mathfrak{g}_{\mathbf{E}}$ (see Eq. (3.1.9)). This relation is exact in the limit $N \to \infty$ and was first derived by Ledoit and Péché in [113]. We emphasize again that this expression remains correct even if $\mu_j$ is an outlier. Since $\Phi(\lambda_i,\mu_j)$ is of order unity whenever $q > 0$, we conclude that the dot product between any bulk eigenvector $\mathbf{u}_i$ of $\mathbf{E}$ and the eigenvectors $\mathbf{v}_j$ of $\mathbf{C}$ is of order $N^{-1/2}$, i.e vanishes

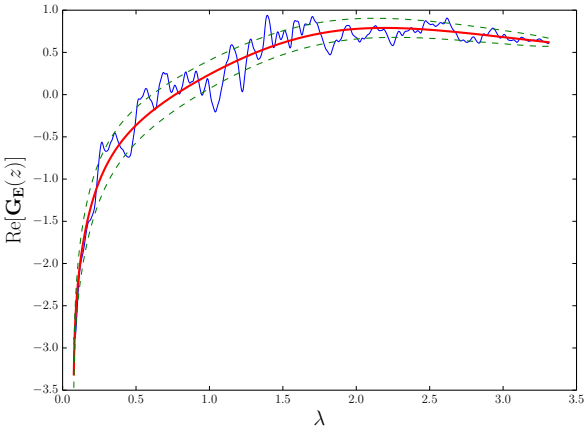
[4] Note that the Gaussian assumption is not needed either within the Replica method presented in Section 3.



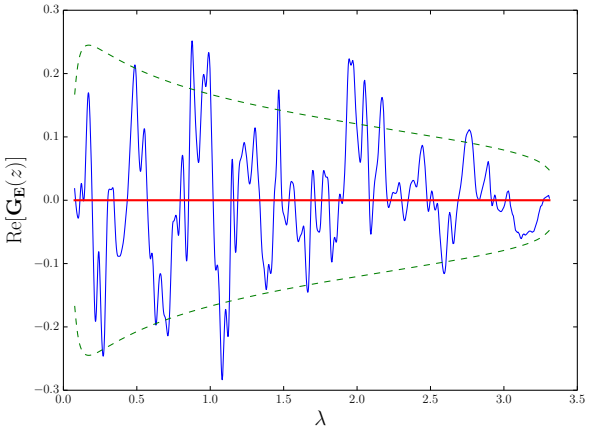

(A) Diagonal entry of $\mathrm{Im}[\mathbf{G}_{\mathbf{E}}(z)]$ with $i = j = 1000$.




(B) Off diagonal entry of $\mathrm{Im}[\mathbf{G}_{\mathbf{E}}(z)]$ with $i = 999$ and $j = 1001$.




(C) Diagonal entry of $\mathrm{Re}[\mathbf{G}_{\mathbf{E}}(z)]$ with $i = j = 1000$.




(D) Off diagonal entry of $\mathrm{Re}[\mathbf{G}_{\mathbf{E}}(z)]$ with $i = 999$ and $j = 1001$.

FIGURE 5.1.1. Illustration of Eq. (5.1.2). The population matrix is an Inverse Wishart matrix with parameter $\kappa = 5$ and the sample covariance matrix is generated using a Wishart distribution with $T = 2N$ and $N = 2000$. The empirical estimate of $\mathbf{G}_{\mathbf{E}}(z)$ (blue line) is computed for any $z = \lambda_i - \mathrm{i}N^{-1/2}$ with $i \in [\![1, N]\!]$ comes from one sample and the theoretical one (red line) is given by the RHS of Eq. (5.1.1). The green dotted corresponds to the confidence interval whose formula is given by Eq. (5.1.3).

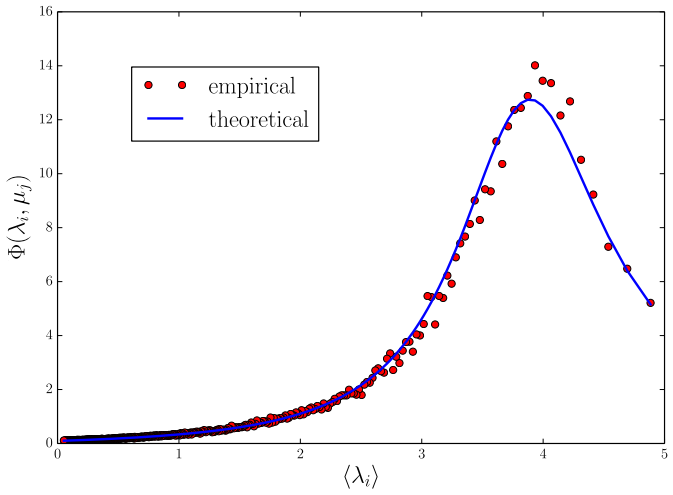



FIGURE 5.1.2. Rescaled mean squared overlaps $\Phi(\lambda_i, \mu_j)$ as a function of $\lambda_i$. We choose $\mathbf{C}$ as an inverse-Wishart matrix with parameter $\kappa = 1.0$ and set $N = 500$, $q = 0.5$. The empirical average (blue points) comes from 500 independent realizations of $\mathbf{E}$. The theoretical prediction (red line) is given by Eq. (5.1.6). The peak of the mean squared overlap is in the vicinity of $\lambda_i \approx \mu_j$ for any $i$.

at large $N$, and therefore non-outlier sample eigenvectors retain very little information about their corresponding true eigenvectors. This implies that any bulk eigenvector is a extremely poor estimator of the true one in the high-dimensional regime. We provide in Figure 5.1.2 an illustration of Eq. (5.1.6) for $N = 500$ and $\mathbf{C}$ an Inverse Wishart matrix with $\kappa = 1$. The empirical average comes from 500 independent realization of $\mathbf{E}$ and we see that it agrees perfectly with the asymptotic theoretical prediction, Eq. (5.1.6). Note that in the limit $q \to 0$, $\Phi(\lambda_i, \mu_j)$ becomes more and more peaked around $\lambda_i \approx \mu_j$, with an amplitude that diverges for $q = 0$. Indeed, in this limiting case, one should find that $\mathbf{u}_i \to \pm \mathbf{v}_j \delta_{ij}$, i.e. the sample eigenvectors become equal to the population ones.

**5.1.2. Outliers.** *This section is based on a work in progress with Antti Knowles [109].*

By construction, the spiked correlation model of Section 4.1.3 is such that the top $r$ eigenvalues $[\lambda_i]_{i \in [\![1,r]\!]}$ lie outside the spectrum of $\rho_{\mathbf{E}}$. What can be said about the statistics of the associated spike eigenvectors $[\mathbf{u}_i]_{i \in [\![1,r]\!]}$? If we think of these outliers as a finite-rank deformation of a (fictitious) spikeless matrix $\underline{\mathbf{E}}$, then by Weyl's eigenvalue interlacing inequalities [187], the asymptotic density $\rho_{\mathbf{E}}$ is not influenced by the presence of non-macroscopic spikes, by which we mean that $\rho_{\mathbf{E}}(\lambda_i) = 0$ for any outlier eigenvalues. We saw in the previous section that for non-outlier eigenvectors, the main ingredients to compute the overlap are (i) the self-averaging property and (ii) the inversion formula (5.1.5). Both implicitly rely on the continuous limit being valid, which is however not the case for outliers. Hence, we expect the statistics of outlier eigenvectors to be quite different from the bulk eigenvectors as confirmed for the null hypothesis case $\underline{\mathbf{C}} = \mathbf{I}_N$ [92, 145]. In this section, we present the analytical tools to analyze these overlaps for outliers in the case of an arbitrary population covariance, following the lines of [43].

From Eq. (4.3.12) we saw that each outlier eigenvalues $[\lambda_i]_{i \in [\![1,r]\!]}$ of $\mathbf{E}$ converges to a deter-

ministic limit $\theta(\mu_i)$, where $\mu_i$ is the corresponding population spike and $\theta$ is a certain function related to the Marčenko-Pastur equation. Consequently, for isolated spikes $i \in [\![1, r]\!]$ we can define the closed disc $D_i$ in the complex plane, centered at $\theta(\mu_i)$ with radius chosen such that each it encloses no other point in the set $[\theta(\mu_j)]_{j\in[\![1,r]\!]}$ (see [43] for details). Then, defining $\Gamma_i$ to be the boundary of the closed disc $D_i$, we can obtain the squared overlap for outlier eigenvectors using Cauchy's integral formula

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = \frac{1}{2\pi \mathrm{i}} \oint_{\Gamma_i} \langle \mathbf{v}_j, \mathbf{G}_{\mathbf{E}}(z)\mathbf{v}_j \rangle \mathrm{d}z, \tag{5.1.7}$$

for $i, j \in [\![1, r]\!]$. We emphasize there is no expectation value in Eq. (5.1.7) (compare to our definition of the overlap in Eq. (5.0.3)). The evaluation of the integral is highly non-trivial since $\mathbf{G}_{\mathbf{E}}$ is singular in the vicinity of $\theta(\mu_j)$ for any $j \in [\![1, r]\!]$ and finite $N$. To bypass this problem, we reconsider the spikeless population covariance matrix $\underline{\mathbf{C}}$ defined in (4.3.6) and the corresponding spikeless sample covariance matrix by $\underline{\mathbf{E}}$. Clearly, the resolvent $\mathbf{G}_{\underline{\mathbf{E}}}$ is no longer singular in the vicinity of $\theta(\mu_j)$, by construction. Moreover, as we said above, the global statistics of the eigenvalues of $\underline{\mathbf{E}}$ and $\mathbf{E}$ are identical in the limit $N \to \infty$. Lastly, we can relate any projection of $\mathbf{G}_{\mathbf{E}}$ onto the outlier population covariance eigenbasis using Schur complement formula (see B for a reminder):

$$\mathbf{V}^{(r)*}\mathbf{G}_{\mathbf{E}}(z)\mathbf{V}^{(r)} = -\frac{1}{z}\left[\mathbf{D}^{-1} - \frac{\sqrt{\mathbf{I}_N + \mathbf{D}}}{\mathbf{D}}\left(\mathbf{D}^{-1} + \mathbf{I}_N - z\mathbf{V}^{(r)*}\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V}^{(r)}\right)^{-1}\frac{\sqrt{\mathbf{I}_N + \mathbf{D}}}{\mathbf{D}}\right]. \tag{5.1.8}$$

This identity has been used in several studies that deal with related problems [27, 43] and references therein. Its derivation only needs linear algebra arguments and can be found at the end of this section. With this identity, the statistics of the outliers of $\mathbf{E}$ is seen to only rely on the spikeless matrix $\underline{\mathbf{E}}$. In particular, the integrand of (5.1.7) can be rewritten using the spikeless resolvent which is analytic everywhere outside the spectrum of $\underline{\mathbf{E}}$. Since the global law of resolvent of $\underline{\mathbf{E}}$ is the same than $\mathbf{E}$ in the large $N$ limit, we can again use the estimate (5.1.1). By plugging (5.1.1) into (5.1.8), one obtains

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = -\frac{1}{2\pi \mathrm{i}} \oint_{\theta(\Gamma_i)} \frac{1}{z}\left[\frac{1}{d_j} - \frac{1+d_j}{d_j^2}\frac{1}{d_j^{-1} + 1 - z\langle \mathbf{v}_j, \mathbf{G}_{\mathbf{E}_0}(z)\mathbf{v}_j \rangle}\right] dz. \tag{5.1.9}$$

Then, using Eq. (4.3.8) and Cauchy's theorem, one eventually finds [43]

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = \delta_{ij}\mu_i \frac{\theta'(\mu_i)}{\theta(\mu_i)} + \mathcal{O}(N^{-1/2}) = \delta_{ij}\mu_i \frac{\theta'(\mu_i)}{\lambda_i} + \mathcal{O}(N^{-1/2}), \tag{5.1.10}$$

for any $i, j \in [\![1, r]\!]$ and where we used (4.3.12) in the denominator in the last step. Therefore, we conclude that the sample outlier eigenvector $\mathbf{u}_i$ is concentrated on a cone around $\mathbf{v}_i$ with aperture $2\arccos(\mu_i\theta'(\mu_i)/\theta(\mu_i))$. We also deduce from Eq. (5.1.10) that $\mathbf{u}_i$ is delocalized in all directions $\mathbf{v}_j$ associated to different spikes $\mu_j \neq \mu_i$.

An interesting application of (5.1.10) is to reconsider the spiked covariance matrix model introduce in the previous chapter. Let us assume for simplicity a single spike ($r = 1$) and from equation (4.3.13), one gets, for $\mu_1 > 1 + \sqrt{q}$

$$\theta(\mu_1) = \mu_1 + q + \frac{q}{\mu_1 - 1},$$

and plugging this result into equation (5.1.10) yields

$$\langle \mathbf{u}_1, \mathbf{v}_1 \rangle^2 = \frac{\mu_1}{\theta(\mu_1)} \left( 1 - \frac{q}{(\mu_1 - 1)^2} \right) + \mathcal{O}(T^{-1/2}), \tag{5.1.11}$$

which is the expected result [23, 25, 27, 136, 144]. This result shows that the coherence between the population spike and its sample counterpart becomes progressively lost when $\mu_1 \to 1 + \sqrt{q}$ as it should be from the result (4.3.14).

The same analysis can be applied for the overlap between the sample spikes and the population bulk eigenvalues $j > r$. More precisely, using the explicit error bound of $T^{-1/2+\varepsilon}$ when $\mathrm{Re}[z] \notin \mathrm{supp}[\rho_{\mathbf{E}}]$ (see [109, Theorem 3.7] for details), we may compute the overlaps from the identity (5.1.9) and a resolvent expansion of $\mathbf{G}_{\mathbf{E}}$ around its deterministic equivalent, given in (5.1.2). Then, one can show that the first term that contributes scales at $\mathcal{O}(T^{-1})$ and the final reads[5]:

$$\Phi(\lambda_i, \mu_j) = q \frac{\mu_j}{\lambda_i (1 - \mu_j/\mu_i)^2} + \mathcal{O}(T^{-1/2}), \qquad i \in [\![1, r]\!],\ j \in [\![r+1, N]\!]. \tag{5.1.12}$$

The rigorous derivation of this result is given in the working paper [43]. As expected, any outlier eigenvector $\mathbf{u}_i$ has only $\sim N^{-1/2}$ overlap with any eigenvector of $\mathbf{C}$ except its "parent" from $\mathbf{v}_i$. We illustrate Eq. (5.1.12) in Figure 5.1.3 as a function of the population eigenvalues $[\mu_i]_i$ with $i > 2$ as $i = 1$ corresponds to the spike, whose overlap is given by Eq. (5.1.10). In our example $\mathbf{C}$ is an Inverse Wishart matrix with parameter $\kappa = 1$ and we add a rank one perturbation such that $\lambda_1 \approx 10$. The empirical average comes from 200 realizations of $\mathbf{E}$ and we see that the agreement with the theoretical prediction in excellent.

### 5.1.3. Derivation of the identity (5.1.8).

The derivation of the identity (5.1.8) is the central tool in order to deal with the outliers of the sample covariance matrix $\mathbf{E}$. It relies purely on linear algebra arguments (see Appendix (B) for a reminder). In order to lighten the notations, let us rename $\mathbf{V} \equiv \mathbf{V}^{(r)}$ in this section. The first step is to write the following identity from Eq. (4.3.6):

$$\begin{aligned} \sqrt{\underline{\mathbf{C}}}\, \mathbf{C}^{-1} \sqrt{\underline{\mathbf{C}}} - \mathbf{I}_N &= (\mathbf{I}_N + \mathbf{V}\mathbf{D}\mathbf{V}^*)^{-1} - \mathbf{I}_N \\ &= -(\mathbf{I}_N + \mathbf{V}\mathbf{D}\mathbf{V}^*)^{-1} \mathbf{V}\mathbf{D}\mathbf{V}^* \\ &= -\mathbf{V}\mathbf{D}(\mathbf{I}_r + \mathbf{D})^{-1} \mathbf{V}^* \end{aligned} \tag{5.1.13}$$

where we used the resolvent identity (5.2.9) in the second line. This allows us to get (omitting the argument $z$)

$$\begin{aligned} \underline{\mathbf{C}}^{-1/2} \mathbf{C}^{1/2} \mathbf{G}_{\mathbf{E}} \mathbf{C}^{1/2} \underline{\mathbf{C}}^{-1/2} &= \underline{\mathbf{C}}^{-1/2} \big( z\mathbf{C}^{-1} - \mathbf{X}\mathbf{X}^* \big)^{-1} \underline{\mathbf{C}}^{-1/2} \\ &= \big( z(\underline{\mathbf{C}}^{1/2} \mathbf{C}^{-1} \underline{\mathbf{C}}^{1/2} - \mathbf{I}_N) + z\mathbf{I}_N - \underline{\mathbf{E}} \big)^{-1} \\ &= \big( -z\mathbf{V}\mathbf{D}(I + \mathbf{D})^{-1} \mathbf{V}^* + \mathbf{G}_{\underline{\mathbf{E}}}^{-1} \big)^{-1}, \end{aligned} \tag{5.1.14}$$

where we invoked the previous identity Eq. (5.1.13) in the last step. From (B.2.1), we have with $\mathbf{A} \equiv z\mathbf{I}_N - \underline{\mathbf{E}}$, $\mathbf{B} \equiv -z\mathbf{V}$, $\mathbf{D} \equiv \mathbf{D}(\mathbf{I}_r + \mathbf{D})^{-1}$ and $\mathbf{C} \equiv \mathbf{V}^*$:

$$\underline{\mathbf{C}}^{-1/2} \mathbf{C}^{1/2} \mathbf{G}_{\mathbf{E}} \mathbf{C}^{1/2} \underline{\mathbf{C}}^{-1/2} = \mathbf{G}_{\underline{\mathbf{E}}} + z\mathbf{G}_{\underline{\mathbf{E}}} \mathbf{V} \Big( \mathbf{D}^{-1} + \mathbf{I}_r - z\mathbf{V}^* \mathbf{G}_{\underline{\mathbf{E}}} \mathbf{V} \Big)^{-1} \mathbf{V}^* \mathbf{G}_{\underline{\mathbf{E}}}. \tag{5.1.15}$$

From there, one has

$$(\mathbf{I}_N + \mathbf{D})^{1/2} \mathbf{V}^* \mathbf{G}_{\mathbf{E}} \mathbf{V} (\mathbf{I}_N + \mathbf{D})^{1/2} = \mathbf{V}^* \mathbf{G}_{\underline{\mathbf{E}}} \mathbf{V} + z\mathbf{V}^* \mathbf{G}_{\underline{\mathbf{E}}} \mathbf{V} \Big( \mathbf{D}^{-1} + \mathbf{I}_r - \mathbf{V}^* \mathbf{G}_{\underline{\mathbf{E}}} \mathbf{V} \Big)^{-1} \mathbf{V}^* \mathbf{G}_{\underline{\mathbf{E}}} \mathbf{V}. \tag{5.1.16}$$

[5] Recall that $\Phi$ is the rescaled overlaps.

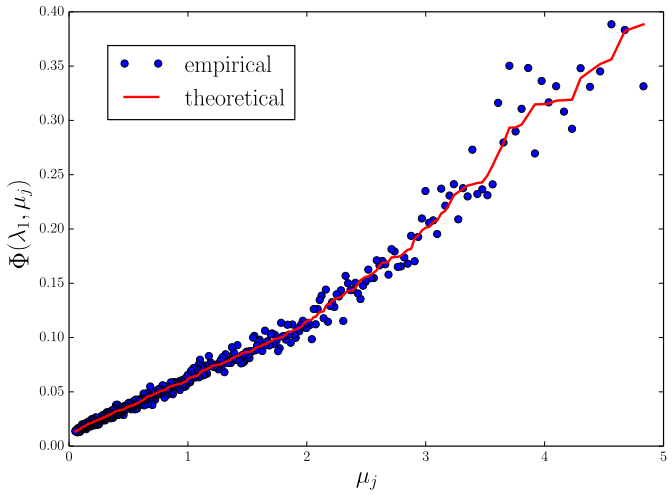



FIGURE 5.1.3. Rescaled mean squared overlap $\Phi^2(\lambda_1, \mu_j)$ as a function of $\mu_j$ for $j > 1$. We chose the spikeless population matrix $\underline{\mathbf{C}}$ to be an Inverse-Wishart matrix with parameter $\kappa = 1.0$ and $N = 500$. We add a rank one perturbation such that $\lambda_1 \approx 10$ is isolated from the others. The sample matrix $\mathbf{E}$ is given by a Wishart matrix with $q = 0.5$. We compare the empirical average (blue points) comes from 200 independent realizations of $\mathbf{E}$. The theoretical prediction (red line) is given by Eq. (5.1.12).

We then use the identity

$$\mathbf{A} - \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A} = \mathbf{B} - \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}, \tag{5.1.17}$$

with $\mathbf{A} = V^* \mathbf{G}_{\underline{\mathbf{E}}} V$ and $\mathbf{B} = -(\mathbf{D}^{-1} + \mathbf{I}_r)/z$ to obtain

$$(\mathbf{I}_r + \mathbf{D})^{1/2} \mathbf{V}^* \mathbf{G}_{\mathbf{E}} \mathbf{V} (\mathbf{I}_r + \mathbf{D})^{1/2} = -\frac{1}{z}\left[\frac{\mathbf{I}_r + \mathbf{D}}{\mathbf{D}} + \frac{\mathbf{I}_r + \mathbf{D}}{\mathbf{D}}\left(-(\mathbf{D}^{-1} + \mathbf{I}_r) + z\mathbf{V}^* \mathbf{G}_{\underline{\mathbf{E}}} \mathbf{V}\right)^{-1} \frac{\mathbf{I}_r + \mathbf{D}}{\mathbf{D}}\right]. \tag{5.1.18}$$

By rearranging the terms, we finally get

$$\mathbf{V}^* \mathbf{G}_{\mathbf{E}} \mathbf{V} = -\frac{1}{z}\left[\mathbf{D}^{-1} - \frac{\sqrt{\mathbf{I}_r + \mathbf{D}}}{\mathbf{D}}\left(\mathbf{D}^{-1} + \mathbf{I}_r - z\mathbf{V}^* \mathbf{G}_{\underline{\mathbf{E}}} \mathbf{V}\right)^{-1} \frac{\sqrt{\mathbf{I}_r + \mathbf{D}}}{\mathbf{D}}\right], \tag{5.1.19}$$

which is precisely Eq. (5.1.8).

## 5.2 Overlaps between the eigenvectors of correlated sample covariance matrices

We now consider the second problem of this chapter, that is to say how much information can we learn about the structure of $\mathbf{C}$ from the sample eigenvectors? Differently said, imagine one measures the sample covariance matrix of the same process but on two independent time intervals, how close are the corresponding eigenvectors expected to be? To answer this question, let us denote by $\mathbf{E}$ and $\tilde{\mathbf{E}}$ the independent sample estimates of the same population matrix $\mathbf{C}$ defined as

$$\mathbf{E} := \sqrt{\mathbf{C}}\boldsymbol{\mathcal{W}}\sqrt{\mathbf{C}}, \qquad \tilde{\mathbf{E}} := \sqrt{\mathbf{C}}\tilde{\boldsymbol{\mathcal{W}}}\sqrt{\mathbf{C}}, \tag{5.2.1}$$

where $\boldsymbol{\mathcal{W}}$ and $\tilde{\boldsymbol{\mathcal{W}}}$ are two independent white Wishart matrix with parameter $q$ and $q'$ respectively. As in Section 5.1, we can investigate this problem through the mean squared overlaps.

In this section, we provide exact, explicit formulas for these overlaps in the high dimensional regime, and perhaps surprisingly, we will see that they may be evaluated without any prior knowledge on the spectrum of $\mathbf{C}$. More specifically, we will show that Eq. (5.0.4) exhibits yet again a self-averaging behavior in the large $N$ limit, i.e. independent from the realization of $\mathbf{E}$ and $\tilde{\mathbf{E}}$. We will moreover see that the overlaps (5.0.4) significantly depart from the trivial null hypothesis as soon as the population $\mathbf{C}$ has a non-trival structure. Hence, this suggests that we might be able to infer the correlation structure of very large databases using empirical quantities only.

All these results have been obtained in the recent work [41] and we shall only give here the main steps. For the sake of clearness, we use the notations $\tilde{\lambda}_1 \geqslant \tilde{\lambda}_2 \geqslant \ldots \geqslant \tilde{\lambda}_N$ to denote the eigenvalues of $\tilde{\mathbf{E}}$ and by $\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \ldots, \tilde{\mathbf{u}}_N$ the associated eigenvectors. Note that we will again index the eigenvectors by their corresponding eigenvalues for convenience.

The central tool in this section is an inversion formula for (5.0.4) as it is usually done in RMT. To that end, we define the bivariate complex function

$$\psi(z,\tilde{z}) \;:=\; \left\langle \frac{1}{N}\,\mathrm{Tr}\left[(z-\mathbf{E})^{-1}(\tilde{z}-\tilde{\mathbf{E}})^{-1}\right]\right\rangle_{\mathcal{P}}, \tag{5.2.2}$$

where $z, \tilde{z} \in \mathbb{C}$ and $\langle\cdot\rangle_{\mathcal{P}}$ denotes the average with respect to probability measure associated to $\mathbf{E}$ and $\tilde{\mathbf{E}}$. Then, by a spectral decomposition of $\mathbf{E}$ and $\tilde{\mathbf{E}}$, one has

$$\psi(z,\tilde{z}) \;=\; \left\langle \frac{1}{N}\sum_{i,j=1}^{N} \frac{1}{z-\lambda_i}\frac{1}{\tilde{z}-\tilde{\lambda}_j}\langle \mathbf{u}_i\,,\tilde{\mathbf{u}}_j\rangle^2 \right\rangle_{\mathcal{P}}, \tag{5.2.3}$$

where $\mathcal{P}$ denotes the probability density function of the noise part of $\mathbf{E}$ and $\tilde{\mathbf{E}}$. For large random matrices, we expect the eigenvalues of $[\lambda_i]_{i\in[\![1,N]\!]}$ and $[\tilde{\lambda}_i]_{i\in[\![1,N]\!]}$ stick to their *classical* locations, i.e. smoothly allocated with respect to the quantile of the spectral density (see Section 4.2.1) so that the sample eigenvalues become deterministic in the large $N$ limit. Hence, we obtain after taking the continuous limit

$$\psi(z,\tilde{z}) \sim \int\int \frac{\rho(\lambda)}{z-\lambda}\frac{\tilde{\rho}(\tilde{\lambda})}{\tilde{z}-\tilde{\lambda}}\Phi(\lambda,\tilde{\lambda})\mathrm{d}\lambda\mathrm{d}\tilde{\lambda}, \tag{5.2.4}$$

where $\rho$ and $\tilde{\rho}$ are respectively the spectral density of $\mathbf{E}$ and $\tilde{\mathbf{E}}$, and $\Phi$ denotes the mean squared overlap defined in (5.0.4) above. Then, it suffices to compute

$$\psi(x-\mathrm{i}\eta, y\pm\mathrm{i}\eta) \;\sim\; \int\int \frac{(x-\lambda+\mathrm{i}\eta)}{(x-\lambda)^2+\eta^2}\frac{(y-\tilde{\lambda}\mp\mathrm{i}\eta)}{(y-\tilde{\lambda})^2+\eta^2}\rho(\lambda)\tilde{\rho}(\tilde{\lambda})\Phi(\lambda,\tilde{\lambda})\mathrm{d}\lambda\mathrm{d}\tilde{\lambda} \tag{5.2.5}$$

from which, one deduces that

$$\mathrm{Re}\big[\psi(x-\mathrm{i}\eta, y+\mathrm{i}\eta)-\psi(x-\mathrm{i}\eta, y-\mathrm{i}\eta)\big] \sim 2\int\int \frac{\eta\rho(\lambda)}{(x-\lambda)^2+\eta^2}\frac{\eta\tilde{\rho}(\tilde{\lambda})}{(y-\tilde{\lambda})^2+\eta^2}\Phi(\lambda,\tilde{\lambda})\mathrm{d}\lambda\mathrm{d}\tilde{\lambda}. \tag{5.2.6}$$

Finally, the inversion formula follows from Sokhotski-Plemelj identity

$$\lim_{\eta\to 0} \mathrm{Re}\big[\psi(x-\mathrm{i}\eta, y+\mathrm{i}\eta) - \psi(x-\mathrm{i}\eta, y-\mathrm{i}\eta)\big] \sim 2\pi^2 \rho(x)\tilde{\rho}(y)\Phi(x,y). \tag{5.2.7}$$

Note that the derivation holds for any models of $\mathbf{E}$ and $\tilde{\mathbf{E}}$ as long as its spectral density converges to a well-defined deterministic limit.

The inversion formula (5.2.7) allows us to study the mean squared overlap (5.0.4) through the asymptotic behavior of the bivariate function $\psi(z,\tilde{z})$. Moreover, since we are able control each entry of the resolvent of $\mathbf{E}$ and $\tilde{\mathbf{E}}$ (see Eq. (5.1.1)), the evaluation of Eq. (5.2.2) is immediate and leads to

$$\psi(z,\tilde{z}) \sim \frac{1}{z\tilde{z}}\frac{1}{N}\,\mathrm{Tr}\big[Z(z)(Z(z)-\mathbf{C})^{-1}\tilde{Z}(\tilde{z})(\tilde{Z}(\tilde{z})-\mathbf{C})^{-1}\big], \tag{5.2.8}$$

where $Z(z)$ is defined in (5.1.1) and $\tilde{Z}(z)$ is obtained from $Z$ by replacing $q$ and $\mathfrak{g}_{\mathbf{E}}$ by $\tilde{q}$ and $\mathfrak{g}_{\tilde{\mathbf{E}}}$. Then, we use the identity

$$\Big(Z(z)-\mathbf{C}\Big)^{-1}\Big(\tilde{Z}(\tilde{z})-\mathbf{C}\Big)^{-1} = \frac{1}{\tilde{Z}(\tilde{z})-Z(z)}\Big[\Big(Z(z)-\mathbf{C}\Big)^{-1} - \Big(\tilde{Z}(\tilde{z})-\mathbf{C}\Big)^{-1}\Big] \tag{5.2.9}$$

to obtain

$$\psi(z,\tilde{z}) \sim \frac{Z(z)\,\tilde{Z}(\tilde{z})}{z\tilde{z}}\frac{1}{\tilde{Z}(\tilde{z})-Z(z)}\frac{1}{N}\,\mathrm{Tr}\Big[\Big(Z(z)-\mathbf{C}\Big)^{-1} - \Big(\tilde{Z}(\tilde{z})-\mathbf{C}\Big)^{-1}\Big]. \tag{5.2.10}$$

From this last equation and using Marčenko-Pastur equation (4.2.1), we finally conclude that

$$\psi(z,\tilde{z}) \sim \frac{1}{\tilde{Z}(\tilde{z})-Z(z)}\left[\frac{\tilde{Z}(\tilde{z})}{\tilde{z}}\mathfrak{g}_{\mathbf{E}}(z) - \frac{Z(z)}{z}\mathfrak{g}_{\tilde{\mathbf{E}}}(\tilde{z})\right]. \tag{5.2.11}$$

One notices that Eq. (5.2.11) only depends on *a priori* observable quantities, i.e. they do not involve explicitly the unknown matrix $\mathbf{C}$. Once we characterized the asymptotic behavior of the bivariate function $\psi(z,\tilde{z})$, we can then apply the inversion formula Eq. (5.2.7) in order to retrieve the mean squared overlap (5.0.4). Before stating the main result of this section, we first rewrite (5.2.11) as a function of the Stieltjes transform $\mathfrak{g}_{\mathbf{S}}$ of the $T\times T$ dual matrix $\mathbf{S} = T^{-1}\mathbf{X}^*\mathbf{C}\mathbf{X}$ that satisfies $\mathbf{X}\mathbf{X}^* = \boldsymbol{\mathcal{W}}$ and Eq. (4.2.25). Similarly, we define $\tilde{\mathbf{S}} = T^{-1}\tilde{\mathbf{X}}^*\mathbf{C}\tilde{\mathbf{X}}$ with $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^* = \tilde{\boldsymbol{\mathcal{W}}}$. Using (4.2.25) and omitting the argument $z$ and $\tilde{z}$, we can rewrite (5.2.11) as

$$\psi(z,\tilde{z}) \;\sim\; \frac{1}{q\tilde{q}z\tilde{z}}\left[\frac{(\tilde{q}z-q\tilde{z})\mathfrak{g}_{\tilde{\mathbf{S}}}^2}{\mathfrak{g}_{\mathbf{S}}-\mathfrak{g}_{\tilde{\mathbf{S}}}} + \frac{(q-\tilde{q})\mathfrak{g}_{\tilde{\mathbf{S}}}}{\mathfrak{g}_{\mathbf{S}}-\mathfrak{g}_{\tilde{\mathbf{S}}}}\right] + \frac{\mathfrak{g}_{\mathbf{S}}+\mathfrak{g}_{\tilde{\mathbf{S}}}}{q\tilde{z}} - \frac{1-q}{qz\tilde{z}}. \tag{5.2.12}$$

We see from (5.2.7) that it now suffices to consider the limit $\eta\to 0$ in order to get the desired result. To lighten the notations, let us define

$$m_0(\lambda) \equiv \lim_{\eta\to 0}\mathfrak{g}_{\mathbf{S}}(\lambda-\mathrm{i}\eta) = m_R(\lambda) + \mathrm{i}m_I(\lambda) \tag{5.2.13}$$

with

$$m_R(\lambda) = q\mathfrak{h}_{\mathbf{E}}(\lambda) + \frac{1-q}{\lambda}, \qquad m_I(\lambda) = q\rho_{\mathbf{E}}(\lambda) + (1-q)\delta_0, \tag{5.2.14}$$

where $\mathfrak{h}_{\mathbf{E}}$ is the Hilbert transform of $\rho_{\mathbf{E}}$. Note that this relation follows from Eq. (4.2.1). We also define $\tilde{m}_0(\lambda) = \lim_{\eta\to 0}\mathfrak{g}_{\tilde{\mathbf{S}}}(\lambda-\mathrm{i}\eta)$ and denote by $\tilde{m}_R, \tilde{m}_I$ the real and imaginary part,

respectively. Then, the asymptotic behavior of Eq. (5.0.4) for any $\lambda \in \operatorname{supp} \varrho$ and $\tilde{\lambda} \in \tilde{\varrho}$ is given by (see [41] for a detailed derivation)

$$\Phi_{q,\tilde{q}}(\lambda,\tilde{\lambda}) = \frac{2(\tilde{q}\lambda - q\tilde{\lambda})\big[m_R|\tilde{m}_0|^2 - \tilde{m}_R|m_0|^2\big] + (\tilde{q}-q)\big[|\tilde{m}_0|^2 - |m_0|^2\big]}{\lambda\tilde{\lambda}\Big[(m_R-\tilde{m}_R)^2 + (m_I+\tilde{m}_I)^2\Big]\Big[(m_R-\tilde{m}_R)^2 + (m_I-\tilde{m}_I)^2\Big]}. \tag{5.2.15}$$

An interesting consistency check is when $\tilde{q} = 0$ in which case the sample eigenvalues coincide with the true ones for the tilde matrices, i.e. $\tilde{\lambda} \to \mu$. In this case we fall back on the framework of the previous section, i.e. obtaining the overlaps between the eigenvectors of $\mathbf{E}$ and $\mathbf{C}$. In that case, one can easily check that $\tilde{m}_R = 1/\mu$ and $\tilde{m}_I = 0$. Hence, we deduce from (5.2.15) that

$$\Phi_{q,\tilde{q}=0}(\lambda,\mu) = \frac{q}{\lambda\mu\big[(m_R - 1/\mu)^2 + m_I^2\big]} = \frac{q\mu}{\lambda|1-\mu m_0(\lambda)|^2}, \tag{5.2.16}$$

which is another way to write (5.1.6) after applying the formula (4.2.25) in the limit $\eta \to 0$. It therefore shows that the result (5.2.15) generalizes Eq. (5.1.6) in the sense that we are able to study the mean squared overlaps between two possibly noisy sample estimates. Note that in the case $\tilde{q} = q$, Eq. (5.2.15) can be somewhat simplified to:

$$\Phi(\lambda,\tilde{\lambda}) = \frac{(\lambda-\lambda')\big(m_R(\lambda)|m_0(\lambda')|^2 - m_R(\lambda')|m_0(\lambda)|^2\big)}{\lambda\tilde{\lambda}\Big[(m_R-\tilde{m}_R)^2 + (m_I+\tilde{m}_I)^2\Big]\Big[(m_R-\tilde{m}_R)^2 + (m_I-\tilde{m}_I)^2\Big]}, \tag{5.2.17}$$

that becomes when $\tilde{\lambda} = \lambda$ [41],

$$\Phi(\lambda,\lambda) = \frac{q}{2\lambda^2}\frac{|m_0(\lambda)|^4\partial_\lambda\big[m_R(\lambda)/|m_0(\lambda)|^2\big]}{m_I^2(\lambda)|\partial_\lambda m_0(\lambda)|^2}. \tag{5.2.18}$$

This last "self-overlap" result quantifies the stability of the eigenvectors $\mathbf{u}_i$ and $\tilde{\mathbf{u}}_j$ associated to the very same eigenvalue $\lambda$ when they both come from the same population matrix $\mathbf{C}$.

Now that we have all the theoretical formulas, let us now give some applications of the formula (5.2.17) as they will highlight that we can indeed find genuine information about the spectrum of $\mathbf{C}$ from the mean squared overlap (5.0.4). We emphasize that all the following applications are performed in the case $q = \tilde{q}$ in order to give more insights about the results. As usual, we begin with the null hypothesis $\mathbf{C} = I_N$ as it will serve as the benchmark when we shall deal with more structured spectrum. As we shown in Section (3.1.2), the Stieltjes transform $\mathfrak{g}_{\mathbf{E}}$, and thus $\mathfrak{g}_{\mathbf{S}}$ is explicit and obtained from the Marčenko-Pastur density. More precisely, we deduce from Eq. (3.1.41) and (4.2.25) that $\mathfrak{g}_{\mathbf{S}}$ is given by

$$G_{\mathbf{S}}(z) = \frac{z+q-1-\mathrm{i}\sqrt{4zq-(z+q-1)^2}}{2z} \tag{5.2.19}$$

for any $z \in \mathbb{C}_-$. It is easy to see using the definition (5.2.13) that we have

$$m_R(\lambda) = \frac{\lambda+q-1}{2z}, \qquad m_I(\lambda) = \frac{\sqrt{4zq-(z+q-1)^2}}{2z}. \tag{5.2.20}$$

Hence, one obtains $|m_0(\lambda)|^2 = \lambda^{-1}$ and $|m_0'(\lambda)|^2 = q/(2\lambda^2)$, and by plugging this expressions into Eq. (5.2.18), we eventually get

$$\Phi_{q,q}(\lambda,\lambda) = 1, \tag{5.2.21}$$

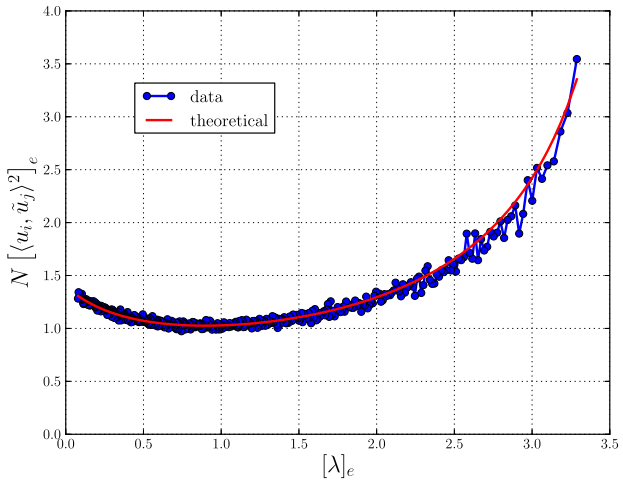



FIGURE 5.2.1. Evaluation of $N\mathbf{E}\langle \mathbf{u}_i\,, \tilde{\mathbf{u}}_i\rangle^2$ with $N = 500$ and $q = \tilde{q} = 0.5$. The population matrix $\mathbf{C}$ is given by an Inverse-Wishart with parameter $\kappa$ and the sample covariance matrices $\mathbf{S}$ and $\tilde{\mathbf{S}}$ are generated from a multivariate Gaussian distribution. The empirical average (blue points) is taken over 200 realizations and the theoretical prediction Eq. (5.2.18) (red line) is evaluated at any $[\lambda_i]_e$.

for any $\lambda \in [(1-\sqrt{q})^2, (1+\sqrt{q})^2]$. This simple result was expected as it corresponds to the case where the spectrum of $\mathbf{C}$ has *no* genuine structure, so all the anisotropy in the problem is induced by the noise, which is independent in the two samples.

Next, we consider a more structured example of a population correlation matrix $\mathbf{C}$. A convenient case that can be treated analytically is when $\mathbf{C}$ to be an inverse Wishart matrix, i.e. distributed according to (3.1.58) with $\kappa > 0$ defined in Eq. (3.1.54). As we saw in the previous chapter, the Stieltjes transform $\mathfrak{g}_{\mathbf{E}}(z)$ is explicit in this case (see Eq. (4.2.33)). Going back to Eq. (5.2.18), one can readily obtain from Eq. (4.2.33),

$$m_R(\lambda) \;=\; \frac{\lambda(1+q\kappa)+q\kappa(1-q)}{\lambda(\lambda+2q\kappa)}, \qquad m_I(\lambda) \;=\; q\frac{\sqrt{\lambda-\lambda_-^{\rm iw}}\sqrt{\lambda_+^{\rm iw}-\lambda}}{\lambda(\lambda+2q\kappa)}, \tag{5.2.22}$$

with $\lambda \in [\lambda_-^{\rm iw}, \lambda_+^{\rm iw}]$ where $\lambda_\pm^{\rm iw}$ is defined in (4.2.34). Plugging these expressions into Eq. (5.2.18) and after elementary computations, one finds

$$\Phi_{q,q}(\lambda,\lambda) = \frac{(1+q\kappa)(\lambda+2q\kappa)^2}{2q\kappa\big[2\lambda(1+\kappa(1+q)) - \lambda^2\kappa + \kappa(-1+2q(1+q\kappa))\big]}. \tag{5.2.23}$$

The immediate consequence of this last formula is that in the presence of *anisotropic* correlations, the mean squared overlap (5.0.4) clearly deviates from the null hypothesis $\Phi(\lambda,\lambda) = 1$. In the nearly isotropic limit $\kappa \to \infty$, that corresponds to the limit $\mathbf{C} \to \mathbf{I}_N$, one gets [41]

$$\Phi(\lambda,\tilde{\lambda}) \underset{\kappa\to\infty}{\sim} \left[1 + \frac{(\lambda-1)(\tilde{\lambda}-1)}{2q^2\kappa} + \mathcal{O}(\kappa^{-2})\right], \tag{5.2.24}$$

which is in fact *universal* in this limit (i.e. independent of the precise statistical properties of the matrix $\mathbf{C}$), provided the eigenvalue spectrum of $\mathbf{C}$ has a variance given by $(2\kappa)^{-1} \to 0$ [41]. In the general case, we provide a numerical illustration of this last statement in Figure 5.2.1 with $\kappa = 5$, $N = 500$ and $q = 0.5$. As we expect $\lambda_i \approx \tilde{\lambda}_i$ for any $i \in [\![1, N]\!]$, we compare our theoretical result (5.2.23) with the empirical average $[\langle \mathbf{u}_i \, , \tilde{\mathbf{u}}_i \rangle^2]_e$ taken over 200 realizations of $\mathbf{S}$ and we see that the agreement is again excellent. We therefore conclude that a possible application of (5.2.15) is to estimate directly the statistical texture of $\mathbf{C}$ using only sample eigenvectors: see Section 8 for an interesting example.

We now present an alternative derivation of $\Phi_{q,\tilde{q}}$ that uses the result of the Section 5.1. The following argument is very general and might be useful when considering the overlaps between the eigenvectors of more general random matrices. The starting point is the orthonormality of the true eigenbasis, i.e. $\mathbf{V}\mathbf{V}^* = \mathbf{I}_N$ for $\mathbf{V} := [\mathbf{v}_1, \ldots, \mathbf{v}_N]$. Hence, we may always write

$$\langle \mathbf{u}_i \, , \tilde{\mathbf{u}}_j \rangle = \left\langle \mathbf{u}_i \, , \, \left( \sum_{k=1}^{N} \mathbf{v}_k \mathbf{v}_k^* \right) \tilde{\mathbf{u}}_j \right\rangle = \sum_{k=1}^{N} \langle \mathbf{u}_i \, , \mathbf{v}_k \rangle \langle \mathbf{v}_k \, , \tilde{\mathbf{u}}_j \rangle \tag{5.2.25}$$

Using the results of Section 5.1, we rename the overlaps $\langle \mathbf{u}_i \, , \mathbf{v}_k \rangle = \sqrt{\Phi(\lambda_i, \mu_k)/N} \times \varepsilon(\lambda_i, \mu_k)$ where $\Phi(\lambda, \mu)$ is defined in (5.0.3) and $\varepsilon(\lambda, \mu)$ are random variables of unit variance. Hence, we have

$$\langle \mathbf{u}_i \, , \tilde{\mathbf{u}}_j \rangle = \frac{1}{N} \sum_{k=1}^{N} \sqrt{\Phi(\lambda_i, \mu_k) \Phi(\tilde{\lambda}_j, \mu_k)} \; \varepsilon(\lambda_i, \mu_k) \varepsilon(\tilde{\lambda}_j, \mu_k). \tag{5.2.26}$$

As noticed in [41], by averaging over the noise and making an "ergodic hypothesis" [64] – according to which all signs $\varepsilon(\mu, \lambda)$ are in fact independent from one another in the large $N$ limit – one ends up with the following rather intuitive convolution result for the square overlaps:

$$\Phi_{q,\tilde{q}}(\lambda_i, \tilde{\lambda}_j) = \frac{1}{N} \sum_{k=1}^{N} \Phi(\lambda_i, \mu_k) \Phi(\tilde{\lambda}_j, \mu_k) \tag{5.2.27}$$

It turns out that this expression is completely general and exactly equivalent to Eq. (5.2.17) if we replace the overlaps function $\Phi$ by (5.1.6). However, whereas this expression still contains some explicit dependence on the structure of the pure matrix $\mathbf{C}$, it has completely disappeared in Eq. (5.2.17). An interesting application of the formula (5.2.27) is when the spectrum of $\mathbf{E}$ (and $\tilde{\mathbf{E}}$) contains a finite number of outliers. Using the results (5.1.10) and (5.1.12) yields in the LDL and for $i \leqslant r$:

$$\Phi_{q,\tilde{q}}(\lambda_i, \tilde{\lambda}_i) \approx \mu_1^2 \frac{\theta'(\mu_1) \tilde{\theta}'(\mu_1)}{\theta(\mu_1) \tilde{\theta}(\mu_1)}, \tag{5.2.28}$$

where we recall that the function $\theta$ is defined in (4.3.12) and we define $\tilde{\theta}$ accordingly by replacing $q$ with $\tilde{q}$. Note that we can express (5.2.28) in terms of observable variables by noticing that

$$\mu_1 = \frac{1}{\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1)}, \qquad \theta'(\mu_1) = \frac{-1}{\mathfrak{g}'_{\underline{\mathbf{S}}}(\theta(\mu_1)) \mu_i^2}, \tag{5.2.29}$$

that we plug into (5.2.28) to conclude that

$$\Phi_{q,\tilde{q}}(\lambda_1, \tilde{\lambda}_1) \approx \frac{\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1)}{\lambda_1 \mathfrak{g}'_{\underline{\mathbf{S}}}(\lambda_1)} \frac{\mathfrak{g}_{\underline{\tilde{\mathbf{S}}}}(\lambda_1)}{\tilde{\lambda}_1 \mathfrak{g}'_{\underline{\tilde{\mathbf{S}}}}(\lambda_1)}. \tag{5.2.30}$$

This expression becomes even simpler when $q = \tilde{q}$ as it becomes

$$\Phi_{q,q}(\lambda_1, \tilde{\lambda}_1) \approx \left( \frac{\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1)}{\lambda_1 \mathfrak{g}'_{\underline{\mathbf{S}}}(\lambda_1)} \right)^2 . \tag{5.2.31}$$

One further deduces from (5.1.10) and (5.1.12) that for $i \leqslant r$, $\Phi_{q,\tilde{q}}(\lambda_i, \tilde{\lambda}_j) \sim \mathcal{O}(N^{-1})$ for any $j \neq i$.

# Chapter 6

# Bayesian Random Matrix Theory

We saw in the previous chapter that RMT allows one to make precise statements about large empirical covariance matrices. In particular, we emphasized that the classical sample estimator $\mathbf{E}$ is not consistent in the high-dimensional limit as the sample spectral density $\varrho_{\mathbf{E}}$ deviates significantly from the true spectrum whenever $q = \mathcal{O}(1)$. There have been many attempts in the literature to correct this "curse of dimensionality" using either heuristics or decision theoretic arguments (see Section 8.2 for a summary of these attempts). Despite the strong differences in these approaches, all of them fall into the class of so-called *shrinkage* estimators, to wit, one seeks the best way to "clean" the sample eigenvalues in such a way that the estimator is as robust as possible to the measurement noise.

In the previous chapter, we insisted that the bulk sample eigenvectors are delocalized, with a projection of order $N^{-1/2}$ in all directions, which means that they are extremely noisy estimators of the population eigenvectors. As a consequence, the naive idea of replacing the sample eigenvalues by the estimated true ones, obtained by inverting the Marčenko-Pastur equation, will not necessarily lead to satisfactory results – it would only be the optimal strategy if we had a perfect knowledge of the eigenvectors of $\mathbf{C}$. Hence, we are left with a very complicated problem: how can estimate "accurately" the matrix $\mathbf{C}$ is the high-dimensional regime knowing that the eigenvalues are systematically biased and the eigenvectors nearly completely unknown?

The aim of the present chapter and the following one is to answer this question by developing an optimal strategy to estimate $\mathbf{C}$, consistent with the quality ratio $q$. By optimal, we mean that the estimator we aim to construct has to minimize a given loss function. A natural optimality criteria is the squared distance between the estimator – called $\Xi(\mathbf{E})$ henceforth – and the true matrix $\mathbf{C}$. As for the James-Stein estimator, we expect that "mixed" estimators provide better performance than "classical" ones (like the Pearson estimator) in high-dimension. In that respect, we introduce a Bayesian framework which, loosely speaking, allows one to introduce probabilistic models that encode the available data through the notion of *prior* belief.

The fact that probabilities represent degrees of belief is at the heart of Bayesian inference and as explained in the introduction to this chapter, this theory has enjoyed many success, especially in a high-dimensional framework. The central tool of this theory is the well known Bayes formula that allows one to introduce the concept of conditional probability. There are many different ways to make use of this formula and the corresponding schools of thought are referred to as empirical, subjective or objective Bayesians (see e.g. [81] for an exhaustive presentation). Here we shall not discuss these different points of view but rather focus on the inference part of the problem. More precisely, our aim in this chapter is to construct a Bayesian

estimator for $\Xi(\mathbf{E})$. We therefore organize this chapter as follows. In the first part, we recall some basic results on Bayesian inference and introduce the estimator that will be of interest to us. We then re-consider the famous "linear shrinkage" estimator, mentioned in Eq. (2.1.9), that interpolates linearly between the sample estimator and the identity matrix through the notion of *conjugate priors*. Finally, we consider the class of rotational invariant prior where the RMT formalism introduced in the previous chapters is applied to derive an optimal estimator for $\mathbf{C}$, which will turn out to be more efficient that all past attempts – see Chapter 9.

## 6.1 Bayes optimal inference: some basic results

**6.1.1. Posterior and joint probability distributions.** Bayesian theory allows one to answer, at leasr in principle, the following question: given the observation matrix $\mathbf{Y}$, how can we best estimate $\mathbf{C}$ if some prior knowledge of the statistics of $\mathbf{C}$ is available? This notion of prior information has been the subject of many controversies but is a cornerstone to Bayes inference theory. More precisely, the main concept of Bayesian inference is the well-known Bayes formula

$$\mathcal{P}(\mathbf{C}|\mathbf{Y}) = \frac{\mathcal{P}(\mathbf{Y}|\mathbf{C})\mathcal{P}(\mathbf{C})}{\mathcal{P}(\mathbf{Y})} \tag{6.1.1}$$

where

- $\mathcal{P}(\mathbf{C}|\mathbf{Y})$ is the *posterior* probability for $\mathbf{C}$ given the measurements $\mathbf{Y}$.
- $\mathcal{P}(\mathbf{Y}|\mathbf{C})$ is the *likelihood* function, modelling the measurement process.
- $\mathcal{P}(\mathbf{C})$ is called the prior probability of $\mathbf{C}$, that is to say the prior belief (or knowledge) about what $\mathbf{C}$ should look like before being corrupted by the measurement noise.
- $\mathcal{P}(\mathbf{Y})$ is the marginal distribution, sometimes called the *evidence*.

Note that the marginal distribution is often considered as a mere normalization constant (or partition function) since it is given by

$$\mathcal{P}(\mathbf{Y}) = \int \mathcal{D}\mathbf{C}\mathcal{P}(\mathbf{C})\mathcal{P}(\mathbf{Y}|\mathbf{C}). \tag{6.1.2}$$

Furthermore, we shall often use the concept of *joint* probability distribution defined by

$$\mathcal{P}(\mathbf{C}, \mathbf{Y}) = \mathcal{P}(\mathbf{Y}|\mathbf{C})\mathcal{P}(\mathbf{C}). \tag{6.1.3}$$

Thus, the two crucial inputs in a Bayesian model is the likelihood process and the prior distribution. Learning using a Bayesian framework can actually be split in two different steps, which in our context are:

1. Set a joint probability distribution $\mathcal{P}(\mathbf{C}, \mathbf{Y})$ defined as the product of the prior distribution and the likelihood function, i.e.

$$\mathcal{P}(\mathbf{C}, \mathbf{Y}) = \mathcal{P}(\mathbf{Y}|\mathbf{C})\mathcal{P}(\mathbf{C}). \tag{6.1.4}$$

2. Test the consistency of the posterior distribution $\mathcal{P}(\mathbf{C}|\mathbf{Y})$ on the available data.

We emphasize that the presence of a prior distribution does not imply that $\mathbf{C}$ is stochastic, it simply encodes the degree of belief about the structure of $\mathbf{C}$. The main advantage of adopting this point of view is that it facilitates the interpretation of the statistical results. For instance, a Bayesian (probability) interval tells us how probable is the value of a parameter we attempt to estimate. This is in contrast to the frequentist interval, which is only defined with respect to a sequence of similar realizations (confidence interval). We will discuss the difference between these points of view in the next paragraph.

**6.1.2. Bayesian inference.** The notion of Bayesian inference is related to the concept of the so-called *Bayes risk*. In our problem, we want to estimate the true covariance matrix $\mathbf{C}$ given our sample data $\mathbf{Y}$; we shall denote by $\Xi(\mathbf{Y})$ this estimator. There are two ways to think about this problem: the frequentist and the Bayesian approach. We will detail the difference between these two in this section.

Let us introduce a loss function $\mathcal{L}(\mathbf{C}, \Xi(\mathbf{Y}))$ that quantifies how far is the estimator from the true quantity $\mathbf{C}$. In general, this loss function is assumed to be a nonnegative convex function with $\mathcal{L}(\mathbf{C}, \mathbf{C}) = 0$. The traditional *frequentist* approach is to evaluate the performance of a given estimator by averaging the loss function over different sets of observations, for a fixed $\mathbf{C}$.

An alternative point of view is to think that the precise nature of $\mathbf{C}$ is unknown. This change in the point of view has to be encoded in the inference problem and one way to do it is to look at the average value of the loss function over all the *a priori* possible realizations of $\mathbf{C}$, and not on the realizations of $\mathbf{Y}$ itself. This is Bayes optimization strategy and the corresponding the decision rule is the so-called *Bayes risk function* that is defined as:

$$R^{\text{Bayes}}(\mathcal{L}(\mathbf{C}, \Xi(\mathbf{Y}))) \;:=\; \Big\langle \mathcal{L}(\mathbf{C}, \Xi(\mathbf{Y})) \Big\rangle_{\mathcal{P}(\mathbf{C},\mathbf{Y})}, \tag{6.1.5}$$

where, unlike the frequentist approach, the expectation value is taken over the joint probability of $\mathbf{Y}$ and $\mathbf{C}$. One of the most commonly used loss function is the squared Hilbert-Schmidt (or Euclidean) $L_2$ norm, i.e.,

$$\mathcal{L}^{L_2}(\mathbf{C}, \Xi(\mathbf{Y})) = \text{Tr}\left[(\mathbf{C} - \Xi(\mathbf{Y}))(\mathbf{C} - \Xi(\mathbf{Y}))^*\right]. \tag{6.1.6}$$

Using that covariance matrices are symmetric and applying Bayes rule, we see that

$$\begin{aligned} R^{\text{Bayes}} &= \Big\langle \Big\langle \text{Tr}\left[(\mathbf{C} - \Xi(\mathbf{Y}))^2\right] \Big\rangle_{\mathcal{P}(\mathbf{Y}|\mathbf{C})} \Big\rangle_{\mathcal{P}(\mathbf{C})} \\ &= \Big\langle \Big\langle \text{Tr}\left[(\mathbf{C} - \Xi(\mathbf{Y}))^2\right] \Big\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{Y})} \Big\rangle_{\mathcal{P}(\mathbf{Y})}, \end{aligned} \tag{6.1.7}$$

where we have used that marginal distributions are positive in order to interchange the order of integration in the second line.

The optimal Bayes estimator is defined as follows: let us denote by $\mathcal{M}_N(\mathbf{Y})$ is the set of $N \times N$ positive definite matrices which are functions of $\mathbf{Y}$. This defines the set of admissible estimators of $\mathbf{C}$. Then the Bayes estimator associated to the loss function (6.1.6) is given by the *minimum mean squared error* (MMSE) condition, i.e.

$$\Xi^{\text{MMSE}} \;\equiv\; \Xi^{\text{MMSE}}(\mathbf{Y}) \;:=\; \underset{\Xi(\mathbf{Y}) \in \mathcal{M}_N(\mathbf{Y})}{\text{argmin}} \Big\langle \mathcal{L}^{L_2}(\mathbf{C}, \Xi(\mathbf{Y})) \Big\rangle_{\mathcal{P}(\mathbf{C},\mathbf{Y})}, \tag{6.1.8}$$

Expanding (6.1.7), it is readily seen that the MMSE estimator is given by the posterior mean:

$$\Xi^{\text{MMSE}} = \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{Y})}. \tag{6.1.9}$$

Note that the natural choice of the loss function may depend on the nature of the problem. Other loss functions often lead to different Bayes estimators, but we do not investigate such generalizations here.

## 6.2 Setting the Bayesian framework

Now that we have derived the optimal estimator we are looking for, we still need to parametrize the joint probability function $\mathcal{P}(\mathbf{C}, \mathbf{Y})$. There are thus two inputs in the Bayesian model: the likelihood function and the prior distribution, and we focus on the former quantity in this section.

In a multivariate framework, the most common assumption (but not necessarily the most realistic) is that the measurement process $\mathbf{Y}$ is Gaussian, that is to say,

$$\mathbb{P}(\mathbf{Y}|\mathbf{C}) = \frac{1}{(2\pi)^{\frac{NT}{2}} \det(\mathbf{C})^{\frac{T}{2}}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \sum_{i,j=1}^{N} Y_{it} \mathbf{C}^{-1}_{i,j} Y_{jt} \right\}. \tag{6.2.1}$$

It is easy to see that this is of the Boltzmann type, as in Eq. (3.1.1). More precisely, using the cyclic property of the trace operator one gets

$$\sum_{t=1}^{T} \sum_{i,j=1}^{N} Y_{it} \mathbf{C}^{-1}_{ij} Y_{jt} = \text{Tr}\left[\mathbf{Y}\mathbf{C}^{-1}\mathbf{Y}^*\right] = T\text{Tr}\left[\mathbf{E}\mathbf{C}^{-1}\right].$$

Thus, the $N$-variate Gaussian likelihood function can be written as

$$\mathcal{P}(\mathbf{Y}|\mathbf{C}) = \frac{1}{(2\pi)^{\frac{NT}{2}}} \exp\left\{ -\frac{T}{2} \text{Tr}\big[\log(\mathbf{C}) + \mathbf{E}\mathbf{C}^{-1}\big] \right\} \equiv \mathcal{P}(\mathbf{E}|\mathbf{C}), \tag{6.2.2}$$

where we used Jacobi's formula $\det(\mathbf{A}) = \exp[\text{Tr}\log\mathbf{A}]$ for any square matrix $\mathbf{A}$. As a result, we can rewrite the inference problem as a function of the sample covariance matrix $\mathbf{E}$, and in particular, the MMSE estimator becomes

$$\Xi^{\text{MMSE}} \equiv \Xi^{\text{MMSE}}(\mathbf{E}) := \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}. \tag{6.2.3}$$

After a little thought, this set-up agrees perfectly with the framework developed in the Chapters 4 and 5 above. Indeed, in those sections we studied the spectral properties of the sample covariance matrix $\mathbf{E}$ given the limiting spectral distribution of $\mathbf{C}$ (the so-called "direct problem" introduced in Section 4.2.1). Differently said, the Marčenko-Pastur equation (4.2.1) has a natural Bayesian interpretation: it provides the (limiting) spectral density of $\mathbf{E}$ conditional to a population covariance matrix $\mathbf{C}$ that we choose within a specific prior probabilistic ensemble.

## 6.3 Conjugate prior estimators

Once we have set the likelihood function, the next step is to focus on the prior distribution $\mathcal{P}(\mathbf{C})$, keeping in mind that the ultimate goal is to compute the Bayes posterior mean estimator

(6.2.3). Unfortunately, the evaluation of the posterior distribution often leads to non trivial computations and closed-form estimators are thus scarce. Nonetheless, there exists some classes of prior distributions where the posterior distribution can be computed exactly. The one that interests us is known as the class of 'conjugate priors' in Statistics. Roughly speaking, suppose that we know the likelihood distribution $\mathcal{P}(\mathbf{E}|\mathbf{C})$, then the prior distribution $\mathcal{P}(\mathbf{C})$ and the posterior distribution $\mathcal{P}(\mathbf{C}|\mathbf{E})$ are said to be conjugate if they belong to the same family of distribution.

As an illustration, let us consider a warmup example example before going back to the estimation of the covariance. Suppose that we want to estimate the mean vector – say $\boldsymbol{\mu}$ – given the $N$-dimensional vector data $\mathbf{y}$ we observe. Moreover, assume that the likelihood function is a multivariate Gaussian distribution with a known covariance matrix $\sigma^2\mathbf{I}_N$. Then, by taking a Gaussian prior on $\boldsymbol{\mu}$ with zero "mean" and "covariance" matrix $\tau^2\mathbf{I}_N$, one can easily check that

$$\mathcal{P}(\boldsymbol{\mu}|\mathbf{y}) = \mathcal{N}_N\left(\frac{\tau^2}{\tau^2+\sigma^2}\mathbf{y}, \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\mathbf{I}_N\right). \tag{6.3.1}$$

Therefore, the Bayes MMSE (6.1.9) of $\boldsymbol{\mu}$ is given by

$$\langle\boldsymbol{\mu}\rangle_{\mathcal{P}(\boldsymbol{\mu}|\mathbf{y})} = \left(1 - \frac{\sigma^2}{\sigma^2+\tau^2}\right)\mathbf{y}, \tag{6.3.2}$$

that is – loosely speaking – the celebrated James-Stein estimator [96]. In fact, the James-Stein estimator follows using the evidence $\mathcal{P}(\mathbf{y})$, and this approach is known as *empirical Bayes* (see at the end of this section for more details).

One can now wonder whether we can generalize this conjugate prior property to the case of covariance matrices under a measurement process characterized by the likelihood function $\mathcal{P}(\mathbf{E}|\mathbf{C})$ given in Eq. (6.2.2). Again, we will see that conjugate prior approach yields a very interesting result. Using the potential theory formalism introduced in (3.1.1) and in Section 3.1.2, it is easy to see from Eq. (6.2.2) that the potential function associated to a Gaussian likelihood function reads

$$V_q(\mathbf{E},\mathbf{C}) = \frac{1}{2q}\left[\log(\mathbf{C}) + \mathbf{E}\mathbf{C}^{-1}\right], \tag{6.3.3}$$

that is clearly the Inverse-Wishart distribution encountered in (3.1.58) in the presence of an external field $\mathbf{E}$. Hence, let us introduce an inverse-Wishart ensemble with two hyperparameters $\{\gamma,\kappa\}$ as a prior for $\mathbf{C}$:[1]

$$\mathcal{P}(\mathbf{C}) = Z\exp\left\{-N\mathrm{Tr}\left[\gamma\log\mathbf{C} + \kappa\mathbf{C}^{-1}\right]\right\},$$

with $Z$ a normalization constant that depends on $\gamma,\kappa$ and $N$. For simplicity, we impose that $\langle\mathbf{C}\rangle_{\mathcal{P}(\mathbf{C})} = \mathbf{I}_N$, we easily obtain (omitting term in $\mathcal{O}(N^{-1})$) that $\gamma = \kappa+1$. This is the convention that we adopt henceforth. Using Bayes rule and the Gaussian likelihood function (6.2.2), we find that the posterior distribution is also an inverse-Wishart distribution of the form:

$$\mathcal{P}(\mathbf{C}|\mathbf{E}) \propto \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[(T+\nu+N+1)\log\mathbf{C} + T(2q\kappa\mathbf{I}_N + \mathbf{E})\mathbf{C}^{-1}\right]\right\}, \tag{6.3.4}$$

[1] More precisely, it is an inverse Wishart distribution $\mathcal{I}W_N(N, N(2\gamma-1)-1, 2N\kappa\mathbf{I}_N)$ defined in Eq. (3.1.58).

where we have defined $\nu := N(2\kappa + 1) - 1$. As a consequence, we expect the Bayes estimator to be explicit like the James-Stein estimator (6.3.2) and the final result for $\Xi^{\text{MMSE}}$ is obtained from (3.1.59):

$$\Xi^{\text{MMSE}} = \frac{T}{T + \nu - N - 1}(2q\kappa \mathbf{I}_N + \mathbf{E}). \tag{6.3.5}$$

We see that the estimator we get is in the same spirit than the so-called James-Stein estimator in the sense that the estimator *shrinks* the sample covariance $\mathbf{E}$ toward the identity with intensity fixed by the hyperparameters $\gamma$ and $\kappa$. This estimator is known as the *linear shrinkage* estimator, first obtained in [87],

$$\Xi^{\text{lin}} := \frac{T}{T + \nu - N - 1}(2q\kappa \mathbf{I}_N + \mathbf{E}) \approx \frac{1}{1 + 2q\kappa}\mathbf{E} + \frac{2q\kappa}{1 + 2q\kappa}\mathbf{I}_N + \mathcal{O}(T^{-1}), \tag{6.3.6}$$

where we used that $T \to \infty$ with $q = N/T$ finite in the RHS. All in all, we have derived the linear shrinkage estimator:

$$\Xi^{\text{lin}} = \alpha_s \mathbf{E} + (1 - \alpha_s)\mathbf{I}_N \quad \text{where} \quad \alpha_s := \frac{1}{1 + 2q\kappa} \in [0, 1], \ \kappa > 0\,. \tag{6.3.7}$$

As for the James-Stein estimator, this estimator tells us to *shrink* the sample covariance matrix $\mathbf{E}$ toward the identity matrix (our prior) with an intensity given by $\alpha_s$. We give a simple illustration of how this estimator transforms the eigenvalues in Figure 6.3.1. In particular, we see that small eigenvalues are lifted upwards while the top ones are shrunk downwards. Furthermore, it is easy to see this estimator shares the same eigenvectors than the sample covariance matrix $\mathbf{E}$. This property will be important in the following.

The remaining question is how can we consistently choose the parameter $\kappa$ (or directly $\alpha_s$) in order to use this estimator in practice? In [87], Haff promoted an empirical Bayes approach similar to the work of James and Stein [96]. In the high-dimensional regime, the Ledoit & Wolf [115] noticed that this approach may suffer from the fact that classical estimator becomes unreliable and consequently proposed a consistent estimator of $\alpha_s$. There also exists more straightforward methods to estimate the parameter $\kappa$ directly from the data, using RMT tools. We summarize all these approaches in Section 8.2.1.

One may finally remark that the above derivation of the linear shrinkage estimator can be extended to the case where the prior is different from the identity matrix. Suppose that the prior distribution of $\mathbf{C}$ is a generalized inverse-Wishart distribution:

$$\mathcal{P}(\mathbf{C}) = Z \exp\left\{-N \text{Tr}\left[\gamma \log \mathbf{C} + \kappa \mathbf{C}_0 \mathbf{C}^{-1}\right]\right\},$$

where $\mathbf{C}_0$ is a certain matrix (referred as a *fundamental* or *prior* matrix) with a possibly non-trivial structure encoding what we believe about the problem at hand. In this case, it is easy to see that the above linear estimator still holds, with:

$$\Xi^{\text{lin}} = \alpha_s \mathbf{E} + (1 - \alpha_s)\mathbf{C}_0 \qquad \alpha_s \in [0, 1]. \tag{6.3.8}$$

Note that when $\mathbf{C}_0 \neq \mathbf{I}_N$, $\mathcal{P}(\mathbf{C})$ is no longer rotationally invariant. A simple example is to choose $\mathbf{C}_0 = (1 - \rho)\mathbf{I}_N + \rho \mathbf{J}$, where $\mathbf{J}$ has all its elements equal to unity. This corresponds to a one-factor model in financial applications, where the correlations between any pair of stocks is constant. This can be seen as a spike correlation model, as was shown in (4.3.6) above, with $\underline{\mathbf{C}} = \mathbf{I}_N$, $r = 1$, $v_1 = (1, 1, \ldots, 1)$ and $d_1 = (N - 1)\rho$.

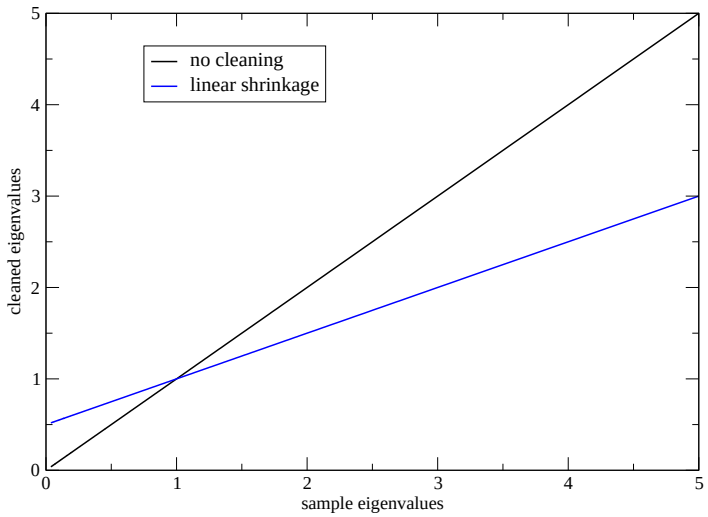



FIGURE 6.3.1. Impact of the linear shrinkage (6.3.7) with $\alpha_s = 0.5$ on the eigenvalues (blue line), compared to the sample eigenvalues (black line). We see that the small eigenvalues are shifted upward and the large ones are brought downward.

We now present the empirical Bayes approach through the "non-observable" James-Stein estimator (6.3.2). This approach can be useful in order to estimate parameters directly from the data but it requires that the marginal distribution can be computed exactly. If we reconsider the framework of the estimator (6.3.2), it is not hard to see that the evidence $\mathcal{P}(\mathbf{y})$, defined in (6.1.2), is given by

$$\mathcal{P}(\mathbf{y}) \sim \mathcal{N}_N(0, (\sigma^2 + \tau^2)\mathbf{I}_N)\,. \tag{6.3.9}$$

Recall from (6.3.2) that our aim is to estimate the ratio $\sigma^2/(\sigma^2+\tau^2)$ where $\sigma^2$ is known. To that end, we notice from (6.3.9) that

$$\|\mathbf{y}\|_2^2 \sim (\sigma^2 + \tau^2)\chi_N^2\,, \tag{6.3.10}$$

where $\|\cdot\|_2$ is the $\mathbb{L}_2$ norm and $\chi_N^2$ is the chi-square distribution with $N$ degrees of freedom. Therefore, we can conclude by maximum likelihood estimation that

$$\frac{\sigma^2 \times \max(N-2, 0)}{\|\mathbf{y}\|_2^2} \approx \frac{\sigma^2}{\sigma^2+\tau^2}\,, \tag{6.3.11}$$

which yields an estimator of the unobservable term in Eq. (6.3.2). Hence, if we plug this sample estimate into (6.3.2), it yields the celebrated James-Stein estimator:

$$\hat{\boldsymbol{\mu}}_{\text{JS}} = \left(1 - \frac{\sigma^2 \times \max(N-2, 0)}{\|\mathbf{y}\|_2^2}\right)\mathbf{y}\,, \tag{6.3.12}$$

that provides an improvement upon the maximum likelihood estimator of the mean of a Gaussian population whenever $N \geqslant 3$.

## 6.4 Rotational invariant prior estimators

The major drawback of the above conjugate prior class of estimator is that it does not make use of the enormous amount of information contained, for large $N$, in the observed spectral density of the sample correlation matrix $\mathbf{E}$. In fact, we know that its Stieltjes transform $\mathfrak{g}_{\mathbf{E}}(z)$ must obey the Marčenko-Pastur equation relating it to $\mathfrak{g}_{\mathbf{C}}(z)$, and there is no guarantee whatsoever

that this relation can be obeyed for any $\mathbf{C}$ belonging to an Inverse-Wishart ensemble. More precisely, the likelihood that $\mathfrak{g}_{\mathbf{E}}(z)$ indeed corresponds to a certain $\mathfrak{g}_{\mathbf{C}}(z)$ with $\mathbf{C}$ an Inverse-Wishart matrix is exponentially small in $N$, even for the optimal choice of the parameter $\kappa$. This is the peculiarity of the Bayesian approach in the large $N$ limit: the ensemble to which $\mathbf{C}$ belongs is in fact extremely strongly constrained by the Marčenko-Pastur relation. In this section and in the next chapter, we discuss how these constraints can be implemented in practice, allowing us to construct a truly consistent estimator of $\mathbf{C}$.

Let us consider a class of *rotationally invariant prior* distributions that belong to the Boltzmann class, Eq. (3.1.1), i.e. ,

$$\mathcal{P}(\mathbf{C}) \propto \exp[-N \operatorname{Tr} V_0(\mathbf{C})] \tag{6.4.1}$$

where $V_0$ denotes the potential function. Therefore, it is easy to see that $\mathbf{C} \overset{\text{law}}{=} \mathbf{\Omega C \Omega}^*$ for any $N \times N$ orthogonal matrix $\mathbf{\Omega} \in \mathbf{O}(N)$. In other words, the eigenbasis of $\mathbf{C}$ is not biased in any specific direction. Moreover, using the Gaussian likelihood function (6.2.2), the posterior distribution reads:

$$\mathcal{P}(\mathbf{C}|\mathbf{E}) = \frac{1}{Z} \exp\Big[-N \operatorname{Tr} \mathcal{V}(\mathbf{C}, \mathbf{E})\Big], \qquad \mathcal{V}(\mathbf{C}, \mathbf{E}) \;:=\; V_q(\mathbf{C}, \mathbf{E}) + V_0(\mathbf{C}), \tag{6.4.2}$$

where $V_q$ is defined in Eq. (6.3.3). As a result, one can derive the following identity:

$$\mathcal{P}(\mathbf{C}|\mathbf{E}) = \mathcal{P}(\mathbf{\Omega C \Omega}^*|\mathbf{\Omega E \Omega}^*), \tag{6.4.3}$$

Therefore, the Bayes MMSE estimator Eq. (6.1.9) obeys the following property:

$$\begin{aligned} \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})} &= \int \mathbf{\Omega C' \Omega}^* \mathcal{P}(\mathbf{\Omega C' \Omega}^*|\mathbf{E}) \mathcal{D}\mathbf{C}' \\ &= \mathbf{\Omega} \left[ \int \mathbf{C}' \mathcal{P}(\mathbf{C}'|\mathbf{\Omega}^* \mathbf{E \Omega}) \mathcal{D}\mathbf{C}' \right] \mathbf{\Omega}^* \equiv \mathbf{\Omega} \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{\Omega}^* \mathbf{E \Omega})} \mathbf{\Omega}^* \end{aligned} \tag{6.4.4}$$

where we changed variables $\mathbf{C} \to \mathbf{\Omega C' \Omega}^*$ and used Eq. (6.4.3) in the last step. Now we can always choose $\mathbf{\Omega} = \mathbf{U}$ such that $\mathbf{U}^* \mathbf{E U}$ is diagonal. In this case, it is not difficult to convince oneself using symmetry arguments that $\langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{U}^* \mathbf{E U})}$ is then also diagonal. The above result then simply means that in general, the MMSE estimator of $\mathbf{C}$ is diagonal in the same basis as $\mathbf{E}$ – see Takemura [166] and references therein:

$$\Xi^{\text{MMSE}} = \mathbf{U} \mathbf{\Gamma}(\Lambda) \mathbf{U}^*, \tag{6.4.5}$$

where $\mathbf{U} \in \mathbb{R}^{N \times N}$ is the eigenvectors of $\mathbf{E}$ and $\mathbf{\Gamma}(\mathbf{\Lambda}) = \operatorname{diag}(\gamma_1(\mathbf{\Lambda}), \ldots, \gamma_N(\mathbf{\Lambda}))$ is a $N \times N$ diagonal matrix whose entries are functions of the sample eigenvalues $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$. We see that assuming a rotationally invariant prior, the Bayesian estimation problem is reduced to finding a set of optimal eigenvalues $\gamma_i(\Lambda)$. This framework agrees perfectly with the linear shrinkage estimator (6.3.7), for which $\gamma_i(\Lambda) := \alpha_s \lambda_i + (1 - \alpha_s)$, and can be seen as a generalized shrinkage estimator.

Before going into details on the explicit form of the $\mathbf{\Gamma}(\mathbf{\Lambda})$, let us motivate when one may impose rotational invariance for the prior distribution of $\mathbf{C}$. In simple terms, it means that we have no prior information on a possible privileged directions in the N-dimensional space that would allow one to bias the eigenvectors of the estimator $\Xi^{\text{MMSE}}$ in these special directions. In this case, it makes sense that the only reasonable eigenbasis for our estimator $\Xi^{\text{MMSE}}$ must be that the (noisy) observation $\mathbf{E}$ at our disposal. Any estimator satisfying Eq. (6.4.4) will be referred to

as a Rotational Invariant Estimator (RIE). However, we emphasize that such an assumption is not optimal when the components of $\mathbf{E}$ reveal so non-trivial structures. One example is the top eigenvector of financial correlation matrices, which is clearly biased in the $(1, 1, \ldots, 1)$ direction. Dealing with such non-rotational invariant objects is however more difficult (see [43, 136] and Chapter 10 for a discussion on this topic).

We are now in a position to derive the explicit form of our optimal Bayes estimator within the class of RIEs. The eigendecomposition (6.4.5) of the estimator $\Xi^{\text{MMSE}}$ states that the eigenvalues of $\gamma_i \equiv \gamma_i(\Lambda)$ can be written as

$$\gamma_i = \langle \mathbf{u}_i, \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})} \mathbf{u}_i \rangle,$$

where we have used the fact that $\langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}$ is diagonal in the $\mathbf{U}$ basis. After a little thought, it is not hard to see that the following identity holds:

$$\frac{1}{N} \text{Tr} \left[ (z\mathbf{I}_N - \mathbf{E})^{-1} \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})} \right] = \frac{1}{N} \sum_{i=1}^{N} \frac{\gamma_i}{z - \lambda_i}, \tag{6.4.6}$$

which will allow us to extract the $\gamma_i$ we are looking for, i.e. determine the optimal shrinkage function of the Bayes estimator (6.4.5). To that end, we invoke the usual self-averaging property that holds for very large $N$, so that we can take the average value over the marginal probability of $\mathbf{E}$ in the LHS of the last equation, yielding:

$$\begin{aligned} \text{Tr} \left[ (z\mathbf{I}_N - \mathbf{E})^{-1} \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})} \right] &= \left\langle \text{Tr} \left[ (z\mathbf{I}_N - \mathbf{E})^{-1} \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})} \right] \right\rangle_{\mathcal{P}(\mathbf{E})}, \\ &= \left\langle \left\langle \text{Tr} \left[ (z\mathbf{I}_N - \mathbf{E})^{-1} \mathbf{C} \right] \right\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})} \right\rangle_{\mathcal{P}(\mathbf{E})}. \end{aligned} \tag{6.4.7}$$

Using Bayes formula (6.1.1), we rewrite this last equation as

$$\begin{aligned} \text{Tr} \left[ (z\mathbf{I}_N - \mathbf{E})^{-1} \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})} \right] &= \left\langle \left\langle \text{Tr} \left[ (z\mathbf{I}_N - \mathbf{E})^{-1} \mathbf{C} \right] \right\rangle_{\mathcal{P}(\mathbf{E}|\mathbf{C})} \right\rangle_{\mathcal{P}(\mathbf{C})}, \\ &= \left\langle \text{Tr} \left[ \langle (z\mathbf{I}_N - \mathbf{E})^{-1} \rangle_{\mathcal{P}(\mathbf{E}|\mathbf{C})} \mathbf{C} \right] \right\rangle_{\mathcal{P}(\mathbf{C})}. \end{aligned} \tag{6.4.8}$$

We recognize in the last line the definition of the Stieltjes transform of $\mathbf{E}$ for a given population matrix $\mathbf{C}$, which allows us to use the Marčenko-Pastur formalism introduced in Chapters 4 and 5. Therefore, since the eigenvalues $[\lambda_i]_i$ become deterministic in the limit $N \to \infty$ (see Chapter 4), we deduce that for large $N$

$$\frac{1}{N} \text{Tr} \left[ (z\mathbf{I}_N - \mathbf{E})^{-1} \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})} \right] \approx \int \frac{\rho_{\mathbf{E}}(\mathrm{d}\lambda)}{z - \lambda} \left\langle \sum_{j=1}^{N} \mu_j \, \Phi(\lambda, \mu_j) \right\rangle_{\mathbf{C}}, \tag{6.4.9}$$

where $\Phi(\lambda, \mu)$ is the mean squared overlap defined in Eq. (5.0.3). By comparing Eqs. (6.4.6) and (6.4.9), we can readily conclude that

$$\gamma(\Lambda) \equiv \gamma(\lambda) = \left\langle \sum_{j=1}^{N} \mu_j \, \Phi(\lambda, \mu_j) \right\rangle_{\mathbf{C}} \sim \int \mu \, \Phi(\lambda, \mu) \rho_{\mathbf{C}}(\mu) \mathrm{d}\mu, \tag{6.4.10}$$

where we used again an "ergodic hypothesis" [64] as $N \to \infty$ in the last step. Hence, we see that in the large $N$ limit, we are able to find a closed formula for the optimal shrinkage function $\gamma$ of the Bayes estimator (6.4.5) that depends on the mean squared overlap, studied in Chapter 5, and the prior spectral density $\rho_{\mathbf{C}}$. Said differently the final result Eq. (6.4.10) is explicit but still seems to depend on the prior we choose for $\mathbf{C}$. In fact, as we shall see in the next chapter, Eq. (6.4.10) can be estimated from the knowledge of $\mathbf{E}$ itself, i.e. without making an explicit choice for the prior! This is in line with our discussion at the beginning of this section: for large $N$, the observation of the spectral distribution of $\mathbf{E}$ is enough to determine the correct prior ensemble to which $\mathbf{C}$ must belong.

We end this section with a self-consistency check in order to illustrate the result (6.4.10). As alluded above, the nonlinear shrinkage function (6.4.10) generalizes the linear shrinkage (6.3.7) above. To highlight this, we assume that $\mathbf{C}$ is an isotropic Inverse Wishart matrices, such that the prior spectral density $\rho_{\mathbf{C}}$ is given by Eq. (3.1.53). We plot in Fig. 6.4.1 the eigenvalues we obtain using our Bayes estimator (6.3.7) (red dots) coming from a single realization of $\mathbf{E}$ with $\mathbf{C}$ an inverse Wishart matrix of size $N = 500$. The parameter of the prior distribution has been chosen such that the shrinkage intensity is equal to one half. We see that the agreement is excellent, showing the validity of the ergodic hypothesis claim and at the same time, of the RI-Bayes estimator (6.4.10) in this particular case.

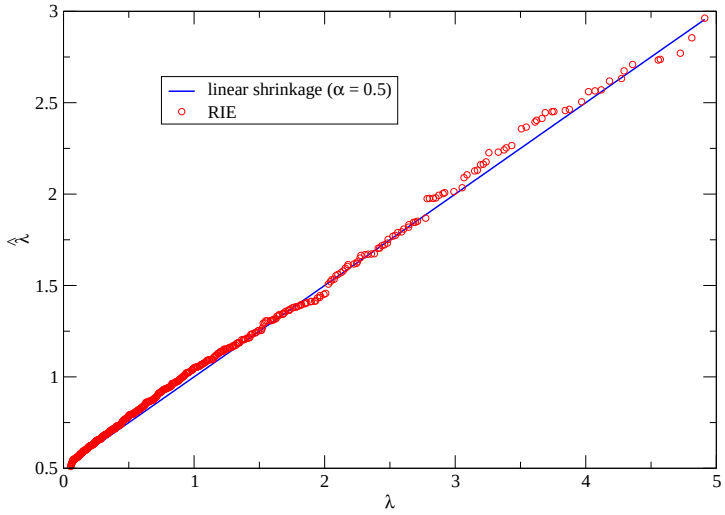



FIGURE 6.4.1. Comparison of our analytical RI-Bayes estimator (6.4.10) (red dots) with the theoretical result Eq. (6.3.7) (blue line) when the prior distribution is an inverse Wishart (3.1.58). The parameters are $N = 500$, $q = 0.5$ and $\alpha_s = 0.5$.

# Chapter 7

# Optimal rotational invariant estimator for general covariance matrices

## 7.1 Oracle estimator

In the previous chapter, we introduced a Bayesian framework to build an estimator of the population correlation matrix $\mathbf{C}$ using the data $\mathbf{Y}$ at our disposal. We showed that using a conjugate prior assumption naturally leads to the class of linear shrinkage estimators, which is arguably among the most influential contributions on this topic. It is successfully used in many contexts as it is a simple way to provide robustness against the noise in a high dimensional setting (see e.g. [87, 165] or [105] for a more recent review). However, the main concern regarding this estimator is that the conjugate prior ensemble is expected to be exponentially improbable (for large $N$) with the data at hand. In order to make full use of the information of the spectral density of the sample correlation matrix, we introduced a class of rotational invariant prior distributions. Within this framework, we have derived an explicit formula for the *minimum mean squared error* (MMSE) estimator valid in the limit of large dimension, which can be seen as a non-linear shrinkage procedure. In this chapter, we want to show that the resulting estimator can be also understood as a so-called "oracle" estimator. This change of viewpoint is quite interesting as it shows that the above Bayes estimator has a much wider basis than anticipated.

Imagine that one actually *knows* the population matrix $\mathbf{C}$ – hence the name "oracle" – but that one decides to create an estimator of $\mathbf{C}$ which is constrained to have a predetermined eigenbasis $\mathbf{U}$. (In practice, this eigenbasis will be that of the sample correlation matrix $\mathbf{E}$). What is the best one can do to estimate the true matrix $\mathbf{C}$? The basic idea might look strange at first sight, since we do not know $\mathbf{C}$ at all! But as we shall see below, the oracle estimator will turn out to coincide with the MMSE estimator which is, for large $N$, entirely expressable in terms of observable quantities. More precisely, let us introduce the set $\mathcal{M}(\mathbf{U})$ of real symmetric definite positive $N \times N$ matrices that are diagonal in the basis $\mathbf{U} = [\mathbf{u}_i]_{i \in qq1,N}$. The optimal estimator of $\mathbf{C}$ in $\mathcal{M}(\mathbf{U})$ in the $L_2$ sense is given by:

$$\Xi^{\text{ora.}} = \underset{\Xi \in \mathcal{M}(\mathbf{U})}{\operatorname{argmin}} \left\| \Xi - \mathbf{C} \right\|^2. \tag{7.1.1}$$

It is trivial to find that the solution of this quadratic optimization problem, as:

$$\Xi^{\text{ora.}} = \sum_{i=1}^{N} \xi_i^{\text{ora.}} \mathbf{u}_i \mathbf{u}_i^*, \qquad \xi_i^{\text{ora.}} = \langle \mathbf{u}_i \, , \mathbf{C}\mathbf{u}_i \rangle. \tag{7.1.2}$$

This provides the best possible estimator of $\mathbf{C}$ given that we are "stuck" with the eigenbasis $[\mathbf{u}_i]_i$. The meaning of this estimator is better understood if we rewrite if a function of the eigenvectors of $\mathbf{C}$, to wit:

$$\xi_i^{\text{ora.}} = \sum_{j=1}^{N} \mu_j \langle \mathbf{u}_i \, , \mathbf{v}_j \rangle^2. \tag{7.1.3}$$

Indeed, we see from this last equation that the oracle estimator is given by a weighted average of the population eigenvalues with weights given by the transition from the imposed basis $\mathbf{u}_i$ to the true basis $\mathbf{v}_j$ with $j \in [\![1, N]\!]$. Hence, the "oracle" estimator (7.1.2) explicitly uses the fact that the estimator lies in a wrong basis.

Coming back to our estimation problem given a sample matrix $\mathbf{E}$, it is clear that if we have no information whatsoever on the true eigenbasis of $\mathbf{C}$, the only possibility is to use the eigenbasis of $\mathbf{E}$ itself as $\mathbf{U}$. This is in some sense equivalent to the assumption of a rotationally invariant prior distribution for $\mathbf{C}$, but we do not rely on any Bayesian argument here. Now, one notices that in the limit $N \to \infty$, the oracle eigenvalues of $[\xi_i^{\text{ora.}}]$ are indeed equivalent to the RI-Bayes MMSE formula (6.4.10), except that in Eq. (7.1.2), the population matrix $\mathbf{C}$ is a (deterministic) general covariance matrix. The equivalence between Bayes estimator (6.4.10) and unconditional estimator is not that surprising in the large $N$ limit and has been mentioned in different contexts [66, 105].

## 7.2 Explicit form of the optimal RIE

For practical purposes, the oracle estimator (7.1.2) looks useless since it involves the matrix $\mathbf{C}$ which is exactly the quantity we wish to estimate. But in the high-dimensional limit a kind "miracle" happens in the sense that the oracle estimator converges to a deterministic RIE that does not involve the matrix $\mathbf{C}$ anymore. Let us derive this formula (that only contains observable quantities) first for bulk eigenvalues, then for outliers – with the further surprise that the final expression is exactly the same in the two cases.

**7.2.1. The bulk.** The derivation of the optimal nonlinear shrinkage function for the bulk eigenvalues in the limit of infinite dimension was considered in different recent works. The first one goes back to the work of Ledoit & Péché [113]. More recently, this oracle estimator was considered in a more general framework [40] (including the case of additive noise models) with the conclusion was that the oracle estimator can be easily computed as soon as the convergence of the mean squared overlap $\Phi(\lambda_i, \mu_j)$ defined in Eq. (5.0.3) can be established.

More precisely, let us fix $i \geqslant r+1$[1], we expect that in the limit of large dimension, the square overlaps $\langle \mathbf{u}_i \, , \mathbf{v}_j \rangle^2$ for any $j = 1, \ldots, N$ will display asymptotic independence so that the law of large number applies, leading to a deterministic result for $\xi_i^{\text{ora.}}$. Hence, we have for large $N$

[1] Recall that the top $r$ eigenvalues are assumed to be outliers.

that, for any $i > r$,

$$\xi_i^{\text{ora.}} = \sum_{j=1}^{N} \mu_j \, \Phi(\lambda_i, \mu_j) \approx \frac{1}{N\pi\rho_{\mathbf{E}}(\lambda_i)} \lim_{\eta\to 0^+} \operatorname{Im} \left[ \sum_{j=1}^{N} \mu_j \, (z_i \mathbf{I}_N - \mathbf{E})^{-1}_{jj} \right], \tag{7.2.1}$$

where we have used the result Eq. (5.1.5) with $z_i = \lambda_i - \mathrm{i}\eta$. One finds using the Marčenko-Pastur (4.2.3) and after simple algebraic manipulations that

$$\xi_i^{\text{ora.}} \sim \frac{1}{q\pi\rho_{\mathbf{E}}(\lambda_i)} \lim_{\eta\to 0^+} \operatorname{Im} \left[ 1 - \frac{1}{1 - q + q z_i \mathfrak{g}_{\mathbf{E}}(z_i)} \right],$$

which can be further simplified to the final Ledoit-Péché formula for the oracle estimators $[\xi_i^{\text{ora.}}]_{i\in[\![r,N]\!]}$:

$$\xi_i^{\text{ora.}} \sim \hat{\xi}(\lambda_i) \qquad \text{with} \qquad \hat{\xi}(\lambda) := \frac{\lambda}{\left|1 - q + q\lambda \lim_{\eta\to 0^+} \mathfrak{g}_{\mathbf{E}}(\lambda - \mathrm{i}\eta)\right|^2}, \tag{7.2.2}$$

where $|\cdot|$ denotes the complex modulus. We notice that the RHS of this last equation does not involve the matrix $\mathbf{C}$ any more and depends only on deterministic quantities. This is the "miracle" of the large $N$ limit we alluded above: the *a priori* non-observable oracle estimator converges to a deterministic quantity that may be estimated directly from the data.

**7.2.2. Outliers.** As usual, the arguments needed to derive the limiting value of the oracle estimator for outlier eigenvalues, i.e., $\xi_i^{\text{ora.}}$ for $i \leqslant r$, are a little bit different from those used above for bulk eigenvalues. Indeed, the latter explicitly needs the density of $\varrho_{\mathbf{E}}(\lambda_i)$ to be non-vanishing (for $N \to \infty$ and as we know from Chapter 4, this it not the case for outliers. Hence, the method of [113] and [40] are not valid anymore. Surprisingly, though, the final result happens to be identical to Eq. (7.2.2)! This has been established recently in [43] and the starting point of their method is to rewrite the oracle solution as

$$\xi_i^{\text{ora.}} = \sum_{j=1}^{r} \mu_j \langle \mathbf{v}_j \, , \mathbf{u}_i \rangle^2 + \sum_{j=r+1}^{N} \mu_j \langle \mathbf{v}_j \, , \mathbf{u}_i \rangle^2, \tag{7.2.3}$$

from which we conclude, using also the results of section 5 , that if $r$ is finite, both terms above will have a non-vanishing contribution for $i \leqslant r$. Roughly speaking, the first sum will contribute in $\mathcal{O}(1)$ for $j = i$ and the second sum gives a term of order $\mathcal{O}((N-r)/N) \sim \mathcal{O}(1)$.

We begin with the easy term which is the first one in the RHS of Eq. (7.2.3). Indeed, recall from Eq. (5.1.10) that any outlier eigenvector $\mathbf{u}_i$ is concentrated on a cone with its axis parallel to $\mathbf{v}_i$ and completely delocalized in any direction orthogonal $\mathbf{v}_j$ with $j \in [\![1, N]\!]$, $j \neq i$ fixed. Hence, the only term that contributes to leading order will be $\langle \mathbf{v}_i \, , \mathbf{u}_i \rangle^2$ and we therefore conclude that

$$\sum_{j=1}^{r} \mu_j \langle \mathbf{v}_j \, , \mathbf{u}_i \rangle^2 \sim \mu_i^2 \frac{\theta'(\mu_i)}{\theta(\mu_i)} \tag{7.2.4}$$

where we used Eq. (4.3.12) in the last step. The second term in Eq. (7.2.3) is trickier to handle. As $r$ is finite and thus much smaller than $N$, we can assume that the second sum will concentrate around its mean value, i.e.

$$\sum_{j=r+1}^{N} \mu_j \langle \mathbf{v}_j \, , \mathbf{u}_i \rangle^2 \sim \sum_{j=r+1}^{N} \mu_j \mathbb{E} \langle \mathbf{v}_j \, , \mathbf{u}_i \rangle^2.$$

The mean squared overlap in the RHS for $j \geqslant r+1$ and $i \leqslant r$ has been evaluated in section 5 and the result is given in Eq. (5.1.12) that we recall for convenience:

$$\mathbb{E}[\langle \mathbf{u}_i \, , \mathbf{v}_j \rangle^2] = \frac{\mu_i^2}{\theta(\mu_i)} \frac{\mu_j}{T(\mu_i - \mu_j)^2}, \qquad i \leqslant r, j \geqslant r+1.$$

Therefore we find for $r \ll N$ [43]

$$\sum_{j=r+1}^{N} \mu_j \langle \mathbf{v}_j \, , \mathbf{u}_i \rangle^2 \sim \frac{\mu_i^2}{\theta(\mu_i)} \frac{1}{T} \sum_{j=1}^{N} \frac{\mu_j^2}{(\mu_i - \mu_j)^2}, \tag{7.2.5}$$

where one notices that the sum of the RHS goes from $j = 1$ to $N$. We can simplify the sum in the RHS of this last equation by using the Marčenko-Pastur equation (4.2.27). Indeed, by setting $z = \theta(\mu_i)$ with $i \leqslant r$ and $\theta$ defined in Eq. (4.3.12), Eq. (4.2.27), becomes

$$\theta(\mu_i) = \mu_i + \frac{1}{T} \sum_{j=1}^{N} \frac{1}{\mu_j^{-1} - \mu_i^{-1}} \tag{7.2.6}$$

and by taking the derivative with respect to $\mu_i$, this yields

$$\frac{1}{T} \sum_{j=1}^{N} \frac{\mu_j^2}{(\mu_i - \mu_j)^2} = 1 - \theta'(\mu_i), \tag{7.2.7}$$

for any $i \leqslant r$. By plugging this identity into Eq. (7.2.5), we then obtain

$$\sum_{j=r+1}^{N} \mu_j \langle \mathbf{v}_j \, , \mathbf{u}_i \rangle^2 \sim \frac{\mu_i^2}{\theta(\mu_i)} \big(1 - \theta'(\mu_i)\big), \tag{7.2.8}$$

for any $i \leqslant r$. All in all, we see by plugging Eqs. (7.2.4) and (7.2.8) into Eq. (7.2.3) that we finally get

$$\xi_i^{\text{ora.}} \sim \frac{\mu_i^2}{\theta(\mu_i)}, \tag{7.2.9}$$

i.e. the oracle estimator for outliers also converge to a deterministic value which is very simple, but depends on the population eigenvalues which are not observable. However, using Eq. (4.3.12), we can rewrite the RHS of Eq. (7.2.9) as a function of the sample eigenvalues. Firstly, one notices that $\theta(\mu_i) = \lambda_i$ for $N \to \infty$ thanks to Eq. (4.3.12). Moreover, we can also invert Eq. (4.3.12) to find

$$\mu_i \sim \frac{1}{\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_i)} = \frac{\lambda_i}{1 - q + q\lambda_i \mathfrak{g}_{\underline{\mathbf{E}}}(\lambda_i)},$$

for any $i \leqslant r$ and where we use relation Eq. (4.2.25) in the last step. Therefore, we deduce that in the high dimensional limit, we can rewrite Eq. (7.2.9) as

$$\xi_i^{\text{ora.}} \sim \frac{\lambda_i}{\big|1 - q + q\lambda_i \mathfrak{g}_{\underline{\mathbf{E}}}(\lambda_i)\big|^2}. \tag{7.2.10}$$

We see that the result is similar to the result for the bulk eigenvalues except that for outliers, we need the Stietjes transform of the spikeless, fictitious sample covariance matrix $\underline{\mathbf{E}}$. But as we

consider the limit $N \to \infty$, we easily deduce using Weyl's interlacing inequalities [187] that we can replace it by the Stieltjes transform of $\mathbf{E}$ so that we finally conclude that for any $i \leqslant r$,

$$\xi_i^{\text{ora.}} \sim \hat{\xi}(\lambda_i)\,, \tag{7.2.11}$$

where the optimal shrinkage function $\hat{\xi}$ is defined in (7.2.2) and we see that the outliers oracle estimator converge to a deterministic function which is exactly the same than for bulk eigenvalues (7.2.2) in the large $N \to \infty$.

To conclude, we found that the oracle estimator converges to a limiting function that does not explicitly require the knowledge of $\mathbf{C}$ and is identical to the Bayes-MMSE estimator obtained in the previous Chapter. Moreover, this function is universal in the sense that the optimal non linear shrinkage needed to clean bulk eigenvalues and outliers is given by the very same function in the limit $N \to \infty$, which is very appealing for practical applications. This function is defined in Eqs. (7.2.2) or (7.2.11) and only requires the knowledge of the Stieljes transform of $\mathbf{E}$, which is observable – see below.

## 7.3 Some properties of the "cleaned" eigenvalues

Even though the optimal nonlinear shrinkage function (7.5.2) seems relatively simple, it is not immediately clear what is the effect induced by the transformation $\lambda_i \to \xi^{\text{ora.}}(\lambda_i)$. In this section, we thus give some quantitative properties of the optimal estimator $\Xi^{\text{ora.}}$ to understand the impact of the optimal nonlinear shrinkage function $\hat{\xi}(\lambda)$.

First let us the consider the moments of the spectrum of $\Xi^{\text{ora.}}$. From Eq. (7.1.3) we immediately derive that:

$$\text{Tr}\Xi^{\text{ora.}} = \sum_{j=1} \mu_j \mathbf{v}_j^* \left( \sum_{i=1} \mathbf{u}_i \mathbf{u}_i^* \right) \mathbf{v}_j = \text{Tr}\mathbf{C}, \tag{7.3.1}$$

meaning that the cleaning operation preserves the trace of the population matrix $\mathbf{C}$, as it should be. For the moment of order 2 of the oracle estimator, we have:

$$\text{Tr}(\Xi^{\text{ora.}})^2 = \sum_{j,k=1}^{N} \mu_j \mu_k \sum_{i=1} \langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 \langle \mathbf{u}_i, \mathbf{v}_k \rangle^2.$$

Now, ff we define the matrix $\mathbf{P}$ as $\{\sum_{i=1} \langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 \langle \mathbf{u}_i, \mathbf{v}_k \rangle^2\}$ for $j, k = 1, N$, it is not hard to see that it is a squared matrix with nonnegative entries and whose rows all sum to unity. The matrix $\mathbf{P}$ is therefore a (bi)stochastic matrix and the Perron-Frobenius theorem tells us that its largest eigenvalues is equal to unity. Hence, we deduce the following general inequality

$$\sum_{j,k=1}^{N} P_{j,k} \mu_j \mu_k \leq \sum_{j=1}^{N} \mu_j^2,$$

which implies that

$$\text{Tr}(\Xi^{\text{ora.}})^2 \leqslant \text{Tr}\mathbf{C}^2 \leqslant \text{Tr}\mathbf{E}^2, \tag{7.3.2}$$

where the last inequality comes from Eq. (4.2.9). In words, this result states that the spectrum of $\Xi^{\text{ora.}}$ is narrower than the spectrum of $\mathbf{C}$, which is itself narrower than the spectrum of $\mathbf{E}$. The optimal RIE therefore tells us that we better be even more "cautious" than simply bringing back

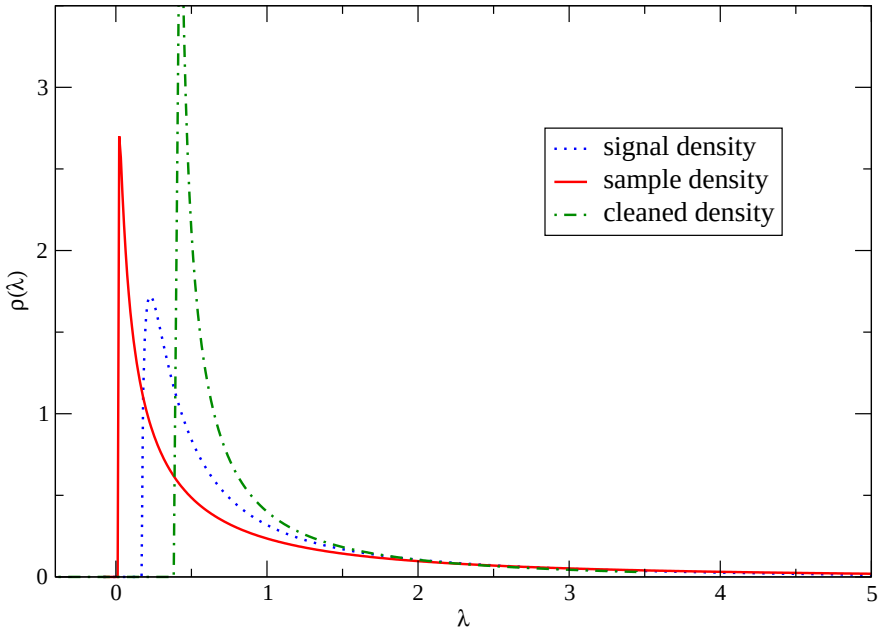



FIGURE 7.3.1. Evaluation of the eigenvalues density of state of the signal, sample and cleaned density when the prior is an inverse Wishart of parameter $\kappa = 1$. We see that the cleaned density is the narrowest one, while the sample is the largest as expected.

the sample eigenvalues to their estimated "true" locations. This is because we have only partial information about the true eigenbasis of $\mathbf{C}$. In particular, one should always shrink downward (resp. upward) the small (resp. top) eigenvalues compared to their "true" locations $\mu_i$ for any $i \in [\![1, N]\!]$, except for the trivial case $\mathbf{C} = \mathbf{I}_N$. As a consequence, estimating the population eigenvalues $[\mu_i]$ is *not* what one should do to obtain an optimal estimator of $\mathbf{C}$ when there is only partial information about its eigenvectors. We provide an illustration in Figure 7.3.1 where we consider $\mathbf{C}$ to be an inverse-Wishart matrix with parameter $\kappa = 1$.

Next, we consider the asymptotic behavior of the oracle estimator for which we recall from Eqs. (7.2.2) and (7.2.11) that

$$\xi_i^{\text{ora.}} \sim \hat{\xi}_i\,, \qquad \text{with} \qquad \hat{\xi}_i \;:=\; \frac{\lambda_i}{|1 - q + q\lambda_i \lim_{\eta \downarrow 0} \mathfrak{g}_{\mathbf{E}}(\lambda_i - \mathrm{i}\eta)|^2}\,.$$

Throughout the following, suppose that we have an outlier at the left of the lower bound of $\operatorname{supp} \rho_{\mathbf{E}}$ and let us assume $q < 1$ so that $\mathbf{E}$ has no exact zero mode[2]. We know since Section 7.2.2 that the estimator (7.2.2) holds for outliers. Moreover, we have that $\lim_{\lambda \to 0^+} \mathfrak{g}_{\mathbf{E}}(\lambda)$ is real and analytic so that we have from Eq. (4.2.15) that $\lambda \mathfrak{g}_{\mathbf{E}}(\lambda) = \mathcal{O}(\lambda)$ for $\lambda \to 0^+$. This allows to conclude from Eq. (7.2.2) that for very small outlier,

$$\lim_{\lambda \to 0^+} \hat{\xi}(\lambda) = \frac{\lambda}{(1-q)^2} + \mathcal{O}(\lambda^2), \tag{7.3.3}$$

which is in agreement with Eq. (7.3.2): small eigenvalues are enhanced for $q \in (0, 1)$.

[2]Recall that we assume $\mathbf{C}$ to be positive definite for the sake of simplicity.

The other asymptotic limit $\lambda \to \infty$ is also useful since it gives us the behavior of the non-linear shrinkage function $\hat{\xi}$ for large outliers. In that case, we know from Eq. (4.2.8) that $\lim_{\lambda\uparrow\infty} \lambda \mathfrak{g}_{\mathbf{E}}(\lambda) \sim 1 + \lambda^{-1}\varphi(\mathbf{E})$, where $\varphi$ denotes the normalized trace operator (3.1.61). Therefore, we conclude that

$$\lim_{\lambda\to\infty} \hat{\xi}(\lambda) \approx \frac{\lambda}{\Big(1 + q\lambda^{-1}\varphi(\mathbf{E}) + \mathcal{O}(\lambda^{-2})\Big)^2} \sim \lambda - 2q\varphi(\mathbf{E}) + \mathcal{O}(\lambda^{-1}), \tag{7.3.4}$$

and if we use that $\operatorname{Tr}\mathbf{E} = \operatorname{Tr}\mathbf{C} = N$, then we simply obtain

$$\lim_{\lambda\to\infty} \hat{\xi}(\lambda) \approx \lambda - 2q + \mathcal{O}(\lambda^{-1}). \tag{7.3.5}$$

It is interesting to compare this with the well-known "Baik-Ben Arous-Péché" (BBP) result on large outliers [15], which reads (see Eq. (4.3.14)) $\lambda \approx \mu + q$ for $\lambda \to \infty$. As a result, we deduce from Eq. (7.3.5) that $\hat{\xi}(\lambda) \approx \mu - q$ and we therefore find the following ordering relation

$$\hat{\xi}(\lambda) < \mu < \lambda, \tag{7.3.6}$$

for an isolated and large eigenvalues $\lambda$ and for $q > 0$. Again, this result is in agreement with Eq. (7.3.2): large eigenvalues should be reduced downward for any $q > 0$, even below the "true" value of the outlier $\mu$. More generally, the non-linear shrinkage function $\hat{\xi}$ interpolates smoothly between $\lambda/(1-q)^2$ for small $\lambda$'s to $\lambda - 2q$ for large $\lambda$'s. Even though we did not manage to prove it, we believe that this is another manifestation of the fact that the limiting optimal nonlinear shrinkage function (7.2.2) is mononotic with respect to the sample eigenvalues.

## 7.4 Some analytical examples

The above general properties of the oracle shrinkage procedure can be given more precise flesh in some exactly solvable cases. In this section we provide two simple toy models where the function $\hat{\xi}(\lambda)$ can be characterized explicitly, before turning to numerical illustrations.

**7.4.1. Null Hypothesis.** The first one is the null hypothesis $\mathbf{C} = \mathbf{I}_N$ where we shall see that, as expected $\xi^{\text{ora.}}(\lambda_i) = 1$ for any eigenvalues $[\lambda_i]_{i\geqslant r+1}$ in the bulk of the distribution. Outside of the spectrum, we observe a "phase transition" phenomena similar to the BBP transition [15].

We begin with the outliers of $\mathbf{E}$. By assumption of our model, all the outliers have a contribution of order $N^{-1}$ so that in the limit $N \to \infty$, $\mathfrak{g}_{\mathbf{E}}$ is real and analytic for any $\lambda_i$ with $i \leqslant r$. Hence, the estimator is easily obtained by plugging the Stieltjes transform (3.1.41) into Eq. (7.2.2), with a result shown in Fig. 7.4.1. For bulk eigenvalues, the computation can be done more explicitly. First, using Eq. (3.1.41), one finds

$$1 - q + qz\mathfrak{g}_{\mathbf{E}}(z) = \frac{(z+1-q) \pm \sqrt{(z+q-1)^2 - 4zq}}{2}.$$

For $z = \lambda - \mathrm{i}\eta$ with $\lambda \in \big[(1-\sqrt{q})^2, (1+\sqrt{q})^2\big]$, we know that the square root in the latter equation becomes imaginary for $\eta \to 0^+$. Hence, if we take the square modulus, one gets

$$\lim_{\eta\to 0} \big|1 - q + q\lambda\mathfrak{g}_{\mathbf{E}}(\lambda - \mathrm{i}\eta)\big|^2 = \frac{(z+1-q)^2 + \big(4\lambda q - (\lambda + q - 1)^2\big)}{4},$$

from which we readily find

$$\lim_{\eta\to 0}\left|1-q+q\lambda \mathfrak{g}_{\mathbf{E}}(\lambda-\mathrm{i}\eta)\right|^2=\lambda,$$

and this gives the expected answer

$$\hat{\xi}(\lambda)=1, \qquad \lambda\in\left[(1-\sqrt{q})^2,(1+\sqrt{q})^2\right]. \tag{7.4.1}$$

We provide an illustration of this phase transition in Figure 7.4.1 in the case where $\mathbf{C}=\mathbf{I}_N$, corresponding to a matrix $\mathbf{E}$ is generated using an isotropic Wishart matrix with $q=0.5$. It also confirms the asymptotic prediction for large and isolated eigenvalue Eq. (7.3.5).

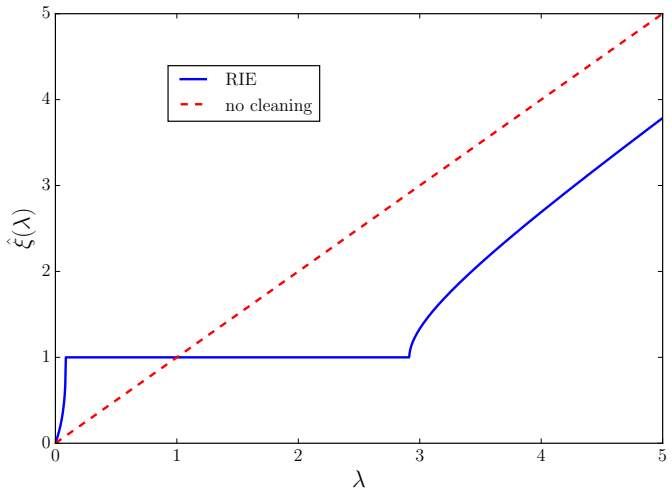



FIGURE 7.4.1. Evaluation of the optimal RIE's eigenvalues for $\mathbf{C}=I_N$ as a function of the sample eigenvalues $[\lambda_i]_{i\in[\![1,N]\!]}$. The nonlinear shrinkage function is plotted with the plain blue line. We see that for $\lambda>(1+\sqrt{q})^2$, a phase transition occurs and the corresponding "cleaned" eigenvalues converges to $\lambda-2q$ as $\lambda$ grows (see red dotted line). We also have a phase transition for any outlier $\lambda<(1-\sqrt{q})^2$ (see Figure 7.4.3).

**7.4.2. Revisiting the linear shrinkage.** In Chapter 6, we saw that the linear shrinkage (towards the identity matrix) is equivalent to assuming that $\mathbf{C}$ itself belongs to an Inverse-Wishart ensemble with some parameter $\kappa$. We want to revisit this result within the framework of the present chapter, and we will see that in the presence of extra spikes, the optimal shrinkage function (7.2.2) again shows a phase transition phenomenon and therefore differs from the linear estimator Eq. (6.3.7) for eigenvalues lying outside the spectrum of $\mathbf{E}$.

As for the null hypothesis case above, there is no particular simplifications for outliers and the result is immediately obtained from Eq. (7.2.2) and (4.2.33). For the bulk component, the square root term in Eq. (4.2.33) becomes imaginary. Hence, setting $z=\lambda-\mathrm{i}\eta$ into Eq. (4.2.33) with $\lambda\in[\lambda_-^{\mathrm{iw}},\lambda_+^{\mathrm{iw}}]$ and $\lambda_\pm^{\mathrm{iw}}$, defined in Eq. (4.2.34), one obtains

$$\left|1-q+q\lambda\lim_{\eta\to 0^+}\mathfrak{g}_{\mathbf{E}}(\lambda-\mathrm{i}\eta)\right|^2=\frac{\left[\lambda_i(1+q\kappa)+\kappa q(1-q)\right]^2+q^2\left[2\lambda_i\kappa(\kappa(1+q)+1)-\kappa^2(1-q)^2-\lambda_i^2\kappa^2\right]}{(\lambda_i+2q\kappa)^2},$$

with $\kappa > 0$. This can be rewritten after expanding the square as

$$\left| 1 - q + q\lambda \lim_{\eta \to 0^+} \mathfrak{g}_{\mathbf{E}}(\lambda - \mathrm{i}\eta) \right|^2 = \frac{\lambda(1+2q\kappa)}{(\lambda + 2q\kappa)}. \tag{7.4.2}$$

By plugging this last equation into Eq. (7.2.2) gives for any $\lambda \in [\lambda_-^{\mathrm{iw}}, \lambda_+^{\mathrm{iw}}]$

$$\xi^{\mathrm{ora.}}(\lambda) = \frac{\lambda_i + 2q\kappa}{1 + 2q\kappa}, \tag{7.4.3}$$

and if we recall the definition $\alpha_s = 1/(1+2q\kappa) \in [0,1]$ of Eq. (6.3.7), we retrieve exactly the linear shrinkage estimator (6.3.7),

$$\xi^{\mathrm{ora.}}(\lambda) \sim \alpha_s \lambda + (1 - \alpha_s), \qquad \lambda \in [\lambda_-^{\mathrm{iw}}, \lambda_+^{\mathrm{iw}}]. \tag{7.4.4}$$

This last result illustrates in a particular case the genuine link between the optimal RIE $\Xi^{\mathrm{ora.}}$ and Bayes optimal inference techniques. In particular, we show that for an isotropic Inverse Wishart matrix, the estimator $\Xi^{\mathrm{ora.}}$ gives the same result than the conjugate prior approach in the high dimensional regime. Nevertheless, this is valid *only for the bulk component* as the presence of outliers induces a phase transition for the optimal RIE, which is absent within the conjugate prior theory that is blind to outliers. We illustrate this last remark in Figure 7.4.2 where $\mathbf{C}$ is an Inverse-Wishart matrix of parameter $\kappa = 2$. The link between Bayesian statistics and RIE in the high-dimensional regime has been noticed in [40] where the case of an additive noise is also considered – see Appendix 11, yielding a generalisation of the well-known Wiener's signal-to-noise ratio optimal estimator [188].

We also illustrate in Figure 7.4.3 the phase transition observed for outliers at the left of the lower bound of the spectrum for both analytical examples. We see that for very small eigenvalues, the theoretical prediction (7.3.3) is pretty accurate. This prediction becomes less and less effective as $\lambda$ moves closer to the left edge.

## 7.5 Optimal RIE at work

In order to conclude this section, we now consider different cases where $\mathfrak{g}_{\mathbf{E}}(z)$ is not explicit, and where the problem is solved numerically. In that case, the main question is to estimate the function $\mathfrak{g}_{\mathbf{E}}(z)$ without imposing any "prior" on $\mathbf{C}$. Indeed, even though the function $\xi^{\mathrm{ora.}}$ only depends on observables quantities, we still need to estimate the fonction $\mathfrak{g}_{\mathbf{E}}(z)$ using only a finite (and random) set of sample eigenvalues.

This question has been addressed recently in [43], where apart from extending the result of [113] to outliers (as reviewed above), the mathematical technique used provided a derivation of Eq. (7.2.2) at a *local* scale and for any large but bounded $N$. As alluded in Chapter 5, the local scale can be understood as an average over small intervals of eigenvalues of width $\eta = \mathrm{d}\lambda \geqslant N^{-1}$. The main result of [43] can be summarized as follows: the limiting Stieltjes transform $\mathfrak{g}_{\mathbf{E}}(z)$ can be replaced by its discrete form

$$\mathfrak{g}_{\mathbf{E}}^N(z) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{z - \lambda_i}, \tag{7.5.1}$$

with *high probability* (see e.g. [109] for the exact statement). Therefore, this yields a fully observable nonlinear shrinkage function and moreover, the choice $\eta = N^{-1/2}$ gives a sharp upper

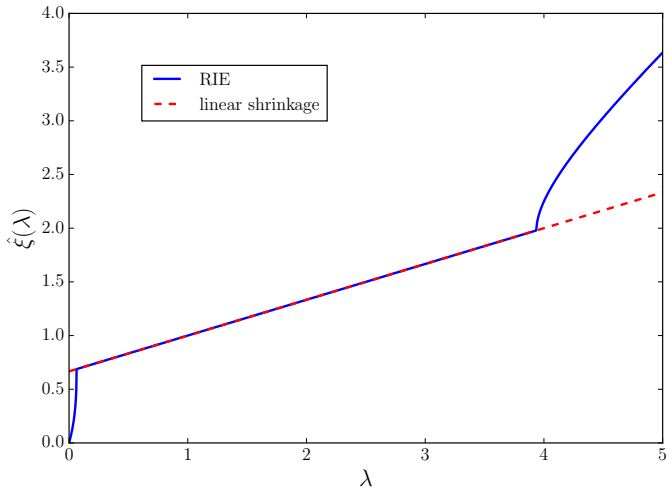



FIGURE 7.4.2. Evaluation of the optimal RIE's eigenvalues for an Inverse Wishart prior with $\kappa = 2$ as a function of the sample eigenvalues $[\lambda_i]_{i\in[\![1,N]\!]}$. The matrix $\mathbf{E}$ is generated using Wishart matrix with parameter $N = 500$ and $q = 0.5$. The nonlinear shrinkage function is plotted with the plain blue line and it coincides with the estimator Eq.(6.3.7) (red dotted line). We nonetheless see that for $\lambda > \lambda_+^{\text{iw}}$, a phase transition occurs and the two estimators split up. The same phenomenon is observed for $\lambda < \lambda_+^{\text{iw}}$ (see Figure 7.4.3).

error bound for any finite $N$ and $T$. Precisely, for $z_i = \lambda_i - \mathrm{i}N^{-1/2}$, there exists a constant $K$ such that for large enough $T$,

$$\left|\xi_i^{\text{ora.}} - \hat{\xi}_i^N\right| \leqslant \frac{K}{\sqrt{T}}, \qquad \hat{\xi}_i^N \equiv \hat{\xi}^N(\lambda_i) := \frac{\lambda_i}{\left|1 - q + qz_i\mathfrak{g}_{\mathbf{E}}^N(z_i)\right|^2}, \tag{7.5.2}$$

provided that $\lambda_i$ is not near zero [43]. We see that Eq. (7.5.2) is extremely simple to implement numerically as it only requires to compute a sum over $N$ terms.

We now test numerically the accuracy of the finite $N$, observable optimal nonlinear shrinkage function (7.5.2) in four different settings for the population matrix $\mathbf{C}$. We choose $N = 500$, $T = 1000$ (which are quite reasonable numbers in real cases, not too small nor too large) and consider the following four different cases:

(i) Diagonal matrix whose ESD is composed of multiple sources with "spikes" located at {8, 15},

$$\rho_{\mathbf{C}} = 0.002\delta_{15} + 0.002\delta_8 + 0.396\delta_3 + 0.3\delta_{1.5} + 0.3\delta_1. \tag{7.5.3}$$

(ii) Deformed GOE, i.e $\mathbf{C} = I_N + \text{GOE}$ (of width $\sigma = 0.2$) with extra spikes located at {3, 3.5, 4.5, 6}.

(iii) Toeplitz matrix with entries $\mathbf{C}_{ij} = 0.6^{|i-j|}$ with spikes located at {7, 8, 10, 11};

(iv) Power-law distributed eigenvalues (see [29] and Chapter 4) with $\lambda_0 = -0.6$ (or $\lambda_{\min} = 0.8$. Using a large $N$ proxy for the classical positions of the $\mu_i$, one gets [29]:

$$\mu_i = -\lambda_0 + \frac{(1+\lambda_0)}{2}\sqrt{\frac{N}{i}} \qquad i \in [\![1, N]\!]\,. \tag{7.5.4}$$

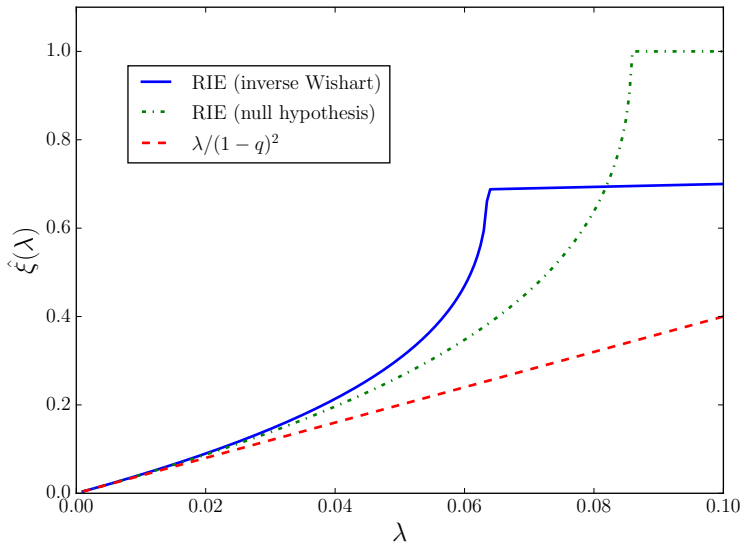



FIGURE 7.4.3. Comparison of the prediction Eq. (7.3.3) (red dashed line) compared to the analytical solution of the null hypothesis (7.4.1) (green dash-dotted line) and the Inverse Wishart prior (7.4.4) with parameter $\kappa = 2$ (blue plain line). In both cases, wet set $q = 0.5$. The asymptotic prediction (7.3.3) becomes less and less accurate as $\lambda$ moves closer to the left edge and the analytic solution (blue line) depicts a phase transition.

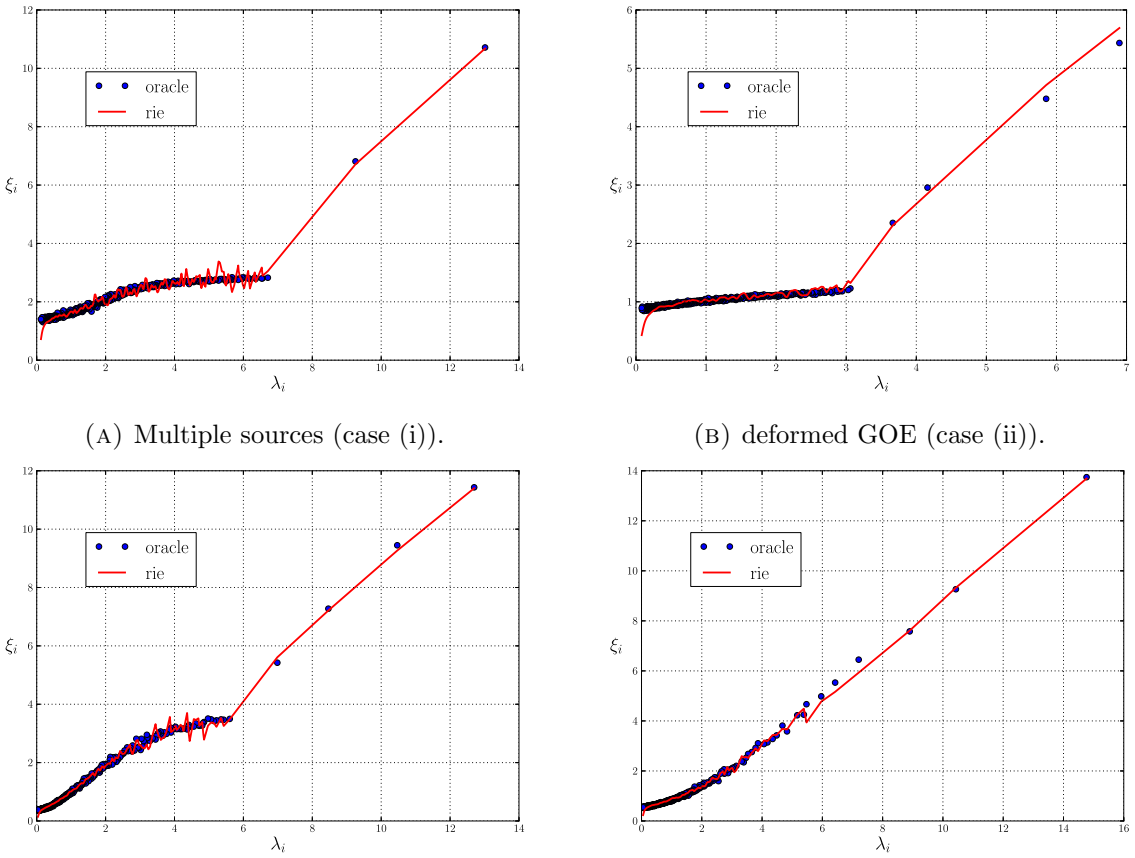
Note that the last power law distribution automatically generates a bounded number of outliers. Moreover, since we work with $N$ and $T$ bounded, the largest eigenvalue of $\mathbf{C}$ is remains bounded. We plot the results obtained with the estimator Eq. (7.5.2) and the oracle estimator Eq. (7.1.2) in Figure 7.5.1.

Overall, the estimator (7.5.2) gives accurate predictions for both the bulk eigenvalues and outliers. We have considered several configurations of outliers. For the case (i), we see that the two isolated outliers are correctly estimated. For the deformed GOE or the Toeplitz case, the outliers are chosen to be a little bit closer to one another and again, the results agree well with the oracle estimator. For the more complex case of a power law distributed spectrum, where there is no sharp right edge, we see that (7.5.2) matches again well with the oracle estimator. We nevertheless notice that the small eigenvalues are *systematically* underestimated by the empirical optimal RIE (7.5.2). This effect will be investigated in more details in Chapter 9.

## 7.6 Extension to the free multiplicative model

As highlighted in [40], the evaluation of optimal RIE for bulk eigenvalues can be extended to more general multiplicative random matrix models (for additive noise models, see Appendix 11). In particular, it is possible to derive (formally) the analog of optimal nonlinear shrinkage function (7.2.2) for the bulk eigenvalues of the measurement model (3.1.80) which encompasses the sample covariance matrix (see Section 4.2.1).

To that end, let us define $\mathbf{M} := \mathbf{C}^{1/2}\mathbf{\Omega}\mathbf{B}\mathbf{\Omega}^*\mathbf{C}^{1/2}$ where $\mathbf{B}$ is a $N \times N$ symmetric rotational invariant noise term and $\mathbf{\Omega}$ is a $N \times N$ random rotation matrix that is distributed according to the Haar




(A) Multiple sources (case (i)).

(B) deformed GOE (case (ii)).

(C) Toeplitz (case (iii))

(D) Power law (case (iv))

FIGURE 7.5.1. Comparison of numerically estimated oracle estimator (7.5.2) (red line) with the exact oracle RIE estimator (7.1.2) (blue points) for the four cases presented at the beginning of Section 7.5 with $N = 500$ and $T = 1000$. The results come from a single realization of $\mathbf{E}$ using a multivariate Gaussian measurement process.

measure. One can easily check from Eq. (3.1.100) that

$$\mathrm{Tr}\left[\mathbf{G}_{\mathbf{M}}(z)\mathbf{C}\right] = N(z\mathfrak{g}_{\mathbf{M}}(z) - 1)\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1)\,. \tag{7.6.1}$$

Using the analyticity of the $\mathcal{S}$-transform, we define the function $\gamma_{\mathbf{B}}$ and $\omega_{\mathbf{B}}$ such that:

$$\lim_{z\to\lambda-\mathrm{i}0^+} \mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1) := \gamma_{\mathbf{B}}(\lambda) + \mathrm{i}\pi\rho_{\mathbf{M}}(\lambda)\omega_{\mathbf{B}}(\lambda)\,, \tag{7.6.2}$$

and as a result, the optimal RIE for bulk eigenvalues of the free multiplicative noise model (3.1.80) may be inferred from (7.2.1):

$$\xi_i^{\text{ora.}} \sim F_2(\lambda_i); \qquad F_2(\lambda) = \lambda\gamma_{\mathbf{B}}(\lambda) + (\lambda\mathfrak{h}_{\mathbf{M}}(\lambda) - 1)\omega_{\mathbf{B}}(\lambda)\,. \tag{7.6.3}$$

Note that one retrieves the estimator (7.2.2) by plugging Eqs. (3.1.44) and (7.6.2) into Eq. (7.6.3). We omit details, which can be found in [40], and we conclude that the formula (7.6.3) indeed generalizes Eq. (7.2.2). Again, we see that the final solution does not depend explicitly on $\mathbf{C}$ but somehow requires a prior on the spectral distribution of the matrix $\mathbf{B}$. It would be quite satisfying to find models in which we may obtain an explicit formula for Eq. (7.6.3) (see Chapter 10 for some relevant applications of this model).

We emphasize in passing that we may also derive the mean squared overlap (5.0.3) in the bulk of the distribution using Eq. (3.1.100). To that end, we invoke the relation (5.1.5) and Eq. (3.1.100) to obtain [40]:

$$\Phi(\lambda,\mu) = \frac{\mu\beta_m(\lambda)}{(\lambda-\mu\alpha_m(\lambda))^2 + \pi^2\mu^2\beta_m(\lambda)^2\rho_{\mathbf{M}}(\lambda)^2}\,, \tag{7.6.4}$$

where we defined the functions $\alpha_m$ and $\beta_m$ as

$$\begin{cases} \alpha_m(\lambda) := \lim\limits_{z\to\lambda-i0^+} \mathrm{Re}\left[\dfrac{1}{\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z)-1)}\right] \\ \beta_m(\lambda) := \lim\limits_{z\to\lambda-i0^+} \mathrm{Im}\left[\dfrac{1}{\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z)-1)}\right] \dfrac{1}{\pi\rho_{\mathbf{M}}(\lambda)}\,, \end{cases} \tag{7.6.5}$$

and the subscript $m$ stands for "multiplication".

We conclude this technical section by mentioning one important open problem which is the extension of these results in the presence of outliers. Indeed, it would be interesting to see whether the optimal RIE formula (7.6.3) remains *universal* in the sense that the cleaning formula for bulk eigenvalues and outliers is identical. The block matrix representation (B.4.8) might be useful in that respect.

# Chapter 8

# Application: Markowitz portfolio theory and previous "cleaning" schemes

## 8.1 Markowitz optimal portfolio theory

For the reader not familiar with Markowitz's optimal portfolio theory [125], we recall in this section some of the most important results. Suppose that an investor wants to invest in a portfolio containing $N$ different assets, with optimal "weights" to be determined. An intuitive strategy is the so-called mean-variance optimization: the investor seeks an allocation such that the overall quadratic risk of the portfolio is minimized given an expected return target. It is not hard to see that this mean-variance optimization can be translated into a simple quadratic optimization program with a linear constraint. Before going into more mathematical details, let us introduce some notations that will be used in the following. We suppose that we observe the return time series of $N$ different stocks. For each stock, we observe a time series of size $T$, where $T$ is often larger than $N$ in practice. This yields the (normalized) $N \times T$ return matrix $\mathbf{Y} = (Y_{it}) \in \mathbb{R}^{N \times T}$ whose true correlation matrix is defined by

$$\langle Y_{it} Y_{jt'} \rangle = \mathbf{C}_{ij} \delta_{t,t'}, \tag{8.1.1}$$

where the absence of correlations in the time direction is a only a first approximation since weak, but persistent linear correlations are known to exist in stock markets.

As natural in the present "Big Data" era, we place ourselves in the high-dimensional regime $N, T \to \infty$ with a finite ratio $q = N/T$. Markowitz's optimal portfolio amounts to solving the following quadratic optimization problem

$$\begin{cases} \min_{\mathbf{w} \in \mathbb{R}^N} \frac{1}{2} \mathbf{w}^* \mathbf{C} \mathbf{w} \\ \text{s.t. } \mathbf{w}^* \mathbf{g} \geq \mathcal{G} \end{cases} \tag{8.1.2}$$

where $\mathbf{g}$ is a $N$-dimensional vector of prediction and $\mathcal{G}$ is the expected gain. This can be easily solved by introducing a Lagrangian multiplier $\gamma$ to rewrite this constrained optimization problem as a unconstrained one[1]:

$$\min_{\mathbf{w} \in \mathbb{R}^N} \left( \frac{1}{2} \right) \mathbf{w}^* \mathbf{C} \mathbf{w} - \gamma \mathbf{w}^* \mathbf{g}. \tag{8.1.3}$$

[1] One can check that the so-called Karush-Kuhn-Tucker conditions are satisfied.

Assuming that $\mathbf{C}$ is invertible, it is not hard to find the optimal solution and the value of $\gamma$ such that overall expected return is exactly $\mathcal{G}$. It is given by

$$\mathbf{w}_{\mathbf{C}} = \mathcal{G}\frac{\mathbf{C}^{-1}\mathbf{g}}{\mathbf{g}^*\mathbf{C}^{-1}\mathbf{g}}, \tag{8.1.4}$$

that requires the knowledge of both $\mathbf{C}$ and $\mathbf{g}$, which are *a priori* unknown. However, forming expectations of future returns is the job of the investor, based on his/her informations and anticipations, so we assume that $\mathbf{g}$ is known. Even if these predictions are completely wrong, it still makes sense to look for the minimum risk portfolio consistent with these expectations. We are still left with the problem of estimating $\mathbf{C}$, or maybe $\mathbf{C}^{-1}$ before applying Markowitz's formula, Eq. (8.1.4). We will see below why one should actually find the best estimator of $\mathbf{C}$ itself before inverting it and determining the weights.

What is the *minimum* risk associated to this allocation strategy, measured as the variance of the returns of the portfolio?[2] If one knew the population correlation matrix, $\mathbf{C}$, the *true* optimal risk associated $\mathbf{w}_{\mathbf{C}}$ would be given by

$$\mathcal{R}^2_{\text{true}} := \langle \mathbf{w}_{\mathbf{C}}\,, \mathbf{C}\mathbf{w}_{\mathbf{C}} \rangle = \frac{\mathcal{G}^2}{\mathbf{g}^*\mathbf{C}^{-1}\mathbf{g}}. \tag{8.1.5}$$

However, the optimal strategy (8.1.4) is not attainable in practice as the matrix $\mathbf{C}$ is unknown. What can one do then, and how badly is the realized risk of the portfolio estimated?

**8.1.1. Predicted and realized risk.** One obvious – but naive – way to use the Markowitz optimal portfolio is to apply (8.1.4) using the empirical matrix $\mathbf{E}$ instead of $\mathbf{C}$. Recalling the results of Chapter 4 and 5, it is not hard to see that this strategy should suffer from strong biases whenever $T$ is not sufficiently large compared to $N$, which is precisely the case we consider here. Notwithstanding, the optimal investment weights using the empirical matrix $\mathbf{E}$ read:

$$\mathbf{w}_{\mathbf{E}} = \mathcal{G}\frac{\mathbf{E}^{-1}\mathbf{g}}{\mathbf{g}^*\mathbf{E}^{-1}\mathbf{g}}, \tag{8.1.6}$$

and the minimum risk associated to this portfolio is thus given by

$$\mathcal{R}^2_{\text{in}} = \langle \mathbf{w}_{\mathbf{E}}\,, \mathbf{E}\,\mathbf{w}_{\mathbf{E}} \rangle = \frac{\mathcal{G}^2}{\mathbf{g}^*\mathbf{E}^{-1}\mathbf{g}}, \tag{8.1.7}$$

which is known as the "in-sample" risk, or the *predicted* risk. Let us assume for a moment that $\mathbf{g}$ is independent from $\mathbf{C}$ (and hence, from $\mathbf{E}$). Then, using the convexity with respect to $\mathbf{E}$ of $\mathbf{g}^*\mathbf{E}^{-1}\mathbf{g}$ we find from Jensen inequality that

$$\mathbb{E}[\mathbf{g}^*\mathbf{E}^{-1}\mathbf{g}] \geqslant \mathbf{g}^*\mathbb{E}\big[\mathbf{E}\big]^{-1}\mathbf{g} = \mathbf{g}^*\mathbf{C}^{-1}\mathbf{g} \tag{8.1.8}$$

because $\mathbf{E}$ is an unbiased estimator of $\mathbf{C}$. Hence, we conclude that the in-sample risk is lower than the 'true' risk and therefore, our optimal portfolio suffers from an in-sample bias: its predicted risk underestimates the true optimal risk, and even more so the future *out-of-sample* or *realized* risk, that is the risk realized in the period subsequent to the estimation period. Let us denote

[2] An equivalent risk measure is the volatility which is simply the square root of the variance of the portfolio strategy.

by $\mathbf{E}'$ the empirical matrix of this out-of-sample period; the *out-of-sample* risk is then naturally defined by:

$$\mathcal{R}^2_{\text{out}} = \langle \mathbf{w_E}, \mathbf{E}' \mathbf{w_E} \rangle = \frac{\mathcal{G}^2 \mathbf{g}^\dagger \mathbf{E}^{-1} \mathbf{E}' \mathbf{E}^{-1} \mathbf{g}}{(\mathbf{g}^\dagger \mathbf{E}^{-1} \mathbf{g})^2}. \tag{8.1.9}$$

For large matrices, we expect the result to be self-averaging and given by its expectation. Since the noise in $\mathbf{w_E}$ can be assumed to be independent from that in $\mathbf{E}'$, we get for large $N$ [141]:

$$w^*_{\mathbf{E}} \mathbf{E}' w_{\mathbf{E}} \approx w^*_{\mathbf{E}} \mathbf{C} w_{\mathbf{E}} \tag{8.1.10}$$

and one readily obtains, from the fact that Eq. (8.1.5) is the minimum possible risk, the following inequality: $\mathcal{R}^2_{\text{true}} \leqslant \mathcal{R}^2_{\text{out}}$. We plot in Figure 8.1.1 an illustration of these inequalities using the so-called efficient frontier where we assumed that $\mathbf{g} = (1, \dots, 1)^*$. For a given $\mathbf{C}$ (here a shifted GOE around the identity matrix, with $\sigma = 0.2$), we build $\mathbf{w_C}$ and $\mathbf{w_E}$ and compare Eqs. (8.1.5), (8.1.7) and (8.1.9) for $q = 0.5$. We see that using $\mathbf{w_E}$ is clearly overoptimisitc and can potentially lead to desastrous results in practice. We emphasize that this conclusion holds for different risk measures [49, 53].

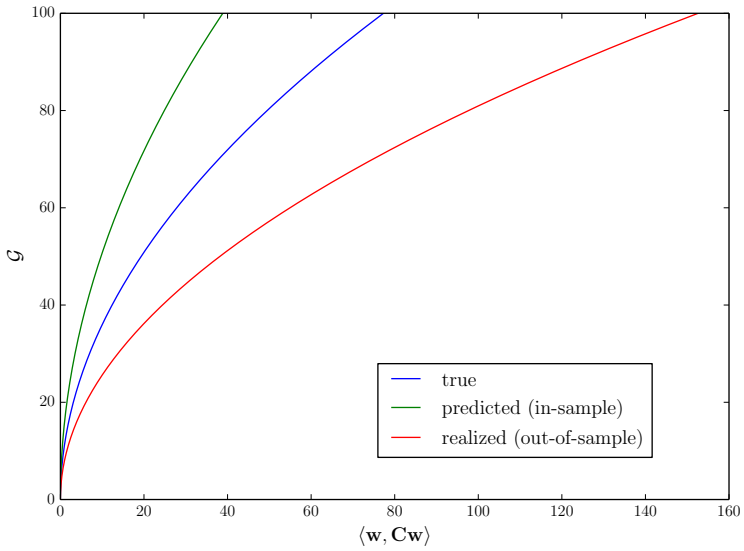



FIGURE 8.1.1. Efficient frontier associated to the mean-variance optimal portfolio (8.1.4) for $\mathbf{g} = (1, \dots, 1)^*$ and $\mathbf{C}$ a shifted GOE around the identity matrix, with $\sigma = 0.2$ and for $q = 0.5$. The blue line depicts the expected gain as a function of the *true* optimal risk (8.1.5) in percentage. The green line the predicted (in-sample) risk while the red line gives the realized (out-of-sample) risk, which is well above the true risk.

**8.1.2. The case of high-dimensional random predictors.** In the limit of large matrices and with some assumptions on the structure $\mathbf{g}$, we can make these inequalities more precise using tools from RMT. In particular, we will show that we can link the true and the realized risk using the

Marčenko-Pastur equation and free probability theory. Let us suppose for simplicity that

$$\mathbf{g} \sim \mathcal{N}_N(0, \mathbf{I}_N), \tag{8.1.11}$$

but the result holds for any random vector $\mathbf{g}$ composed of i.i.d. entries with zero mean, unit variance and a sufficient number of bounded moments. Let $\mathbf{M}$ be a positive definite matrix which is independent from the vector $\mathbf{g}$, then we have in the large $N$ limit,

$$\frac{\mathbf{g}^*\mathbf{M}\mathbf{g}}{N} = \frac{1}{N}\operatorname{Tr}[\mathbf{g}\mathbf{g}^*\mathbf{M}] \underset{\text{freeness}}{=} \frac{\mathbf{g}^*\mathbf{g}}{N}\varphi(\mathbf{M}) \tag{8.1.12}$$

where we recall that $\varphi$ is the normalized trace operator. Thus, from our assumption (8.1.11) we easily deduce,

$$\frac{\mathbf{g}^*\mathbf{M}\mathbf{g}}{N} - \varphi(\mathbf{M}) \underset{N\to\infty}{\to} 0. \tag{8.1.13}$$

Now setting $\mathbf{M} = \{\mathbf{E}^{-1}, \mathbf{C}^{-1}\}$, we apply Eq. (8.1.13) to Eqs. (8.1.7), (8.1.5) and (8.1.9) respectively, to find

$$\begin{aligned}
\mathcal{R}^2_{\text{in}} &\to \frac{\mathcal{G}^2}{N\varphi(\mathbf{E}^{-1})}, \\
\mathcal{R}^2_{\text{true}} &\to \frac{\mathcal{G}^2}{N\varphi(\mathbf{C}^{-1})}, \\
\mathcal{R}^2_{\text{out}} &\to \frac{\mathcal{G}^2\varphi(\mathbf{E}^{-1}\mathbf{C}\mathbf{E}^{-1})}{N\varphi^2(\mathbf{E}^{-1})},
\end{aligned} \tag{8.1.14}$$

where we recall that $\varphi$ is the normalized trace operator defined in Eq. (3.1.61). Let us focus on the first two terms above. For $q < 1$, we know from Eq. (4.2.14) that $\mathfrak{g}_{\mathbf{E}}(0) = -\varphi(\mathbf{E}^{-1})$. The same relation holds for $\mathfrak{g}_{\mathbf{C}}$ if $\mathbf{C}$ has no exact zero mode. Moreover, we showed above that these two quantities are related in the high-dimensional regime through $\varphi(\mathbf{C}^{-1}) = (1-q)\varphi(\mathbf{E}^{-1})$ – see Eq. (4.2.16). As a result, we have, for $N \to \infty$

$$\mathcal{R}^2_{\text{in}} = (1-q)\mathcal{R}^2_{\text{true}}. \tag{8.1.15}$$

Hence, for any $q \in (0,1)$, we see that the in-sample risk associated to $\mathbf{w}_{\mathbf{E}}$ always provides an over-optimistic estimator. Even better, we are able to quantify exactly the risk underestimation thanks to (8.1.15).

Next we would like to find the same type of relation for the "out-of-sample" risk. We recall that under the framework of Chapter 4, we may always rewrite $\mathbf{E} = \mathbf{C}^{1/2}\boldsymbol{\mathcal{W}}\mathbf{C}^{1/2}$ where $\boldsymbol{\mathcal{W}}$ is a white Wishart matrix of parameter $q$ independent from $\mathbf{C}$. Hence, we have for the out-of-sample risk

$$\mathcal{R}^2_{\text{out}} = \frac{\mathcal{G}^2\varphi(\mathbf{C}^{-1}\mathbf{W}^{-2})}{N\varphi^2(\mathbf{E}^{-1})}$$

when $N \to \infty$. Then, the trick is to notice that in the limit of large matrices, $\mathbf{W}$ and $\mathbf{C}$ are *asymptotically free*. This allows us to conclude from the freeness relation (3.1.64) that

$$\varphi(\mathbf{C}^{-1}\mathbf{W}^{-2}) = \varphi(\mathbf{C}^{-1})\,\varphi(\mathbf{W}^{-2}), \tag{8.1.16}$$

Hence, using the asymptotic relation (4.2.16), we find:

$$\mathcal{R}^2_{\text{out}} = \mathcal{G}^2(1-q)^2\frac{\varphi(\mathbf{W}^{-2})}{N\varphi(\mathbf{C}^{-1})}, \tag{8.1.17}$$

Finally, one can readily compute $\varphi(\boldsymbol{\mathcal{W}}^{-2})$ by performing the large $z \to 0$ expansion of the Stieltjes transform of the Marčenko-Pastur density given Eq. in (4.2.16) by replacing $\mathbf{C}$ with $\mathbf{I}_N$, that is to say $\varphi(\boldsymbol{\mathcal{W}}^{-2}) = (1-q)^{-3}$ for $q < 1$. We finally get:

$$\mathcal{R}^2_{\text{out}} = \frac{\mathcal{R}^2_{\text{true}}}{1-q}. \tag{8.1.18}$$

All in all, we obtained the following asymptotic relations:

$$\frac{\mathcal{R}^2_{\text{in}}}{1-q} = \mathcal{R}^2_{\text{true}} = (1-q)\mathcal{R}^2_{\text{out}}, \tag{8.1.19}$$

which holds for a completely general $\mathbf{C}$. Note that similar results have been obtained in a slightly different context in [141] for $\mathbf{C} = I_N$ and later in [56]. Hence, if one invests with the "naive" weights $\mathbf{w_E}$, it turns out that the predicted risk underestimate the realized risk by a factor $(1-q)^2$ and in the extreme case $N = T$ or $q = 1$, the in-sample risk is equal to zero while the out-of-sample risk diverges. We thus conclude that, as announced, the use of the sample covariance matrix $\mathbf{E}$ for the Markowitz optimization problem can lead to disastrous results. This suggests that we should have a more reliable estimator of $\mathbf{C}$ in order to control the 'out-of-sample' risk.

**8.1.3. Out-of-sample risk minimization.** We insisted throughout the last section that the right quantity to control in portfolio management is the realized, out-of-sample risk. It is also clear from Eq. (8.1.19) that using the sample estimate $\mathbf{E}$ is a very bad idea and hence, it is natural to wonder which estimator of $\mathbf{C}$ one should use to minimize this out-of-sample risk? The Markowitz formula (8.1.4) naively suggests that one should look for a faithful estimator of the so-called precision matrix $\mathbf{C}^{-1}$. But in fact, since the expected out-of-sample risk involves the matrix $\mathbf{C}$ linearly, it is that matrix that should be estimated. There are two different approaches to argue that the oracle estimator indeed yields the optimal out-of-sample risk.

The first approach consists in rephrasing the Markowitz problem in terms of conditional expectation. Indeed, the Markowitz problem can be thought as the minimization of the expected future risk given the observations available at the investment date. More formally, it can be written as[3]

$$\begin{cases} \min_{\mathbf{w}} \mathbb{E}\left[ \frac{1}{T_{\text{out}}} \left( \sum_{t'=t+1}^{t+T_{\text{out}}} \langle \mathbf{w}\,, \mathbf{r}_{t'} \rangle \right)^2 \middle| \mathcal{F}(t) \right], \\ \text{s.t. } \mathbf{w}^* \mathbf{g} \geq \mathcal{G}\,, \end{cases} \tag{8.1.20}$$

where $\mathcal{F}(t)$ is all the information available at time $t$ (the investment data), $T_{\text{out}}$ is the out-of-sample period, and $\mathbf{r}$ is the vector of returns of the $N$ stocks in our portfolio. Assuming iid returns means that the optimal weights are independent from the future realizations of $\mathbf{r}$. Moreover, we assume that $\mathcal{P}(\mathbf{r}_{t'}) \propto \mathcal{P}(\mathbf{r}_{t'}|\mathbf{C})\mathcal{P}_0(\mathbf{C})$ for $t' > t$, where $\mathcal{P}_0(\mathbf{C})$ is an (arbitrary) prior distribution on the population covariance matrix $\mathbf{C}$. One then has:

$$\begin{aligned} \mathbb{E}\left[ \frac{1}{T_{\text{out}}} \Big( \sum_{t'=t+1}^{t+T_{\text{out}}} \langle \mathbf{w}\,, \mathbf{r}_{t'} \rangle \Big)^2 \middle| \mathcal{F}(t) \right], &= \left\langle \mathbf{w}\,, \frac{1}{T_{\text{out}}} \sum_{t'} \mathbb{E}\Big[ \mathbf{r}_t \mathbf{r}_t^* \Big| \mathcal{F}(t) \Big] \, \mathbf{w} \right\rangle, \\ &= \left\langle \mathbf{w}\,, \mathbb{E}\Big[ \mathbf{C} \Big| \mathcal{F}(t) \Big] \, \mathbf{w} \right\rangle. \end{aligned} \tag{8.1.21}$$

[3] Recall that we neglect the expected return $\mathbf{g}$ in the calculation of the variance, since the latter is usually small compared to the volatility.

Recalling the results from Chapter 6, we see that $\mathbb{E}[\mathbf{C}|\mathcal{F}(t)] = \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}$ under a multivariate Gaussian assumption on the returns[4] (see Eq. (6.2.2)). Therefore, using the result Eq. (6.2.3), we can conclude that the oracle estimator is the one that minimizes the out-of-sample risk in that specific framework.

There exists another, perhaps more direct derivation of the same result that we shall now present. It is based on the relation (8.1.9) Let us show this explicitely in the context of rotationally invariant estimators, that we considered in Chapter 6 and 7. Let us define our RIE as

$$\Xi = \sum_{i=1}^{N} \xi(\lambda_i) \mathbf{u}_i \mathbf{u}_i^*,$$

where we recall that $[\mathbf{u}_i]_i$ are the sample eigenvectors and $\xi(\cdot)$ is a function that has to be determined. Suppose that we construct our portfolio $\mathbf{w}_\Xi$ using this RIE, that we assume to be independent of the prediction vector $\mathbf{g}$. Again, we assume for simplicity that $\mathbf{g}$ is a Gaussian vector with zero mean and unit variance. Consequently, the estimate (8.1.13) is still valid, such that the realized risk associated to the portfolio $\mathbf{w}_\Xi$ reads for $N \to \infty$:

$$\mathcal{R}^2_{\text{out}}(\Xi) = \mathcal{G}^2 \frac{\text{Tr}\left(\Xi^{-1} \mathbf{C} \Xi^{-1}\right)}{\left(\text{Tr}\, \Xi^{-1}\right)^2}. \tag{8.1.22}$$

using the spectral decomposition of $\Xi$, we can rewrite the numerator as

$$\text{Tr}\left(\Xi^{-1} \mathbf{C} \Xi^{-1}\right) = \sum_{i=1}^{N} \frac{\langle \mathbf{u}_i, \mathbf{C}\mathbf{u}_i \rangle}{\xi^2(\lambda_i)}. \tag{8.1.23}$$

On the other hand, one can rewrite the denominator of Eq. (8.1.22) as

$$\left(\text{Tr}\, \Xi^{-1}\right)^2 = \left( \sum_{i=1}^{N} \frac{1}{\xi(\lambda_i)} \right)^2. \tag{8.1.24}$$

Regrouping these last two equations allows us to rewrite Eq. (8.1.22) as

$$\mathcal{R}^2_{\text{out}}(\Xi) = \mathcal{G}^2 \sum_{i=1}^{N} \frac{\langle \mathbf{u}_i, \mathbf{C}\mathbf{u}_i \rangle}{\xi^2(\lambda_i)} \left( \sum_{i=1}^{N} \frac{1}{\xi(\lambda_i)} \right)^{-2}. \tag{8.1.25}$$

Our aim is to find the optimal shrinkage function $\xi(\lambda_j)$ associated to the sample eigenvalues $[\lambda_j]_{j=1}^N$, such that the out-of-sample risk is minimized. This can be done by solving, for a given $j$, the following first order condition:

$$\frac{\partial \mathcal{R}^2_{\text{out}}(\Xi)}{\partial \xi(\lambda_j)} = 0. \tag{8.1.26}$$

By performing the derivative with respect to $\xi(\lambda_j)$ in (8.1.25), one obtains

$$-2 \frac{\langle \mathbf{u}_j, \mathbf{C}\mathbf{u}_j \rangle \xi'(\lambda_j)}{\xi^3(\lambda_j)} \left( \sum_{i=1}^{N} \frac{1}{\xi(\lambda_i)} \right)^{-2} + 2 \frac{\xi'(\lambda_j)}{\xi^2(\lambda_j)} \left( \sum_{i=1}^{N} \frac{\langle \mathbf{u}_i, \mathbf{C}\mathbf{u}_i \rangle}{\xi^2(\lambda_i)} \right) \left( \sum_{i=1}^{N} \frac{1}{\xi(\lambda_i)} \right)^{-3} = 0, \tag{8.1.27}$$

[4] We expect this result to hold also for the multivariate Student, see Section 4.1.3.

and one can check that the solution is precisely given by

$$\xi(\lambda_j) = \langle \mathbf{u}_j \, , \mathbf{C}\mathbf{u}_j \rangle := \xi_j^{\text{ora.}}, \tag{8.1.28}$$

which is the oracle estimator that we have studied in the chapters 6 and 7.

As a conclusion, the optimal RIE (7.2.2) actually minimizes the out-of-sample risk under the class of rotationally invariant estimators under some distribution assumptions. Moreover, the corresponding "optimal" realized risk is given by

$$\mathcal{R}^2_{\text{out}}(\Xi^{\text{ora.}}) = \frac{\mathcal{G}^2}{\text{Tr}\left[(\Xi^{\text{ora.}})^{-1}\right]}, \tag{8.1.29}$$

where we used the notable property that for any $n \in \mathbb{Z}$:

$$\text{Tr}[(\Xi^{\text{ora.}})^n \mathbf{C}] = \text{Tr}[(\Xi^{\text{ora.}})^{n+1}], \tag{8.1.30}$$

which directly follows from the general formula (7.1.2). Note that this result has also been obtained in [117] where the authors also showed that this estimator maximizes the Sharpe ratio, i.e., the expected return of the strategy divided by its volatility.

**8.1.4. Optimal in and out-of-sample risk for an Inverse Wishart prior.** In this section, we specialize the result (8.1.29) to the case when $\mathbf{C}$ is an Inverse-Wishart matrix with parameter $\kappa > 0$, corresponding to the simple linear shrinkage optimal estimator. Notice that we shall assume throughout this section that there are no outliers ($r = 0$). Firstly, we infer from Eq. (3.1.55) by $z \to 0$ that

$$\varphi(\mathbf{C}^{-1}) = -\mathfrak{g}_{\mathbf{C}}(0) = 1 + \frac{1}{2\kappa}\,, \tag{8.1.31}$$

so that we get from Eq. (8.1.14) that in the large $N$ limit:

$$\mathcal{R}^2_{\text{true}} = \frac{\mathcal{G}^2}{N}\frac{2\kappa}{1+2\kappa}\,. \tag{8.1.32}$$

Next, we see from Eq. (8.1.29) that the optimal out-of-sample risk requires the computation of $\varphi((\Xi^{\text{ora.}})^{-1})$. In general, the computation of this normalized is highly non-trivial but we shall show that some genuine simplifications appear when $\mathbf{C}$ is an inverse Wishart. In the LDL, the final result, whose derivation is postponed at the end of this section, reads:

$$\varphi((\Xi^{\text{ora.}})^{-1}) = -(1+2q\kappa)\mathfrak{g}_{\mathbf{E}}(-2q\kappa) = 1 + \frac{1}{2\kappa(1+q(1+2\kappa))}\,, \tag{8.1.33}$$

and therefore we have from Eq. (8.1.29)

$$\mathcal{R}^2_{\text{out}}(\Xi^{\text{ora.}}) = \frac{\mathcal{G}^2}{N}\frac{2\kappa(1+q(1+2\kappa))}{1+2\kappa(1+q(1+2\kappa))}\,, \tag{8.1.34}$$

from which it is clear from Eqs. (8.1.34) and (8.1.32) that for any $\kappa > 0$:

$$\frac{\mathcal{R}^2_{\text{out}}(\Xi^{\text{ora.}})}{\mathcal{R}^2_{\text{true}}} = 1 + q\frac{2\kappa}{1+2\kappa(1+q(1+2\kappa))} \geqslant 1\,, \tag{8.1.35}$$

where the last inequality becomes an equatlity only when $q = 0$, as it should.

It is also interesting to evaluate the in-sample risk associated to the oracle estimator. It is defined by

$$\mathcal{R}^2_{\text{in}}(\Xi^{\text{ora.}}) = \mathcal{G}^2 \frac{\text{Tr}\big[(\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}\big]}{N\varphi^2((\Xi^{\text{ora.}})^{-1})}\,, \tag{8.1.36}$$

where the most challenging term is the numerator. As above, the computation of this term is, to our knowledge, not trivial in the general case but using the fact that the eigenvalues of $\Xi^{\text{ora.}}$ are given by (7.4.4), we can once again find a closed formula. As above, we relegate the derivation at the end of this section and the result reads:

$$\varphi\big((\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}\big) = -(1-z)^2\big[\mathfrak{g}_{\mathbf{E}}(z) + z\mathfrak{g}'_{\mathbf{E}}(z)\big]\Big|_{z=-2q\kappa} = \frac{(1+2\kappa)(1+2q\kappa)^3}{2\kappa(1+q(1+2\kappa))^3}\,. \tag{8.1.37}$$

Hence by plugging Eqs. (8.1.37) and (8.1.33) into Eq. (8.1.36), we obtain

$$\mathcal{R}^2_{\text{in}}(\Xi^{\text{ora.}}) = \frac{\mathcal{G}^2}{N}\frac{2\kappa(1+2q\kappa)}{(1+2\kappa)(1+q(1+2\kappa))}\,, \tag{8.1.38}$$

and we therefore deduce with Eq. (8.1.32) that for any $\kappa > 0$:

$$\frac{\mathcal{R}^2_{\text{in}}(\Xi^{\text{ora.}})}{\mathcal{R}^2_{\text{true}}} = 1 - \frac{q}{1+q(1+2\kappa)} \leqslant 1\,, \tag{8.1.39}$$

where the inequality becomes an equality for $q = 0$ as above.

Finally, one may easily check from Eqs. (8.1.19), (8.1.35) and (8.1.39), that

$$\mathcal{R}^2_{\text{in}}(\Xi^{\text{ora.}}) - \mathcal{R}^2_{\text{in}}(\mathbf{E}) \geqslant 0, \qquad \mathcal{R}^2_{\text{out}}(\Xi^{\text{ora.}}) - \mathcal{R}^2_{\text{out}}(\mathbf{E}) \leqslant 0\,, \tag{8.1.40}$$

showing explicitely that we indeed reduce the over-fitting by using the oracle estimator instead of the sample covariance matrix in the high dimensional framework.

The aim of this technical section is to derive the results (8.1.33) and (8.1.37). We begin with Eq. (8.1.33) and we use that the eigenvalues of the oracle estimator converges to Eq. (7.4.4) when $N \to \infty$ and $\mathbf{C}$ is an inverse Wishart of parameter $\kappa > 0$. Hence, this yields

$$\varphi((\Xi^{\text{ora.}})^{-1}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\alpha_s(\lambda_i-1)} = \frac{1}{\alpha_s}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\frac{1-\alpha_s}{\alpha_s}+\lambda_i}\,, \tag{8.1.41}$$

and using Eq. (6.3.7), we also have

$$\frac{1}{\alpha_s} = 1+2q\kappa, \quad \text{and} \quad \frac{1-\alpha_s}{\alpha_s} = 2q\kappa\,.$$

We may conclude that

$$\varphi((\Xi^{\text{ora.}})^{-1}) \sim (1+2q\kappa)\mathfrak{g}_{\mathbf{E}}(-2q\kappa)\,, \tag{8.1.42}$$

where we emphasize that the Stieltjes transform is analytic since its argument is non-positive for any $\kappa > 0$. This is the first equality of Eq. (8.1.33) that relates the computation of the normalized trace with the Stieltjes transform of $\mathbf{E}$. When $\mathbf{C}$ is an Inverse Wishart, we know that $\mathfrak{g}_{\mathbf{E}}$ is explicit and given by (4.2.33). Nonetheless, it seems that Eq. (4.2.33) is diverging for $z = -2q\kappa$ so that one has to be careful in the evaluation of $\mathfrak{g}_{\mathbf{E}}(-2q\kappa)$. To that end, we fix $z = -2q\kappa + \varepsilon$ with $\varepsilon > 0$ and expand the numerator of Eq. (4.2.33) as a power of $\varepsilon$ to find:

$$\mathfrak{g}_{\mathbf{E}}(z) = \frac{q-z}{z(1+q-z)} + \mathcal{O}(\varepsilon),$$

meaning that for $\varepsilon = 0$, we obtain

$$\mathfrak{g}_{\mathbf{E}}(-2q\kappa) = -\frac{1+2\kappa}{2\kappa(1+q(1+2\kappa))}. \tag{8.1.43}$$

It is then easy to deduce Eq. (8.1.33) from this last equation and Eq. (8.1.42).

The computation of Eq. (8.1.37) is a bit more tedious but very similar to the derivation of the previous paragraph. Indeed, using that $(\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}$ share the same eigenbasis, we have thanks to Eq. (7.4.4):

$$\varphi((\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i}{(1+\alpha_s(\lambda_i-1))^2}\,, \tag{8.1.44}$$

which gives after some simple manipulations:

$$\varphi((\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}) = \frac{1}{\alpha_s}\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{1+\alpha_s(\lambda_i-1)} - \frac{1-\alpha_s}{(1+\alpha_s(\lambda_i-1))^2}\right]. \tag{8.1.45}$$

Defining $z = -2q\kappa < 0$, one can deduce the first equality of Eq. (8.1.37) using the same identification with the Stieltjes transform (and its derivative with respect to $z$) as above. The derivative of Eq. (4.2.33) reads:

$$\mathfrak{g}'_{\mathbf{E}}(z) = \frac{1}{z^2(z+2q\kappa)^2}\left[z(2\kappa q+z)\left(1+\kappa-\frac{\kappa(\kappa(q-z+1)+1)}{\sqrt{\kappa^2(z+q-1)^2-2\kappa z(1+2\kappa)}}\right) - 2(q\kappa+z)\beta(z)\right], \tag{8.1.46}$$

where $\beta(z)$ is defined by

$$\beta(z) \;:=\; z(1+\kappa)-\kappa(1-q)+\sqrt{\kappa^2(z+q-1)^2-2\kappa z(1+2\kappa)}\,, \tag{8.1.47}$$

which is the denominator of Eq. (4.2.33). We omit further details as the proof of the second equality of Eq. (8.1.37) relies on a Taylor expansion around $-2q\kappa$ in the same spirit than in the previous paragraph. This regularizes the Stieltjes transform and its derivative and one eventually obtains:

$$-2q\kappa\mathfrak{g}'_{\mathbf{E}}(-2q\kappa) = \frac{q(1+2\kappa)\big[q+2(1+\kappa+2q\kappa(1+\kappa))\big]}{2\kappa(1+q(1+2\kappa))^3} \tag{8.1.48}$$

and we find the desired result by plugging this last equation into Eq. (8.1.37).

## 8.2 A short review on previous cleaning schemes

In this section, we give a short survey of the many attempts in the literature to circumvent the above "in-sample" curse by *cleaning* the covariance matrix before using it for i.e. portfolio construction. Even if most of the recipes considered below are not optimal (in a statistical sense), a lot of interesting ideas have been proposed to infer the statistical properties of the unknown population matrix. As we shall see, most of the methods appeared after the seminal work of Marčenko & Pastur [123]. We nonetheless stress that the literature on estimating large covariance matrices is so large that it is impossible to make justice to all the available results here. We will only consider methods for which RMT results offer interesting insights and refer to, e.g. [16, 29, 145] for complementary sources of information.

We shall present four different classes of estimators. The first one is the linear shrinkage. This estimator has been studied in details in Chapters 6 and 7 but here, we focus on the estimation of the shrinkage intensity. As we will see, RMT will provide very simple methods to estimate parameters from the data.

Then we will present the **eigenvalues clipping** method of [111, 151] where the aim is to separate "trustworthy" eigenvalues from "noisy" ones. The basic idea of this method is the

spiked covariance matrix model that we presented in Section 4 where the true eigenvalues consist in a finite number $r$ of spikes and one degenerate eigenvalue $\approx 1 - \mathcal{O}(r/N)$, with multiplicity $N - r$.

The third method, that we name **eigenvalues substitution**, consists in solving the inverse Marčenko-Pastur problem (see Section 4). Roughly speaking, in the presence of a very large number of eigenvectors, one can discretize the Marčenko-Pastur equation and solve the inverse problem using either a parametric [29] or non-parametric approach [104].

The last method concerns **factors models**, or structured covariance estimators, where one tries to explain the correlation matrix through a simplified model of the underlying structure of the data. This is a very popular approach in finance and economics, and we will see how RMT has allowed some recent progress.

All these methods will be tested using real financial data in the next chapter.

**8.2.1. Linear Shrinkage.** We recall that the linear shrinkage is given by

$$\Xi^{\text{lin}} = \alpha_s \mathbf{E} + (1 - \alpha_s)\mathbf{I}_N, \qquad \alpha \in [0, 1]. \tag{8.2.1}$$

As discussed in Chapter 6, this estimator has a long history in high-dimensional statistics [87,115] as it provides a simple proof that the sample estimator $\mathbf{E}$ is inconsistent whenever $N$ and $T$ are both large. A very exhaustive presentation of the properties of this estimator in the high-dimensional regime can be found in [115] or in [105] in a more RMT oriented standpoint. It is easy to see that $\Xi^{\text{lin}}$ shares the same eigenbasis than the sample estimator $\mathbf{E}$, and is thus a rotationally invariant estimator with

$$\Xi^{\text{lin}} = \sum_{i=1}^{N} \xi^{\text{lin}} \mathbf{u}_i \mathbf{u}_i^*, \qquad \xi^{\text{lin}} = 1 + \alpha_s(\lambda_i - 1) \tag{8.2.2}$$

We already emphasized that this estimator exhibits all the expected features: the small eigenvalues are "shrunk" upwards (compare to the sample eigenvalues) while the top eigenvalues are "shrunk "downwards (see Figure 8.2.1). As alluded above, this estimator has been fully investigated in [115]. Most notably, the authors were able to determine an asymptotic optimal formula to estimate $\alpha_s$ directly from the data. Keeping the notations of Section 4, our data set is $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_T) \in \mathbb{R}^{N \times T}$ and we assume that $\mathbb{E}[Y_{it}] = 0$ and $\mathbb{E}[Y_{it}^2] = T^{-1}$ for all $i \in [\![1, N]\!]$. Defining:

$$\begin{aligned} \beta &:= \frac{1}{N} \operatorname{Tr}\left[(\mathbf{E} - \mathbf{I}_N)(\mathbf{E} - \mathbf{I}_N)^*\right] \\ \gamma &:= \max\left(\beta, \frac{1}{T^2} \sum_{k=1}^{T} \frac{1}{N} \operatorname{Tr}\left[(\mathbf{y}_k \mathbf{y}_k^* - \mathbf{E})(\mathbf{y}_k \mathbf{y}_k^* - \mathbf{E})^*\right]\right), \end{aligned} \tag{8.2.3}$$

then

$$\widehat{\alpha}_s = 1 - \frac{\beta}{\gamma}, \tag{8.2.4}$$

is a consistent estimator of $\alpha_s$ in the high-dimensional regime [115].

Using tools from RMT, and more precisely the result of Sections 4 and 5, we can find another consistent estimators of $\alpha_s$ which uses the fact that linear shrinkage implicitely assumes the underlying correlation matrix to be an Inverse-Wishart matrix with parameter $\kappa$, from which

$\alpha_s$ is deduced as $\alpha_s = (1+2q\kappa)^{-1}$. The value of $\kappa$ can be extracted from the data using the relation (valid for $q<1$):

$$\mathfrak{g}_{\mathbf{C}}(0) = (1-q)\mathfrak{g}_{\mathbf{E}}(0) = 1+2\kappa. \tag{8.2.5}$$

where the last equality can be deduced from (3.1.55) and (4.2.16). Therefore, we obtain a simple estimate for $\kappa$ from the trace of $\mathbf{E}^{-1}$ as:

$$\kappa = \frac{1}{2}\left((1-q)\frac{\operatorname{Tr}\mathbf{E}^{-1}}{N} - 1\right). \tag{8.2.6}$$

However, this estimate is only reliable when $\kappa$ is not too large, i.e. when $\mathbf{C}$ is significantly different from the identity matrix (in the opposite case, $(1-q)\operatorname{Tr}\mathbf{E}^{-1} \approx N$ so that one can obtain negative values for $\kappa$). A more robust alternative is to use the "two-samples" test introduced in Chapter 5.2, see Eqs (5.2.17) and [41].

**8.2.2. Eigenvalues clipping.** This method is perhaps the first RMT-based estimator for large covariance matrices. It has been investigated in several papers [110,111,151] where the Marčenko-Pastur distribution is used in a very intuitive way to correct the sample eigenvalues. The idea of the method is as follows: all the eigenvalues that are beyond the largest expected eigenvalue of the empirical matrix $\lambda_+ = (1+\sqrt{q})^2$ (within a null hypothesis) are interpreted as signal while the others are pure noise (see Figure 4.3.1). An alternative interpretation would be that outliers are true factors while the others are meaningless.

In a recent paper [27], this idea has been made rigorous in the sense that if we suppose that $\mathbf{C}$ is a finite rank perturbation of $\mathbf{I}_N$ as defined in (4.3.6), then the reference matrix of the bulk eigenvalues of $\mathbf{E}$ simply corresponds to the (isotropic) Wishart matrix $\boldsymbol{\mathcal{W}}$. Differently said, for this specific model, these bulk eigenvalues should be seen as pure noise, and the right edge $(1+\sqrt{q})^2$ can be interpreted as the filter between noise and signal.

Endowed with a simple rule to isolate the signal eigenvalues, how should one clean the noisy ones ones? Laloux et al. [111] proposed the following rule: first diagonalize the matrix $\mathbf{E}$ and keep the eigenvectors unchanged. Then apply the following scheme in order to denoise the sample eigenvalues:

$$\Xi^{\text{clip.}} := \sum_{i=1}^{N} \xi_i^{\text{c}} \mathbf{u}_i \mathbf{u}_i^*, \qquad \xi_i^{\text{clip.}} = \begin{cases} \lambda_i & \text{if } \lambda_i \geqslant (1+\sqrt{q})^2 \\ \bar{\lambda} & \text{otherwise,} \end{cases} \tag{8.2.7}$$

where $\bar{\lambda}$ is chosen such that $\operatorname{Tr}\Xi^{\text{clip.}} = \operatorname{Tr}\mathbf{E}$. Roughly speaking, this method simply states that the noisy eigenvalues are shrunk toward a (single) constant such that the trace is preserved. This procedure is known as *clipping* and Figure 8.2.1 shows how it shifts upwards the lowest eigenvalues in order to avoid *a priori* abnormal low variance modes.

Nonetheless, the method suffers from several separate problems. First, one often observes empirically, especially with financial data, that the value of $q = N/T$ that is fixed by the dimensionality of the matrix and the length of the time series is significantly different from the "effective" value $q_{\text{eff}}$ that allows one to fit best the empirical spectral density [111]. This effect can be induced either by small temporal autocorrelation in the time series [18, 47, 48] and/or by the inadequacy of the null hypothesis $\mathbf{C} = \mathbf{I}_N$ for the bulk of the distribution. In any case, a simple recipe would be to use a corrected upper edge $\lambda_+ = (1+\sqrt{q_{\text{eff}}})^2$ for the threshold separating wheat from chaff. Another possibility, proposed in [29], is to introduce a fine-tuning parameter $\alpha_c \in [0,1]$ such that the $\lceil N\alpha_c \rceil$ largest eigenvalues are kept unaltered while the others

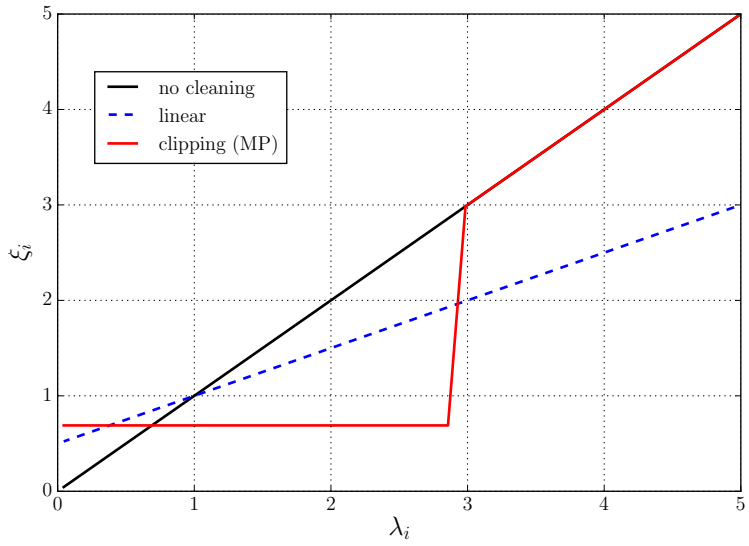



FIGURE 8.2.1. Impact on sample eigenvalues of the eigenvalues clipping (8.2.7) (red plain line) with a threshold given by $(1+\sqrt{q})^2$ with $q = 0.5$ and the linear shrinkage (8.2.2) (blue dashed line) with intensity $\alpha_s = 0.5$. We see that the lowest eigenvalues are shifted upward.

are still replaced by a common $\bar{\lambda}$. It is easy to see that for $\alpha_c = 1$, we get the empirical covariance matrix while for $\alpha_c = 0$, we get the identity matrix. So $\alpha_c$ plays the role of the upper bound $\lambda_+$ of the Marčenko-Pastur density, and allows one to interpolate between $\mathbf{E}$ and the null hypothesis $\mathbf{I}_N$, much like linear shrinkage. Nevertheless, the calibration of the parameter $\alpha_c$ is not based on any theoretical rule.

Another concern about this method is that we know from section 7.3 that the optimal estimator of the large outliers is not their bare empirical value $\lambda_i$. Rather, one should shift them downwards, by a quantity equal to $-2q$ (in the limit $\lambda_i \gg 1$). Hence, at the very least, such a shift should be included in the eigenvalue clipping scheme from Eq. (8.2.7) (see [17] for a related discussion).

**8.2.3. Eigenvalue substitution.** The main idea behind the eigenvalue substitution method is also quite intuitive and amounts to replacing the sample eigenvalues by their corresponding "true" values obtained by inverting the Marčenko-Pastur equation (4.2.1). More formally, we seek the set of true eigenvalues $\{\mu_j\}_{j\in[\![1,N]\!]}$ that solve Eq. (4.2.1) for a *given* set of sample eigenvalues $\{\lambda_j\}_{j\in[\![1,N]\!]}$. As for the eigenvalues clipping procedure, this technique can be seen a nonlinear shrinkage function and has the advantage to lean upon a more robust theoretical framework than the clipping "recipe". However, as we emphasized in Section 4.2.1, inverting the Marčenko-Pastur equation is quite challenging in practice. In this section, we present several possibilities to achieve this goal in the limit of large dimensions.

**Parametrization of Marčenko-Pastur equation.** One way to think about the inverse Marčenko-Pastur problem is to adopt a Bayesian viewpoint (like in Chapter 6). More specifically, we assume that $\mathbf{C}$ belongs to a rotationally invariant ensemble – so that the there is no a priori knowledge about the eigenvectors – and assume a certain structure on the LSD $\rho_{\mathbf{C}}(\mu)$, parameterized by

one or several numbers. The optimal values of these parameters (and the corresponding optimal $\widehat{\rho}_{\mathbf{C}}$) are then fixed by e.g. a maximum likelihood procedure on the associated $\rho_{\mathbf{E}}$, obtained from the direct Marčenko-Pastur equation. Once the fit is done, the *substitution* cleaning scheme reads

$$\lambda_i \to \widehat{\mu}_i \quad \text{such that} \quad \frac{i}{N} = \int_{\widehat{\mu}_i}^{\infty} \widehat{\rho}_{\mathbf{C}}(x)\mathrm{d}x. \tag{8.2.8}$$

Note that under the transformation (8.2.8), we assume that the eigenvalues of $\mathbf{C}$ are allocated smoothly according to the quantile of the limiting density $\widehat{\rho}_{\mathbf{C}}$.

As an illustration of this parametric substitution method, let us consider a power law density (4.2.41) as the prior for $\rho_{\mathbf{C}}(\mu)$. Such a probabilistic model for the population eigenvalues density is thought to be plausible for financial markets, and reflect the power-law distribution of sector sizes in the economy [29, 128]. In that case, the parametric substitution turns out to be explicit in the limit of large dimension. Moreover, the estimation of the unique parameter $\lambda_0$ in this model can be done using e.g. maximum likelihood, as we can compute exactly $\rho_{\mathbf{E}}$ on $\mathbb{R}^+$ using (4.2.42) and (4.2.27). This then yields a parameter $\widehat{\lambda}_0$ and hence $\widehat{\rho}_{\mathbf{C}}$ as well. As a result, the substitution procedure (8.2.8) becomes for $N \to \infty$ [29]:

$$\mu_i = -\widehat{\lambda}_0 + \frac{(1+\widehat{\lambda}_0)}{2}\sqrt{\frac{N}{i}} \qquad i \in [\![1, N]\!]\,. \tag{8.2.9}$$

We present such a procedure in Fig. 8.2.2 using US stocks data. We conclude from this figure that the fit is indeed fairly convincing, i.e. that a power-law density for the eigenvalues of $\mathbf{C}$ is a reasonable assumption.

**Discretization of Marčenko-Pastur equation.** Interestingly, a "quasi" non-parametric procedure is possible under some smoothness assumption on the density $\rho_{\mathbf{C}}$. This algorithm is due to N. El Karoui [104] who proposed to solve an approximate form of the Marčenko-Pastur inverse problem. The starting point is to notice that each eigenvalue of $\mathbf{E}$ satisfies:

$$\left\{ z_j = \frac{1}{\mathfrak{gs}(z_j)} \left[ 1 - q + q \int \frac{\rho_{\mathbf{C}}(\mu) d\mu}{1 - \mu\, \mathfrak{gs}(z_j)} \right], \quad \text{with} \quad z_j = \lambda_j - i\eta \right\}_{j=1}^{N}$$

that follows from Eq. (4.2.27) and where we recall that $\mathbf{S}$ is the $T \times T$ dual matrix of $\mathbf{E}$ defined in (4.2.24). The main assumption of this method is to decompose the density of states $\rho_{\mathbf{C}}$ as a weighted sum of Dirac masses:

$$\rho_{\mathbf{C}}(\mu) = \sum_{k=1}^{N} \widehat{w}_k \delta(\mu - \mu_k), \quad \text{such that} \quad \sum_{k=1}^{N} \widehat{w}_k = 1 \quad \text{and } \widehat{w}_k \geq 0,\, \forall\, k \in [\![1, N]\!]. \tag{8.2.10}$$

Note that this decomposition simply use the discreteness of the eigenvalues that follows from the very definition of an ESD where each eigenvalues are associated with a weight equals to $N^{-1}$. One notices that there are two different sources of uncertainty: the "true" eigenvalues $\{\mu_j\}_j$ and their corresponding weights $\widehat{w}_j$ so that the parametrization looks inextricably complex. In [104], the author suggested to fix the positions $\{\mu_j\}_j$ *a priori* such that we are left with the weights $\widehat{w}_j$ as the only unknown variables in the problem. Within this framework, the author then proposed

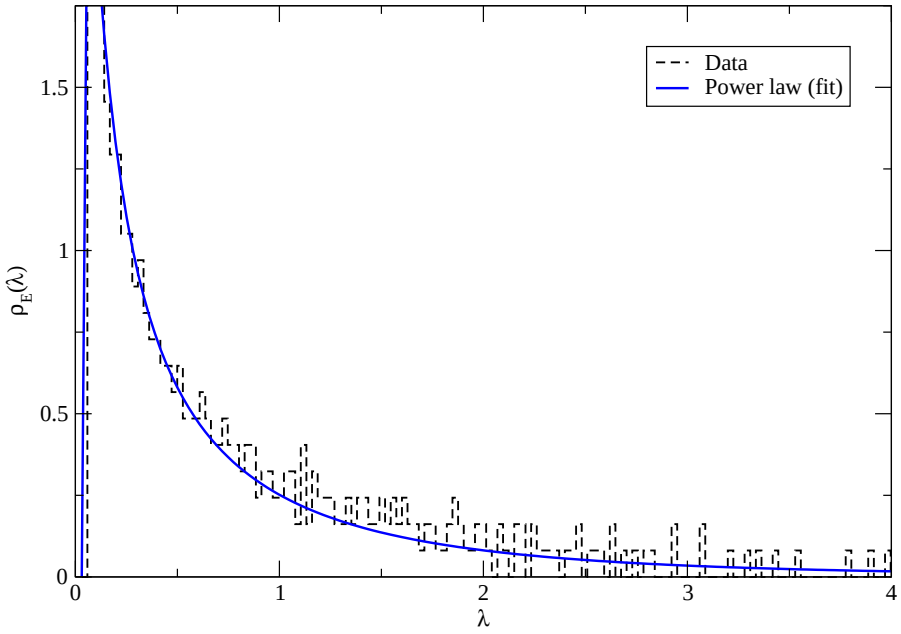



FIGURE 8.2.2. Fit of the power law distribution (4.2.41) on the sample eigenvalues of the 450 most liquid assets of the S&P index from 2006 to 2010 using the Marčenko-Pastur equation (4.2.1). The fit has been performed using a maximum likelihood procedure and yields $\alpha \approx 0.3$. The black dashed histogram represents the empirical spectral density.

to obtain the optimal weights through the following optimization program:

$$\{\widehat{w}_j\}_{j=1}^N = \begin{cases} \underset{\{w_i\}_{i=1}^K}{\operatorname{argmin}} \mathcal{L}\left(\left\{\frac{1}{\mathfrak{g}_{\mathbf{S}}(z_j)}\left[1-q+q\sum_{k=1}^N \frac{w_k}{1-\mu_k\,\mathfrak{g}_{\mathbf{S}}(z_j)}\right]-z_j\right\}_{j=1}^N\right) \\ \text{subject to } \sum_{k=1}^K w_k = 1, \quad \text{and } w_k \geq 0 \quad \forall\, k \in [\![1,N]\!], \end{cases} \tag{8.2.11}$$

where $\mathcal{L}$ is a certain loss function. In addition to the error we make by approximating the true density by a sum of weighted Dirac masses, there are at least two others sources of errors:

1. The approximation $\mathfrak{g}_{\mathbf{E}}(z_j) \approx N^{-1}\mathrm{Tr}(z_j\mathbf{I}_N - \mathbf{E})^{-1}$;
2. The position of the eigenvalues $\{\mu_j\}$ that have to be chosen.

In the large $N$ limit, the first approximation is fairly accurate (see Section 8). However, the second is much more difficult to handle especially in the case of a very diluted spectrum. Note that if we define $e_j$ as the error we make term in (8.2.11) for each $\lambda_j$, then the consistency of the algorithm has been showed in [104] under the norm $L_\infty = \max_{j=1,\dots,N} \max(|\,\mathrm{Re}(e_j)|, |\,\mathrm{Im}(e_j)|)$. Once we get the optimal weight $\{\widehat{w}_k\}$, the cleaning procedure is immediate

$$\lambda_i \to \widehat{\mu}_i \quad \text{where} \quad \widehat{\mu}_i = \min\left\{x \in \mathbb{R}^+ : \sum_{k=1}^N \widehat{w}_k \Theta(\mu_k - x) \geq \frac{i}{N}\right\} \tag{8.2.12}$$

where we have used the approximation

$$\int_x^\infty \rho_{\mathbf{C}}(u)du \approx \sum_{k=1}^N \widehat{w}_k \Theta(\mu_k - x),$$

with $\Theta(x)$ that denotes the Heaviside step function.

While the method is backed by a theoretical framework, it turns out that the error source # 2. above is a strong limitation in practice. A recent proposal to invert the Marčenko-Pastur equation by optimizing directly the eigenvalues $[\mu_j]_j$ has therefore been proposed in [116]. This alternative method, called QuEST, turns out to be much more robust numerically (see [118] and Chapter 9 for an extended discussion and some applications).

As a conclusion, we see that it is possible to solve (approximately) the inverse Marčenko-Pastur equation in a quite general fashion, meaning that we might indeed be able to locate approximately the true eigenvalues $\mu_i$ for any $i = 1, \ldots, N$. As a result, the eigenvalue substitution estimator is then obtained as

$$\Xi^{\text{sub}} = \sum_{k=1}^N \widehat{\mu}_k \mathbf{u}_k \mathbf{u}_k^*. \tag{8.2.13}$$

However, even when a perfect estimation of the true density $\rho_{\mathbf{C}}$ is feasible, we see that this estimator does not take into account the fact that the sample eigenvectors are not consistent estimators of the true ones, as shown in Chapter 5. Therefore, for covariance matrices estimation, it is not advised to use the substitution (8.2.13) since this is not the optimal solution. However, it can be used to compute the optimal RIE (7.2.2) and we refer to Section 9.1.3 for more details.

## 8.3 Factor models

The main idea behind linear factor models is pretty simple: the (normalized) data $Y_{i,t}$ is represented as a linear combination of $M$ common factor $F$

$$Y_{it} = \sum_{k=1}^M \beta_{ik} f_{kt} + \varepsilon_{it} \tag{8.3.1}$$

where the $\beta_{ik}$ are the linear exposures of the variable $i$ to the factors $k = 1, \ldots, M$ at time $t$ and the $N \times T$ matrix $\varepsilon_{it}$ is the idiosyncratic part of $Y_{i,t}$ (or the residual in Statistics), assumed to be of zero mean. The model (8.3.1) in matrix form reads

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{F} + \boldsymbol{\mathcal{E}}, \tag{8.3.2}$$

which is known as *Generalized Linear Model* [130]. It is often assumed that the residuals are i.i.d. across $i$ with $t$ fixed (see e.g. [50] for an application in Finance). It is not hard to see that the covariance matrix under the model (8.3.1), the true covariance matrix is given by

$$\mathbf{C} = \beta \Sigma_F \beta^* + \Sigma_\varepsilon \tag{8.3.3}$$

where $\Sigma_F$ is the covariance matrix of size $M \times M$ of the factor $F$ – which can always be chosen to be proportional to the identity matrix – and $\Sigma_\varepsilon$ is the $N \times N$ covariance matrix of the residuals $\varepsilon$, which is simply the identity in the simplest framework. Within the linear decomposition

(8.3.1), we see that we have generically a number of parameters to estimate of order $\mathcal{O}(NM)$ out of datasets of size $\mathcal{O}(NT)$. Hence, we see that the curse of dimensionality disappears as soon as $M \ll N, T$ which implies that the empirical estimate

$$\mathbf{E} = \frac{1}{T}(\boldsymbol{\beta}\mathbf{F} + \boldsymbol{\mathcal{E}})(\boldsymbol{\beta}\mathbf{F} + \boldsymbol{\mathcal{E}})^*, \tag{8.3.4}$$

becomes more accurate. This is a simple way of cleaning high-dimensional covariance matrices within factor models.

However, this cleaning scheme leaves open at least one question of practical use. How should the number of factor $M$ be chosen? In the case where one has *a priori* information on the factors $F$, we are just left with the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\mathcal{E}}$. But in the general case, this question is still an open problem. Let us treat the general case, in which several authors considered tools from RMT to choose the number of factor $M$.

In [102], the author assumes that the empirical estimator of $\Sigma_\varepsilon$ is given by an isotropic Wishart matrix for which the upper bounds of the spectrum is exactly known. Hence, if there were no tangible factor in the data, one should observe that largest eigenvalues of the matrix $\mathbf{E}$ defined in (8.3.4) cannot exceed

$$\lambda_+^{\text{eff}}(q) := (1 + \sqrt{q})^2 + \delta(q, N) \tag{8.3.5}$$

where the last term $\delta$ is a suitably defined constant as to reflect the width of the Tracy-Widom tail, i.e. $\delta(q, N) \sim T^{-2/3}$ [102]. If however one observes that the largest sample eigenvalue $\lambda_1$ exceeds $\lambda_+^{\text{eff}}$, then a true factor probably exists. In that case, the procedure suggested in [102] is to extract the corresponding largest component from the data:

$$Y_{it}^{(1)} = Y_{it} - \beta_{1,t} f_{1t},$$

which is the residual from a regression of the data on the first principal component. Next, we compare the largest eigenvalue of $\mathbf{Y}^{(1)}\mathbf{Y}^{(1)*}/T$ against the new threshold $\lambda_+^{\text{eff}}(q' = q - 1/T)$ and iterate the procedure until $\mathbf{Y}^{(M)}\mathbf{Y}^{(M)*}/T$ has all its eigenvalues within the Marčenko-Pastur sea. This approach has been generalized in [140] to the case where the empirical estimator of the $\Sigma_\varepsilon$ is an *anisotropic* Wishart matrix for which one has several results concerning the spectrum (see Chapter 4). The procedure is similar to the one above method: the author proposed an algorithm to detect outliers for this anisotropic Wishart matrix using the results of Ref. [146]. We refer to [140] for more details. We can therefore see that RMT allows one to derive some rigourously based heuristics to determine the number of true factors $M$, which are quite similar in spirit to the eigenvalue clipping method described above.

It is also possible that one has some *a priori* insight on the structure of the relevant factors. This for instance is a standard state of affairs in theoretical finance, where the so-called Capital Asset Pricing Model (CAPM) [132] assumes a unique factor corresponds to the market portfolio, or its extension to three factors model by Fama-French [77] (see [169] for further more recent extensions). In that case, one can simplify the problem to the estimation of the $\beta$ by assuming that the the factors $f_k$ and the residuals $\varepsilon_i$ are linearly uncorrelated:

$$\langle f_k f_l \rangle = \delta_{kl} \quad , \quad \langle \varepsilon_i \varepsilon_j \rangle = \delta_{ij} \left( 1 - \sum_l \beta_{li}^2 \right) \quad \text{and} \quad \langle f_k \varepsilon_l \rangle = 0, \tag{8.3.6}$$

such that the true correlation becomes:

$$C_{ij} = \sum_{k=1}^{M} \beta_{ki}\beta_{kj} + \delta_{ij}\left(1 - \sum_{l=1}^{M} \beta_{li}^2\right)$$

that is to say

$$C_{ij} = \begin{cases} 1 \quad \text{if} \quad i = j \\ (\boldsymbol{\beta}\boldsymbol{\beta}^*)_{ij} \quad \text{otherwise.} \end{cases} \tag{8.3.7}$$

Again, we emphasize that we are reduced to the estimation of only $N \times M$ parameters out of $N \times T$ points. We now give an insight on how one can estimate the coefficients of $\beta$ using the sample data, which is due to the recent paper [52]. Note that the eigenvalue clipping (8.3.7) can recovered by setting $\beta \equiv \beta_{PCA}$ where

$$\beta_{PCA} \ := \ \mathbf{U}_{|M}\mathbf{\Lambda}_{|M}^{1/2}\,, \tag{8.3.8}$$

with $\mathbf{U}$ the sample eigenvectors, $\Lambda$ the $N \times N$ diagonal matrix with the sample eigenvalues and the subscript $_{|M}$ denotes that only the $M$ largest components are kept, where $M$ is such that $\lambda_i > (1 + \sqrt{q})^2$ for any $i \leq M$. The method of [52] suggests to find the $\boldsymbol{\beta}$s such that:

$$\widehat{\boldsymbol{\beta}} \ := \ \underset{\beta}{\text{argmin}}\, \mathcal{L}\left(\left\| \frac{1}{T}\mathbf{Y}\mathbf{Y}^* - \boldsymbol{\beta}\boldsymbol{\beta}^* \right\|_{\text{off-diag}}\right), \tag{8.3.9}$$

with $\mathcal{L}$ a given loss function and off-diag to denote the off-diagonal elements. (The diagonal elements are all equal to unity by construction). Numerically, the authors solve the latter equation in the vicinity of the PCA beta's (8.3.8) and with a quadratic norm $\mathcal{L}$. We refer the reader to [52] for more details on the procedure and its implementation, as well as an extension of the model to non-linear (volatility) dependencies.

# Chapter 9

# Numerical Implementation and Empirical results

This chapter aims at putting all the above ideas into practice, the final goal being, in a financial context, to achieve minimum out-of-sample, or forward looking risk. As we have seen above, the Rotationally Invariant Estimator framework is promising in that respect. Still, as one tries to implement this method numerically, some problems arise. For example, we saw in Section 7.5 that the discrete version (7.5.2) of the optimal RIE (7.2.2) deviates systematically from its limiting value for small eigenvalues. But as we discussed in Section 8, the estimation of these small eigenvalues is particularly important since Markowitz optimal portfolios tend to overweight small eigenvalues and hence, inadequate estimators of these small eigenvalues may lead to disastrous results. We will therefore first discuss two different regularization schemes that appeared in the recent literature (see [118] and [42]) that attempt to correct this systematic underestimation of the small eigenvalues. Then we will turn to numerical experiments on synthetic and real financial data, to test the quality of the regularized RIE for real world applications.

## 9.1 Finite $N$ regularization of the optimal RIE (7.5.2)

**9.1.1. Why is there a problem for small-eigenvalues?.** The small eigenvalue bias can be best illustrated using the null hypothesis on the sample covariance matrix. Indeed, we know that for $\mathbf{C} = \mathbf{I}_N$, the optimal RIE (7.2.2) should yield $\widehat{\xi}(\lambda_i) = 1$ exactly as $N \to \infty$ (see Eq. (7.4.1)). We therefore compare the observable shrinkage function $\widehat{\xi}^N$ (7.5.2) for finite $N$ with its limiting value $\widehat{\xi} = 1$. The results are reported in Figure 9.1.1 where the observable estimator Eq. (7.5.2) is represented by the green points and the limiting value is given by the red dotted line. We see that the bulk and the right edge are relatively well estimated, but this is clearly not the case for the left edge, below which the estimated eigenvalues dive towards zero. This highlights, as stated in [43], that the behaviour for small eigenvalues is more difficult to handle compared to the rest of the spectrum.

This underestimation can be investigated analytically. Let us define $z = \lambda - \mathrm{i}\eta$ and we actually see from the Figure 9.1.1 that the discrete RIE $\widehat{\xi}^N$ is a very good approximation of the limiting quantity $\widehat{\xi}(z)$, i.e., with $\eta = N^{-1/2}$ (blue plain line). Hence, the deviation at the left edge is *systematic* for any finite $N$ and it only disappears as $N \to \infty$ ($\eta \to 0^+$). This finite size effect is due to the *hard* left edge as eigenvalues are confined to stay on $\mathbb{R}^+$. Indeed, under the

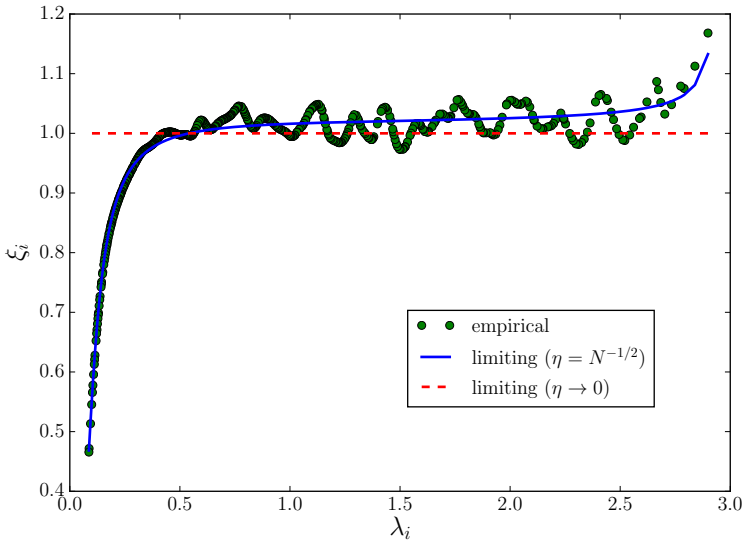



FIGURE 9.1.1. Evaluation of the empirical RIE (7.5.2) (green points) for $\mathbf{C} = \mathbf{I}_N$ with $N = 500$. The matrix $\mathbf{E}$ is generated using Wishart matrix with parameter $q = 0.5$. We compare the the result with its limiting value for $\eta = N^{-1/2}$ (blue line) and $\eta \to 0^+$ (red dotted line).

one-cut assumption, we can always decompose the Stieltjes transform as (see Eq. (3.1.31))

$$\mathfrak{g}_{\mathbf{E}}(z) = \mathfrak{h}(z) + Q(z)\sqrt{d_+(z)}\sqrt{d_-(z)}, \qquad d_\pm(z) \;:=\; z - \lambda_\pm \tag{9.1.1}$$

where $\mathfrak{h}(z)$ is the Hilbert transform of $\rho_{\mathbf{E}}$ and $Q(z)$ is a given function that we assumed be smoothed on $\mathbb{R}^+$. We place ourselves in the situation where $d_-(\lambda) = \varepsilon \ll \eta$, i.e. the eigenvalue $\lambda$ is very close to the left edge. Then, we have

$$\begin{aligned} \mathfrak{g}_{\mathbf{E}}(z) &= \mathfrak{h}(z) + Q(z)\sqrt{-\mathrm{i}\eta}\sqrt{d_+(\lambda) - \mathrm{i}\eta} + \mathcal{O}(\varepsilon) \\ &= \mathfrak{h}(z) - (1+i)Q(z)\sqrt{\frac{\eta|d_+(\lambda)|}{2}} + \mathcal{O}(\varepsilon)\,. \end{aligned} \tag{9.1.2}$$

Specializing this last equation to the null hypothesis $\mathbf{C} = \mathbf{I}_N$, one infers from Eq. (3.1.41) that $1/Q(z) = 2qz$ and $\mathfrak{h}(z) = Q(z)(z + q - 1)$. Then plugging (9.1.2) into (7.2.2) yields, at the left edge:

$$\widehat{\xi}(\lambda_- - \mathrm{i}\eta) = 1 - \sqrt{\frac{2\eta\sqrt{q}}{(1-\sqrt{q})^2}} + \mathcal{O}(\eta), \tag{9.1.3}$$

that is to say, we have a finite size "correction" to the asymptotic result $\widehat{\xi}(z) = 1$ of order $N^{-1/4}$ when $\eta = N^{-1/2}$. This correction is therefore quite significant if $N$ is not large enough. One tempting solution would be to decrease the value of $\eta$ to be arbitrarily small. However, we know that the empirical Stieltjes transform is only a good approximation of the limiting value up to an error of order $(T\eta)^{-1}$, so that $\eta$ cannot be too small either. We conclude that the underestimation effect that we observe in Figures 9.1.1 and 7.5.1 is purely due to a finite size effect and would furthermore occur for any model of $\rho_{\mathbf{C}}$ (see Fig. 7.5.1). We emphasize that this effect is different from cleaning left outliers as displayed in Fig. 7.4.3.

**9.1.2. Denoising the empirical RIE (7.5.2).** There are two ways to address this problem. The first one is to use a simple ad-hoc denoising procedure that we shall now explain; the second is a more sophisticated scheme recently proposed by Ledoit and Wolf (see below).

Firstly, using the fact that the finite size corrections are rather harmless for large eigenvalues (see Figure 9.1.1), we can focus on small sample eigenvalues only. The idea is to use a regularization that would be exact if the true correlation matrix was of the Inverse-Wishart type, with $\rho_{\mathbf{C}}$ to be given by Eq. (3.1.53), for which we know that the associated optimal RIE is the linear shrinkage (7.4.4).[1] Within this specification, the parameter $\kappa$ allows to interpolate $\rho_{\mathbf{C}}$ between the infintely wide measure on $\mathbb{R}^+$ ($\kappa \to 0^+$) and the above null hypothesis ($\kappa \to \infty$).

Our procedure, *for the only purpose of regularization*, is to calibrate $\kappa$ such that the lower edge $\lambda_-^{\rm iw}$ of the corresponding empirical spectrum (and given in Eq. (4.2.33)), coincides with the observed smallest eigenvalue $\lambda_N$. We then rescale the smallest eigenvalues using the exact factor that would be needed if $\mathbf{C}$ was indeed an Inverse-Wishart matrix, i.e.:

$$\widehat{\xi}_i^{\rm reg} = \widehat{\xi}_i^N \times \max(1, \Gamma_i^{\rm iw}), \qquad \Gamma_i^{\rm iw} = \frac{|1 - q + q z_i \mathfrak{g}_{\mathbf{E}}^{\rm iw}(z_i)|^2}{\lambda_i/(1 + \alpha_s(\lambda_i - 1))}, \quad z_i = \lambda_i - {\rm i}N^{-1/2}, \tag{9.1.4}$$

where $\alpha_s = 1/(1+2q\kappa)$ and $\mathfrak{g}_{\mathbf{E}}^{\rm iw}$ is given in Eq. (4.2.33). We give a more precise implementation of this "IW-regularization" in the Algorithm 1, and a numerical illustration for an Inverse Wishart matrix (3.1.58) with parameter $\kappa = 10$ and $q = 0.5$, for which $\alpha_s \approx 0.09$. The results are plotted in Figure 9.1.2 where the empirical points come from a single simulation with $N = 500$.

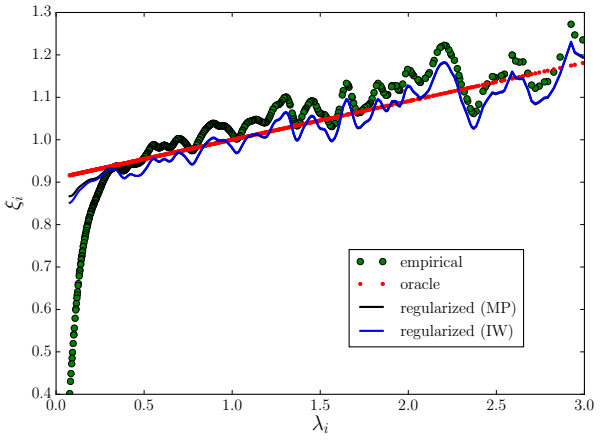



FIGURE 9.1.2. We apply the IW-regularization $\widehat{\xi}/\widehat{\xi}^N(z)$ with $z = \lambda - {\rm i}N^{-1/2}$ in the case where $\mathbf{C}$ is an Inverse-Wishart matrix with $\kappa = 10$ and $q = 0.5$. The finite size effect of the empirical RIE (7.5.2) (green points) is efficiently corrected. The red points correspond to the oracle estimator which is, in this case, the linear shrinkage procedure. We also compare the result of a "rescaled" Marčenko-Pastur spectrum, as proposed in [42].

[1] A yet simpler solution, proposed in [42] is to consider a rescaled Marčenko-Pastur's spectrum in such a way to fit the smallest eigenvalue $\lambda_N$. This is indistinguishable from the IW procedure when $\kappa$ is large enough, and provides very accurate predictions for US stocks return [42]. Nevertheless, in the presence of very small "true" eigenvalues, corresponding to of e.g. very strongly correlated financial contracts, this simple recipe fails.

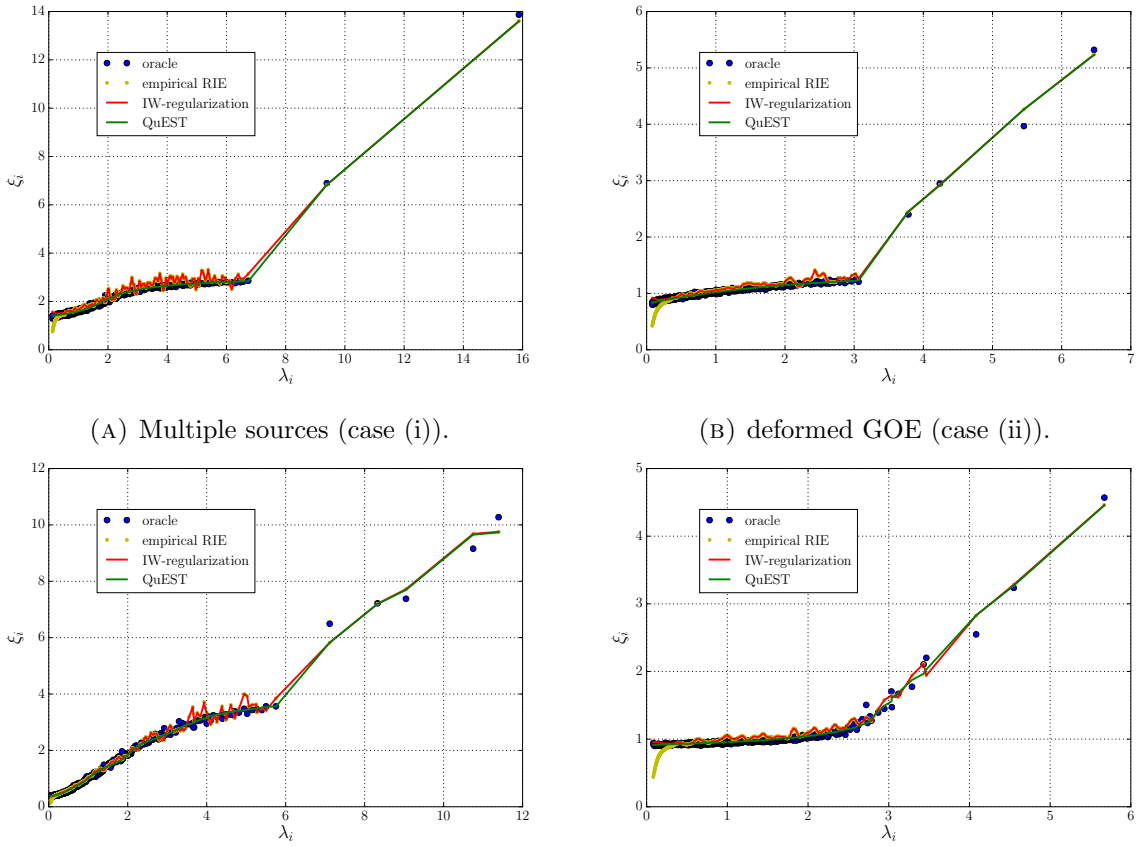
**Algorithm 1** IW-regularization of the empirical RIE (7.5.2)

**function** G_IW$(z, q, \kappa)$:
  $\lambda_\pm \leftarrow \big[(1+q)\kappa + 1 \pm \sqrt{(2\kappa+1)(2q\kappa+1)}\,\big]/\kappa$;
  **return** $\big[z(1+\kappa) - \kappa(1-q) - \sqrt{z-\lambda_+}\sqrt{z-\lambda_-}\big]/(z(z+2q\kappa))$;
**end function**

**function** RIE$(z, q, g)$:
  **return** $\mathrm{Re}[z]/|1-q+qzg|^2$;
**end function**

**function** DENOISING_RIE$(N, q, \{\lambda_i\}_{i=1}^N)$: //$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_N$
  $\kappa \leftarrow 2\lambda_N/\big((1-q-\lambda_N)^2 - 4q\lambda_N\big)$;
  $\alpha \leftarrow 1/(1+2q\kappa)$;
  **for** $i = 1$ to $N$ **do**
    $z \leftarrow \lambda_i - \mathrm{i}N^{-1/2}$;
    $g \leftarrow \big(\sum_{j\neq i}^N 1/(z-\lambda_j)\big)/(N-1)$;
    $\xi_i \leftarrow$ RIE$(z, q, g)$;
    $g \leftarrow$ G_IW$(z, q, \kappa)$
    $\Gamma_i \leftarrow (1+\alpha(\lambda_i - 1))/$RIE$(z, q, g)$;
    **if** $\Gamma_i > 1$ and $\lambda_i < 1$ **then**
      $\xi_i \leftarrow \Gamma_i \xi_i$;
    **end if**
  **end for**
  $s \leftarrow \sum_i \lambda_i / \sum_i \xi_i$; //preserve the trace
  **return** $\{s \times \xi_i\}_{i=1}^N$
**end function**

We now reconsider the numerical examples given in Section 7.5, for which we apply the IW-regularization algorithm (1). The results are plotted in Figure 9.1.3 and we observe that this IW-regularization works perfectly for all four population eigenvalues we consider in our simulations. Indeed, if we look at the left edge region, the regularized eigenvalues have been shifted upwards to coincide with the oracle estimator (blue points) while we had a significant deviation for the fully empirical, bare estimator (green dots). Hence, the IW-regularization (Algorithm 1) provides a very simple way to correct this systematic downside bias which is of crucial importance whether we need to invert the covariance matrix. Note that we can further improve the result by sorting the regularized eigenvalues. This is justified by the fact that we expect the RIE to be monotone with respect to the sample eigenvalues in the limit $N \to \infty$. We will investigate this point numerically in the next section (see Table 9.1).




(A) Multiple sources (case (i)).

(B) deformed GOE (case (ii)).

(C) Toeplitz (case (iii))

(D) Power law (case (iv))

FIGURE 9.1.3. Comparison of the IW-regularization (7.5.2) (red line) with the empirical RIE (7.5.2) (yellow dots) and the oracle estimator (7.1.2) (blue points) for the four cases presented at the beginning of Section 7.5 with $N = 500$ and $T = 1000$. We also plot the estimation we get using QuEST estimator (9.1.10) (green line). The results come from a single realization of $\mathbf{E}$ using a multivariate Gaussian measurement process.

**9.1.3. Quantized Eigenvalues Sampling Transform (QuEST).** An alternative method, recently proposed by Ledoit and Wolf [118] to approximate numerically the optimal RIE (7.2.2), is to work with the Marčenko-Pastur equation (4.2.1). It is somewhat similar to the numerical scheme proposed by N. ElKaroui (see Section 8.2.3) to solve the indirect problem of the Marčenko-Pastur equation.

The method, named as QuEST (Quantized Eigenvalues Sampling Transform), is based on a quantile representation of the eigenvalues. More formally, the key assumption is that the empirical eigenvalues are allocated smoothly according to the quantile of the spectral distribution, i.e.

$$\frac{i}{N} = \int_{-\infty}^{\lambda_i} \rho_{\mathbf{E}}(x)\mathrm{d}x, \tag{9.1.5}$$

and the aim is to find the quantile, as a function of the population eigenvalues $[\mu_i]_i$, such that (9.1.5) holds. Note that the representation (9.1.5) is nothing less than the definition of the classical location of the *bulk* eigenvalues, encountered in Eq. (4.2.32). Hence, for $N \to \infty$, this method does not seem to be appropriate for outliers as we know that the spectral density $\rho_{\mathbf{E}}$ puts no weights on outliers. Nevertheless, for contructing RIEs, this might not be that important since, roughly speaking, all we need to know is the Stieltjes transform of the spikeless covariance matrix $\underline{\mathbf{E}}$ (see Section 7.2.2). That being said, the "quantized" eigenvalues, expected to be close to the empirical eigenvalues, are defined as

$$\tilde{\gamma}_i(\boldsymbol{\mu}) \;:=\; N\int_{(i-1)/N}^{i/N} F_{\mathbf{E}}^{-1}(p)\mathrm{d}p, \qquad i \in [\![1, N]\!],\; p \in [0,1], \tag{9.1.6}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$, and

$$\begin{aligned} F_{\mathbf{E}}^{-1}(p) \;&:=\; \sup\Big\{x \in \mathbb{R} \,:\, F_{\mathbf{E}}(x) \leqslant p\Big\}, \\ F_{\mathbf{E}}(x) \;&:=\; \begin{cases} \max\Big(1 - 1/q, N^{-1}\sum_{i=1}^N \delta_0(\mu_i)\Big) & \text{if } x = 0, \\ \int_0^x \rho_{\mathbf{E}}(u)\mathrm{d}u, & \text{otherwise,} \end{cases} \end{aligned} \tag{9.1.7}$$

with $\rho_{\mathbf{E}}(u) = \lim_{\eta\downarrow 0} \operatorname{Im} \mathfrak{g}_{\mathbf{E}}^N(u - \mathrm{i}\eta)$ and $\mathfrak{g}_{\mathbf{E}}^N$ is the unique solution in $\mathbb{C}^+$ of the discretized Marčenko-Pastur equation (4.2.3)

$$\mathfrak{g}_{\mathbf{E}}^N(z) = \frac{1}{N}\sum_{i=1}^N \frac{1}{z - \mu_i(1 - q + qz\mathfrak{g}_{\mathbf{E}}^N(z))}. \tag{9.1.8}$$

Even if the numerical scheme seems quite intricated, all these quantities are simply a discretized version of the Marčenko-Pastur equation. Indeed, Eq. (9.1.5) is equivalent to Eq. (4.2.3) for large $N$ and (9.1.6) is nothing but a discrete estimator of Eq. (4.2.32).

Finally, the optimization program reads

$$\tilde{\boldsymbol{\mu}} \;:=\; \begin{cases} \operatorname{argmin}_{\boldsymbol{\mu}\in\mathbb{R}_+^N} \sum_{i=1}^N \Big[\tilde{\gamma}_i(\boldsymbol{\mu}) - \lambda_i\Big]^2, \\ \text{s.t.} \;\; \tilde{\gamma}_i(\boldsymbol{\mu}) \;\; \text{satisfies Eqs. (9.1.6), (9.1.7) and (9.1.8).} \end{cases} \tag{9.1.9}$$

From there, the regularization scheme of the empirical RIE (7.5.2) reads

$$\xi_i^{\mathrm{QuEST}} = \frac{\lambda_i}{|1 - q + q\lambda_i \lim_{\eta\downarrow 0} \tilde{\mathfrak{g}}_{\mathbf{E}}^N(\lambda_i - \mathrm{i}\eta)|^2}, \tag{9.1.10}$$

where $\tilde{\mathfrak{g}}_{\mathbf{E}}^N(z) \in \mathbb{C}^+$ is the unique solution of

$$\tilde{\mathfrak{g}}_{\mathbf{E}}^N(z) = \frac{1}{N}\sum_{i=1}^N \frac{1}{z - \tilde{\mu}_i(1 - q + qz\tilde{\mathfrak{g}}_{\mathbf{E}}^N(z))}. \tag{9.1.11}$$

We emphasize that details about the implementation of QuEST are given in [118]. We see that the above regularization scheme allows – in principle – to estimate the limiting RIE (7.2.2)

since we can now set $\eta$ to be arbitrarily small. This means that, contrary to the empirical estimate (7.5.2), the QuEST procedure should not suffer from a systematic underestimation at the left edges. The main advantage of this method is that it also allows us to estimate the population eigenvalues, which can be useful in some particular cases. However, from a numerical standpoint, this algorithm is far more complicated to implement than the above IW-regularization (Algorithm (1)). Indeed, we see that the starting point of the optimization (9.1.9) is the vector of population eigenvalues, which can be problematic for very "diluted" spectrum. Moreover, the algorithm might suffer from instabilities in the presence of very large and isolated eigenvalues. Note that a detailed presentation for the implementation is given in [118], where the authors also advise to sort the cleaned eigenvalues $\{\xi_i^{\text{QuEST}}\}_{i\leqslant N}$ since, as said above, we expect the optimal cleaned eigenvalues to be monotonic with respect to the sample eigenvalues.

**9.1.4. Empirical studies.** We compare the above QuEST numerical scheme with the simple IW-regularization of Section 9.1.2 and the results are plotted in Figure 9.1.3. The eigenvalues coming from the QuEST regularization are depicted by the green line and we see that the results are very satisfactory. In particular, it indeed does not suffer from the systematic bias in the left edge and seems to handle efficiently outliers even if the formula (9.1.5) is a priori not valid for isolated eigenvalues in the large $N$ limit. We nonetheless notice that the algorithm suffers sometimes from instabilities in the presence of "clustered" outliers as seen in the power law example (see Figure 9.1.3d). On the other hand, and perhaps surprisingly, the ad-hoc IW-regularization given in Algorithm 1 provides very similar result to the complicated – but theoretically better founded – QuEST procedure. Nonetheless, this latter method requires to solve a nonlinear and non-convex optimization problem (see Eq. (9.1.9)) which implies heavy numerical computations and may not even converge to the global minimum (if it exists).

We want to further investigate the efficiency of these two regularizations. One direction is to change the number of variables $N$ with $q = 0.5$ fixed. This allows us to assess the finite size performance of the two algorithms. The second direction is to fix $N = 500$ and vary the observation ratio $q$. We shall consider three different regularizations in the following: (i) IW-regularization (Algorithm 1), (ii) IW-regularization + sorting (name "IWs regularization" in the following) and (iii) QuEST procedure. Note that we will focus our study on the power law example of Figure 9.1.3d since this simple prior allows use to generate very complex spectrum with possibly "clustered" outliers. We emphasize the the regularization scheme (ii) is justified by the fact that we expect the estimator to preserve the monotonicity of the sample eigenvalues.

To measure the accuracy and the stability of each algorithm, we characterize the deviation between a given estimator and the Oracle (7.1.2). Using the mean squared error (MSE), we may also analyze the relative performance (RP) in percentage compared to the sample covariance. This is given by

$$\text{RP}(\Xi) \;:=\; 100 \times \left(1 - \frac{\mathbb{E}\|\Xi - \Xi^{\text{ora.}}\|_2}{\mathbb{E}\|\mathbf{E} - \Xi^{\text{ora.}}\|_2}\right), \tag{9.1.12}$$

where $\Xi \equiv \Xi(\mathbf{E})$ is a RIE of $\mathbf{C}$ and $\Xi^{\text{ora.}}$ is the oracle estimator. We also report in each case the average computational time needed to perform the estimation[2].

First, let us assess the usefulness of sorting the cleaned eigenvalues. We report in Table 9.1 the performance we obtained for $N = 500$ and $q = 0.5$ fixed over 100 realizations of $\mathbf{E}$ (which is a Wishart matrix with population covariance matrix $\mathbf{C}$). We conclude from Table 9.1 that it is indeed better to sort the eigenvalues when using the IW-regularization (9.1.4) as the difference is

[2]Simulations are based on an Intel® Core™ i7-4700HQ and CPU of 8 × 2.40 GHz processor.

statistically significant, while being nearly equally efficient in terms of computational time. For large $N$, the QuEST procedure yields the best accuracy score but the difference with the IWs eigenvalues is not significant and the QuEST requires much more numerical operations than the ad-hoc IWs algorithm. Note that the performance improvement over to the sample covariance is very substantial.

TABLE 9.1. We reconsider the setting of Figure 9.1.3d and check the consistency over 100 samples. The population density $\rho_{\mathbf{C}}$ is drawn from (7.5.4) with $\lambda_0 = -0.6$ and $N = 500$ and the sample covariance matrix is obtained from the Wishart distribution. MSE stands for the mean squared error with respect to the oracle estimator (7.1.2), stdev stands for the standard deviation of the squared error and the RP defined in Eq. (9.1.12). Running time shows the average time elapsed for the cleaning of one sample set of eigenvalues of size $N$.

| Method | MSE | stdev | RP | Running time (sec) |
|---|---|---|---|---|
| IW-regularization | 0.64 | 0.13 | 99.69 | 0.02 |
| IWs-regularization | 0.45 | 0.12 | 99.78 | 0.03 |
| QuEST | 0.44 | 0.15 | 99.79 | 33.5 |

We now investigate how these conclusions change when $N$ varies with $q = 0.5$ fixed. The results are given in Table 9.2. First, we stress that the RP with respect to the sample covariance matrix is already greater than 98% for $N = 100$ which is why we did not report these values in the table. As above, for any value of $N \geqslant 100$, it appears that sorting the eigenvalues improves significantly the mean squared error with respect to the oracle estimator. We also emphasize that for $N = 1000$, it takes 0.06 seconds to get the dressed RIE while the QuEST algorithm requires 80 seconds in average. We see that as the size $N$ grows to infinity, the high degree of complexity needed to solve the nonlinear and non-convex optimization (9.1.9) becomes very restrictive, while improvement over the simple IWs method is no longer significant.

TABLE 9.2. Check of the consistency of the three regularizations with respect to the dimension $N$. The population density $\rho_{\mathbf{C}}$ is drawn from (7.5.4) with $\lambda_0 = -0.6$ and the sample covariance matrix is obtained from the Wishart distribution with $T = 2N$. We report in the table the mean squared error with respect to the oracle estimator (7.1.2) and the standard deviation in parenthesis as a function of $N$.

| Method | $N = 100$ | $N = 200$ | $N = 300$ | $N = 400$ | $N = 500$ | $N = 1000$ |
|---|---|---|---|---|---|---|
| IW-regularization | 0.53 (0.17) | 0.56 (0.15) | 0.64 (0.16) | 0.65 (0.14) | 0.64 (0.14) | 0.74 (0.14) |
| IWs-regularization | 0.35 (0.14) | 0.39 (0.14) | 0.45 (0.14) | 0.45 (0.13) | 0.46 (0.12) | 0.53 (0.12) |
| QuEST | 0.26 (0.16) | 0.33 (0.15) | 0.39 (0.15) | 0.4 (0.15) | 0.45 (0.15) | 0.5 (0.13) |

We now look at the second test in which $N = 500$ is fixed and we vary $q = 0.25, 0.5, 0.75, 0.95$. For each $q$, we perform the same procedure than Table 9.2 and the results are reported in Table 9.3. It is easy to see that the conclusions of the first consistency test are still valid for the three regularization schemes as a function of $q$ with $N = 500$. Note that we do not consider here

the case $q \geqslant 1$ which is less immediate more complicated since $\mathbf{E}$ generically possess $(N - T)$ zero eigenvalues. Both regularization schemes, IWs-regularization and QuEST algorithm, fail to handle this case and we shall come back to this problem in Chapter 10.

TABLE 9.3. Check of the consistency of the three regularizations with respect to the dimension ratio $q$. The population density $\rho_{\mathbf{C}}$ is drawn from (7.5.4) with $\lambda_0 = -0.6$ and $N = 500$ and the sample covariance matrix is obtained from the Wishart distribution with parameter $T = N/q$. We report in the table the mean squared error with respect to the oracle estimator (7.1.2) and the standard deviation in parenthesis as a function of $q$.

| Method | $q = 0.25$ | $q = 0.5$ | $q = 0.75$ | $q = 0.95$ |
|---|---|---|---|---|
| IW-regularization | 0.31 (0.06) | 0.65 (0.14) | 1.2 (0.18) | 1.78 (0.44) |
| IWs-regularization | 0.28 (0.05) | 0.46 (0.12) | 0.71 (0.17) | 0.94 (0.39) |
| QuEST | 0.25 (0.05) | 0.45 (0.15) | 0.72 (0.17) | 0.98 (0.35) |

To conclude, we observed through these examples with synthetic data that we are able to estimate accurately the oracle estimator for bounded $N$ both for small eigenvalues and outliers. The QuEST procedure is found to behave efficiently for any $N$ and any $q < 1$, and allows one to estimate the both the population eigenvalues and the limiting Stieltjes transform with high precision. However, as far as the estimation of large sample covariance matrices is concerned, the improvement obtained by solving the nonlinear and non-convex optimization problem (9.1.9) becomes insignificant as $N$ increases (see Tables 9.2 and 9.3). Furthermore, the computational time of the QuEST algorithm increases considerably as $N$ grows. We shall henceforth use the IWs RIE as our estimator of $\mathbf{C}$ for the applications below. Nonetheless, whenever $N$ is not very large, the QuEST procedure is clearly advised as it yields a significant improvement with an acceptable computational time.

## 9.2 Optimal RIE and out-of-sample risk for optimized porfolios

As alluded to above (see Section 8.1), the concept of correlations between different assets is a cornerstone of Markowitz' optimal portfolio theory, especially for risk management purposes [126]. It is therefore of crucial importance to use a correlation matrix that faithfully represents *future* risks, and not past risks – otherwise the over-allocation on spurious low risk combination of assets might prove disastrous. In that respect, we saw in Section 8.1.3 that the best estimator inside the space of estimators restricted to possess the sample eigenvectors is precisely the oracle estimator (7.1.2) which is not observable a priori. However, if the number of variables is sufficiently large, we know – thanks to the numerical study of the previous section – that it is possible to estimate very accurately the oracle estimator using only observable variables. The main objective in the present section is to investigate the IWs RIE procedure for financial stock market data.

Let us now explain the construction of our test. We consider a universe made of $N$ different financial assets – say stocks – that we observe at – say – the daily frequency, defining a vector of returns $\mathbf{r}_t = (r_{1t}, r_{2t}, \ldots, r_{Nt})$ for each day $t = 1, \ldots, T$. It is well known that volatilities

of financial assets are heteroskedastic [30] and we therefore focus specifically on *correlations* and not on volatilities in order to study the systemic risk. To that end, we standardize these returns as follows: (i) we remove the sample mean of each asset; (ii) we normalize each return by an estimate $\widehat{\sigma}_{it}$ of its daily volatility: $\widetilde{r}_{it} = r_{it}/\widehat{\sigma}_{it}$. There are many possible choices for $\widehat{\sigma}_{it}$, based e.g. on GARCH or FIGARCH models historical returns, or simply implied volatilities from option markets, and the reader can choose his/her favourite estimator which can easily be combined with the correlation matrix cleaning schemes discussed below. For simplicity, we have chosen here the cross-sectional daily volatility, that is

$$\widehat{\sigma}_{it} := \sqrt{\sum_j r_{jt}^2}\,, \tag{9.2.1}$$

to remove a substantial amount of non stationarity in the volatilities. The final standardized return matrix $\mathbf{Y} = (Y_{it}) \in \mathbb{R}^{N\times T}$ is then given by $Y_{it} := \widetilde{r}_{it}/\sigma_i$ where $\sigma_i$ is the sample estimator of the $\widetilde{r}_i$ which is now, to a first approximation, stationary.

We may now compute construct the sample covariance matrix $\mathbf{E}$ as in Eq. (4.1.3). We stress that the Marčenko and Pastur result does not require multivariate normality of the returns, which can have fat-tailed distributions. In fact, the above normalisation by the cross-sectional volatility can be seen as a proxy for a robust estimator of the covariance matrix (4.1.8) with $U(x) = x^{-1}$ which can be studied using the tools of Chapters 4 and 5 (see Section 4.1.3 for a discussion on this point). All in all, we are able to construct the optimal RIE either using IWs-regularization (Algorithm 1 + sorting) or the QuEST regularization, the latter allowing us to estimate the population eigenvalue spectrum as well.

For our simulations, we consider an international pools of stocks with daily data:

(i) US: 500 most liquid stocks during the training period of the S&P 500 from 1966 until 2012;

(ii) Japan: 500 most liquid stocks during the training period of the all-shares TOPIX index from 1993 until 2016;

(iii) Europe: 500 most liquid stocks during the training period of the Bloomberg European 500 index from 1996 until 2016.

We chose $T = 1000$ (4 years) for the training period, i.e. $q = 0.5$, and $T_{\text{out}} = 60$ (three months) for the out-of-sample test period. Let us first analyze the optimal RIE for US stocks and we plot in Figure 9.2.1 the average nonlinear shrinkage curve for the IWs-regularization (blue line) and for the QuEST regularization (red dashed line) – where we sort the eigenvalues in both cases – and compare it with the estimated population eigenvalues obtained from (9.1.9). We see that IWs-regularization and QuEST still yields very similar results. Furthermore, we notice that the spectrum of the cleaned eigenvalues is as expected narrower than the spectrum of the population eigenvalues.

Interestingly, the oracle estimator (7.1.2) can be estimated empirically and used to directly test the accuracy of the IWs-regularized RIE (9.1.4). The trick is to remark that the oracle eigenvalues (7.1.2) can be interpreted as the "true" (out-of-sample) risk associated to a portfolio whose weights are given by the $i$-th eigenvector. Hence, assuming that the data generating process is stationary, we estimate the oracle estimator through the realized risk associated to such eigen-portfolios [141]. More precisely, we divide the total length of our time series $T_{\text{tot}}$ into

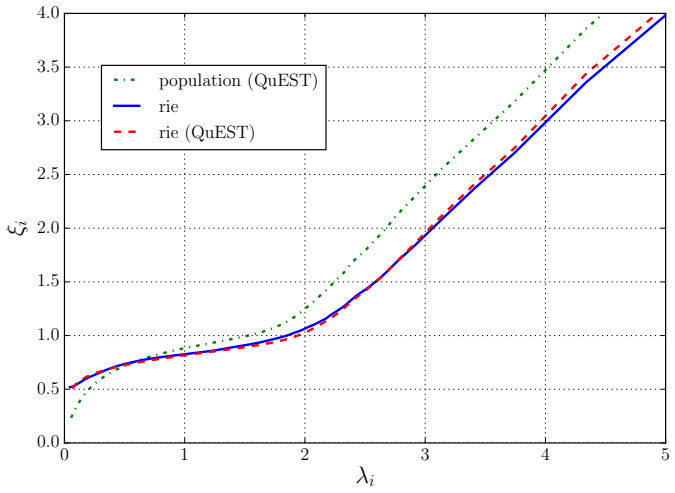



FIGURE 9.2.1. Comparison of the IWs-regularization (9.1.4) (blue) with the QuEST procedure (9.1.10) (red dashed line) using 500 US stocks from 1970 to 2012. The agreement between those two regularizations is quite remarkable. We also provide the estimation of the population eigenvalues obtained from (9.1.9) (green dashed-dotted line).

$n$ consecutive, non-overlapping samples of length $T_{\text{out}}$. The "training" period has length $T$, so $n$ is given by:

$$n := \lfloor \frac{T_{\text{tot}} - T - 1}{T_{\text{out}}} \rfloor. \tag{9.2.2}$$

The oracle estimator (7.1.2) is then computed as:

$$\xi_i^{\text{ora.}} \approx \frac{1}{n} \sum_{j=0}^{n-1} \mathcal{R}^2_{\text{out}}(t_j, \mathbf{u}_i) \quad i = 1, \ldots, N, \tag{9.2.3}$$

for $t_j = T + j \times T_{\text{out}} + 1$ and $\mathcal{R}(t, \mathbf{w})$ denotes the out-of-sample variance of the returns of portfolio $\mathbf{w}$ built at time $t$, that is to say

$$\mathcal{R}^2_{\text{out}}(t, \mathbf{w}) := \frac{1}{T_{\text{out}}} \sum_{\tau=t+1}^{t+T_{\text{out}}} \left( \sum_{i=1}^{N} \mathbf{w}_i Y_{i\tau} \right)^2, \tag{9.2.4}$$

where $Y_{i\tau}$ denotes the rescaled realized returns. Again, as we are primarily interested in estimating correlations and not volatilities, both our in-sample and out-of-sample returns are made approximately stationary and normalized. This implies that $\sum_{i=1}^{N} \mathcal{R}^2_{\text{out}}(t, \mathbf{u}_i) = N$ for any time $t$. We plot our results for the estimated oracle estimator (9.2.3) using US data in Fig 9.2.2 that we compare with the IWs-regularized RIE. The results are, we believe, quite remarkable: the RIE formula (9.1.4) (red dashed line) tracks very closely the average realized risk (blue triangles), specially in the region where there is a lot of eigenvalues.

We may now repeat the analysis for the other pools of stocks as well. We begin with the TOPIX where we plot in Figure 9.2.3a the estimation of the population (using Eq. (9.1.9)) and the regularized RIE (using Algorithm 1 or Eq. (9.1.10)). Again, the results we get from the

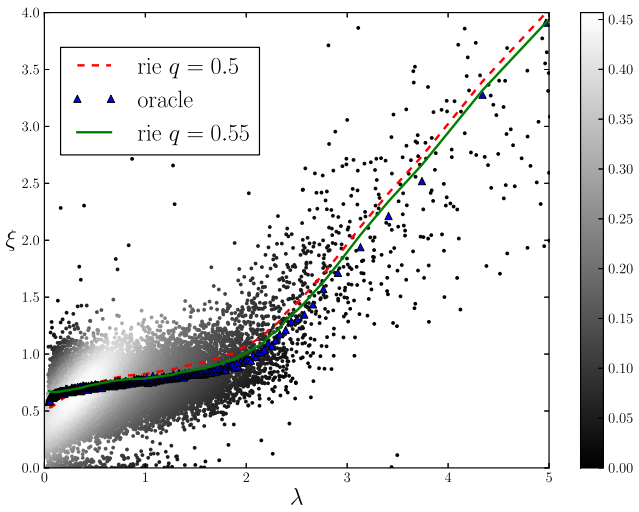
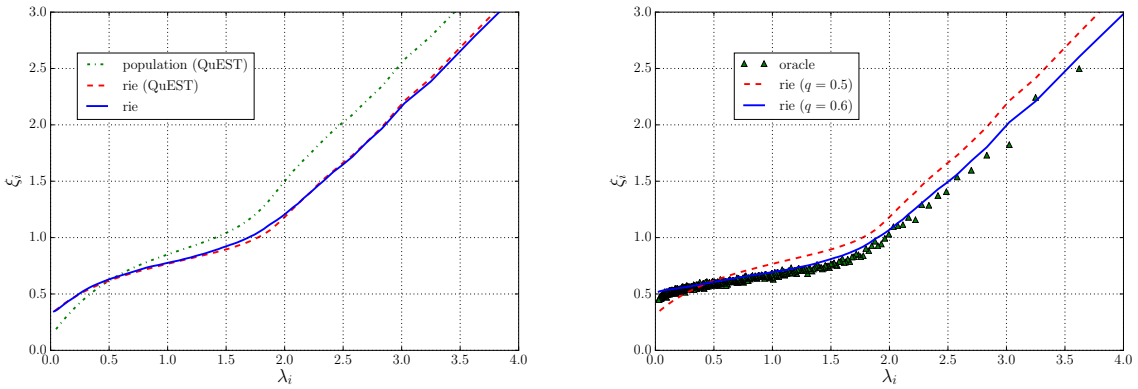



FIGURE 9.2.2. Comparison of the IWs-regularized RIE (9.1.4) with the proxy (9.2.3) using 500 US stocks from 1970 to 2012. The points represent the density map of each realization of (9.2.3) and the color code indicated the density of data points. The average IWs-regularized RIE is plotted with the red dashed line and the average realized risk in blue. We also provide the prediction of the IWs-regularized RIE with an effective observation ratio $q_{\text{eff}}$ which is slightly bigger than $q$ (green plain line). The agreement between the green line and the average oracle estimator (blue triangle) is quite remarkable.

simple IWs-regularization and QuEST procedure are nearly indistinguishable. This is another manifestation of the robustness of both algorithms at a finite $N$. We then plot in Figure 9.2.3b the comparison between the IWs-regularized RIE (red dashed line) and the Oracle estimator, approximated by (9.2.3) (green triangles). We observe that the overall estimation is not as convincing as for US stocks (Figure 9.2.2) but as above, we notice that the deviation may be explained by the presence of autocorrelations. Indeed, there also exists an effective ratio $q_{\text{eff}} = 1.2q$ such that the estimation is extremely good (see blue line in Figure 9.2.3b).

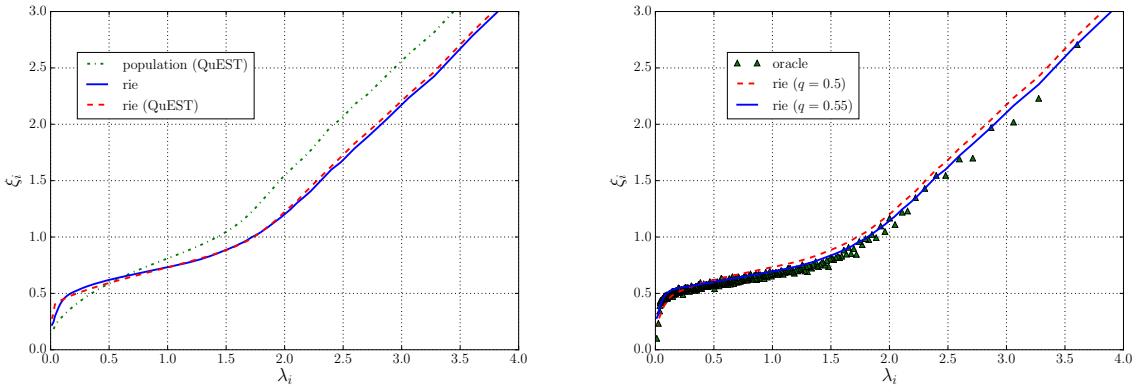
Finally we look at European stocks where the conclusion are similar than for the US stocks. In particular, we notice in Figure 9.2.4b that the estimation we obtained for the IWs-regularized RIE with the observed $q = 0.5$ (red dashed line) yields a very good approximation of the Oracle estimator (green triangle). We can nonetheless improve the estimation with an effective ratio $q_{\text{eff}} = 1.1q$ (blue plain line).

All in all, we see that both the simple IWs-regularization and the QuEST regularization allow one to estimate accurately the (approximated) Oracle estimator using only observables variables. This study highlights that the optimal RIE estimator is robust regarding the data generating process, as financial stock markets are certainly not Gaussian. Note that this last observation is not valid anymore if we use the historical estimator of the variance as a measure of volatility as the data are still highly influenced by heteroskedastic effects. Still, the cross sectional volatility estimator (9.2.1) does not remove entirely the temporal dependence of the variables since it appears that one can choose an *effective* observation ratio $q_{\text{eff}} > q$ for which the IWs-regularized RIE and the Oracle estimate nearly coincide. This effect may be understood by the presence of autocorrelations in the stock returns that are not taken into account in the




(A) Population and optimal RIE bulk eigenvalues. (B) Comparison with Oracle estimator (9.2.3).

FIGURE 9.2.3. Left figure: analysis of the population (green dashed line) and optimal RIE bulk eigenvalues (red dashed line for Eq. (9.1.10) and blue plain line for the IWs-regularization) using the 500 most liquid stocks during the training period of the all-shares TOPIX index from 1993 until 2016. Right figure: Comparison between the IWs-regularized RIE (red dashed line) with the Oracle estimator (9.2.3) (green triangle). We also provide the plot of the IWs-regularized RIE with an effective observation ratio (blue line).




(A) Population and optimal RIE bulk eigenvalues. (B) Comparison with Oracle estimator (9.2.3).

FIGURE 9.2.4. Left figure: analysis of the population (green dashed line) and optimal RIE bulk eigenvalues (red dashed line for Eq. (9.1.10) and blue plain line for the IWs-regularization) using the 500 most liquid stocks during the training period of the Bloomberg European 500 index from 1996 until 2016. Right figure: Comparison between the IWs-regularized RIE (red dashed line) with the Oracle estimator (9.2.3) (green triangle). We also provide the plot of the IWs-regularized RIE with an effective observation ratio (blue line)

model of $\mathbf{E}$. The presence of autocorrelations has been shown to widen the spectrum of the sample matrix $\mathbf{E}$ [48]. Since the agreement reached with the naive value $q = N/T$ is already very good, we shall come back to the open problem of calibrating $q_{\text{eff}}$ on empirical data in the Chapter 10.

## 9.3 Out-of-sample risk minimization

It is interesting to compare the optimal shrinkage function that maps the empirical eigenvalue $\lambda_i$ onto its "cleaned" RIE counterpart $\xi_i$. We show these functions in Figure 9.3.1 for the three schemes we retained here, i.e. linear shrinkage, clipping and RIE, using the same data set as in Figure 9.2.2. This figure clearly reveals the difference between the three schemes. For clipping (red dashed line), the intermediate eigenvalues are quite well estimated but the convex shape of the optimal shrinkage function for larger $\lambda_i$'s is not captured. Furthermore, the larger eigenvalues are systematically overestimated. For the linear shrinkage (green dotted line), it is immediate from Figure 9.3.1 why this method is not optimal for any shrinkage parameters $\alpha_s \in [0, 1]$ (that fixes the slope of the line).

We now turn to optimal portfolio construction using the above three cleaning schemes, with the aim of comparing the (average) realized risk of optimal Markowitz portfolios constructed as:

$$\mathbf{w} := \frac{\widehat{\mathbf{\Sigma}}^{-1}\mathbf{g}}{\mathbf{g}^*\widehat{\mathbf{\Sigma}}^{-1}\mathbf{g}}, \tag{9.3.1}$$

where $\mathbf{g}$ is a vector of *predictions* and $\widehat{\mathbf{\Sigma}}$ is the cleaned covariance matrix $\widehat{\Sigma}_{ij} := \sigma_i\sigma_j\widehat{\Xi}_{ij}$ for any $i, j \in [\![1, N]\!]$. Note again that we consider here returns normalized by an estimator of their volatility: $\widetilde{r}_{it} = r_{it}/\widehat{\sigma}_{it}$. This means that our tests are immune against an overall increase or decrease of the the volatility in the out-of-period, and are only sensitive to the quality of the estimator of the correlation matrix itself.

In order to ascertain the robustness of our results in different market situations, we consider the following four families of predictors $\mathbf{g}$:

(i) The minimum variance portfolio, corresponding to $g_i = 1$, $\forall i \in [\![1, N]\!]$

(ii) The omniscient case, i.e. when we know exactly the realized returns on the next out-of-sample period for each stock. This is given by $g_i = \mathcal{N}\,\tilde{r}_{i,t}(T_{\text{out}})$ where $r_{i,t}(\tau) = (P_{i,t+\tau} - P_{i,t})/P_{i,t}$ with $P_{i,t}$ the price of the $i$th asset at time $t$ and $\widetilde{r}_{it} = r_{it}/\widehat{\sigma}_{it}$.

(iii) Mean-reversion on the return of the last day: $g_i = -\mathcal{N}\,\widetilde{r}_{it}$ $\forall i \in [\![1, N]\!]$.

(iv) Random long-short predictors where $\mathbf{g} = \mathcal{N}\,\mathbf{v}$ where $\mathbf{v}$ is a random vector uniformly distributed on the unit sphere.

The normalisation factor $\mathcal{N} := \sqrt{N}$ is chosen to ensure $\mathbf{w}_i \sim \mathcal{O}(N^{-1})$ for all $i$. The out-of-sample risk $\mathcal{R}^2$ is obtained from Eq. (9.2.4) by replacing the matrix $\mathbf{X}$ by the normalized return matrix $\widetilde{\mathbf{R}}$ defined by $\widetilde{\mathbf{R}} := (\widetilde{r}_{it}) \in \mathbb{R}^{N\times T}$. We report the average out-of-sample risk for these various portfolios in Table 9.4, for the three above cleaning schemes and the three geographical zones, keeping the same value of $T$ (the learning period) and $T_{\text{out}}$ (the out-of-sample period) as above. The linear shrinkage estimator uses a shrinkage intensity $\alpha$ estimated from the data

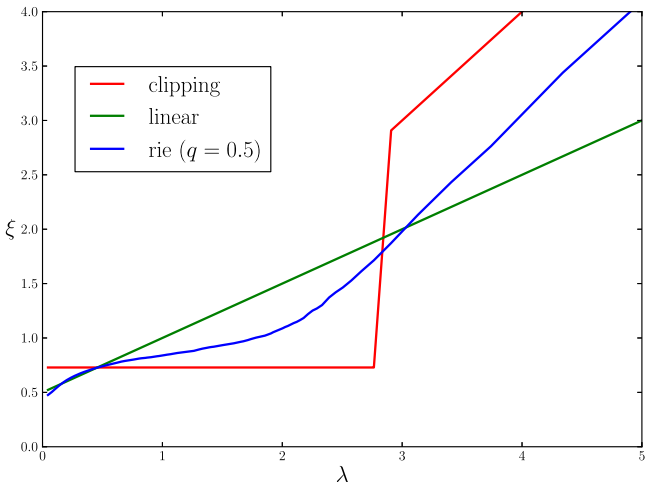



FIGURE 9.3.1. Comparison of the debiased RIE (9.1.4) (blue line) with clipping at the edge of the Marčenko-Pastur (red dashed line) and the linear shrinkage with $\alpha = 0.5$ (green dotted line). We use here the same data set as in Figure 9.2.2.

following [114] (LW). The eigenvalues clipping procedure uses the position of the Marčenko-Pastur edge, $(1 + \sqrt{q})^2$, to discriminate between meaningful and noisy eigenvalues. The second to last line gives the result obtained by taking the identity matrix (*total* shrinkage, $\alpha_s = 0$) and the last one is obtained by taking the uncleaned, in-sample correlation matrix ($\alpha_s = 1$).

These tables reveal that: (i) it is always better to use a cleaned correlation matrix: the out-of-sample risk without cleaning is, as expected, always higher than with any of the cleaning schemes, even with four years of data; (ii) in all cases but one (Minimum risk portfolio in Japan, where the LW linear shrinkage outperforms), the regularized RIE is providing the lowest out-of-sample risk, independently of the type of predictor used. Note that these results are statistically significant everywhere, except perhaps for the minimum variance strategy with Japanese stocks: see the standard errors that are given between parenthesis in Table 9.4. Finally, we test the robustness in the dimension $N$ by repeating the same test for $N = \{100, 200, 300\}$. We focus on relatively small values of $N$ as the conclusions are unchanged in all cases as soon as $N \geqslant 300$. We see that aside some fluctuations for $N = 100$, the result for out-of-sample test with the RIE is robust in the dimension $N$ as indicated in the Table 9.5.

## 9.4 Testing for stationarity assumption

In this section, we investigate in more details the stationarity assumption underlying the Marčenko-Pastur framework, i.e. that the future (out-of-sample) is statistically identical to the past (in-sample), in the sense that the empirical correlation matrices $\mathbf{E}_{\text{in}}$ and $\mathbf{E}_{\text{out}}$ are generated by the same underlying statistical process characterized by a unique correlation matrix $\mathbf{C}$. We will use the two-sample eigenvector test introduced in Section 5.2.

Let us reconsider the two-sample self-overlap formula (5.2.18) for which the key object is the *limiting* Stieltjes transform (5.2.13). As we saw in Section 9.1.2, using the "raw" empirical Stieltjes transform yields a systematic bias for small eigenvalues which can be problematic when

TABLE 9.4. Annualized average volatility (in %) of the different strategies. Standard deviations are given in bracket.

Minimum variance portfolio

| $\langle\mathcal{R}\rangle_e$ | US | Japan | Europe |
|---|---|---|---|
| RIE (IWs) | **10.4** (0.12) | 30.0 (2.9) | **13.2** (0.12) |
| Clipping MP | 10.6 (0.12) | 30.4 (2.9) | 13.6 (0.12) |
| Linear LW | 10.5 (0.12) | **29.5** (2.9) | 13.2 (0.13) |
| Identity $\alpha_s = 0$ | 15.0 (0.25) | 31.6 (2.92) | 20.1 (0.25) |
| In sample $\alpha_s = 1$ | 11.6 (0.13) | 32.3 (2.95) | 14.6 (0.2) |

Omniscient predictor

| $\langle\mathcal{R}\rangle_e$ | US | Japan | Europe |
|---|---|---|---|
| RIE (IWs) | **10.9** (0.15) | **12.1** (0.18) | **9.38** (0.18) |
| Clipping MP | 11.1 (0.15) | 12.5 (0.2) | 11.1 (0.21) |
| Linear LW | 11.1 (0.16) | 12.2 (0.18) | 11.1 (0.22) |
| Identity $\alpha_s = 0$ | 17.3 (0.24) | 19.4 (0.31) | 17.7 (0.34) |
| In sample $\alpha_s = 1$ | 13.4 (0.25) | 14.9 (0.28) | 12.1 (0.28) |

Mean reversion predictor

| $\langle\mathcal{R}\rangle_e$ | US | Japan | Europe |
|---|---|---|---|
| RIE (IWs) | **7.97** (0.14) | **11.2** (0.20) | **7.85** (0.06) |
| Clipping MP | 8.11 (0.14) | 11.3 (0.21) | 9.35 (0.09) |
| Linear LW | 8.13 (0.14) | 11.3 (0.20) | 9.26 (0.09) |
| Identity $\alpha_s = 0$ | 17.7 (0.23) | 24.0 (0.4) | 23.5 (0.2) |
| In sample $\alpha_s = 1$ | 9.75 (0.28) | 15.4 (0.3) | 9.65 (0.11) |

Uniform predictor

| $\langle\mathcal{R}\rangle_e$ | US | Japan | Europe |
|---|---|---|---|
| RIE | **1.30** (8e-4) | **1.50** (1e-3) | **1.23** (1e-3) |
| Clipping MP | 1.31 (8e-4) | 1.55 (1e-3) | 1.32 (1e-3) |
| Linear LW | 1.32 (8e-4) | 1.61 (1e-3) | 1.27 (1e-3) |
| Identity $\alpha_s = 0$ | 1.56 (2e-3) | 1.86 (2e-3) | 1.69 (2e-3) |
| In sample $\alpha_s = 1$ | 1.69 (1e-3) | 2.00 (2e-3) | 2.7 (0.01) |

applying Eq. (5.2.18). Hence, we shall split the numerical computation of the overlap formula (5.2.17) or (5.2.18) into two steps. The first step is to estimate the population eigenvalues using the QuEST method of Ledoit and Wolf (see Section 9.1.3). Since these eigenvalues are designed to solve the Marčenko-Pastur equation, the second step consists in extracting from Eq. (9.1.8) an estimation of the Stieltjes transform of $\mathbf{E}$ for an arbitrarily small imaginary part $\eta$, that we denote by $\widehat{\mathfrak{g}}_{\mathbf{E}}(z)$ for any $z \in \mathbb{C}_-$. Using $\widehat{\mathfrak{g}}_{\mathbf{E}}(z)$ in Eq. (5.2.13) allows us to obtain the overlaps.

**9.4.1. Synthetic data.** We test this procedure on synthetic data first. Our numerical procedure is a follows. As in Section 5.2, we consider 100 independent realization of the Wishart noise $\boldsymbol{\mathcal{W}}$ with parameter $T$ and covariance $\mathbf{C}$. Then, for each pair of samples, we compute the smoothed

TABLE 9.5. Annualized average volatility (in %) of the different strategies as a function of $N$ with $q = 0.5$. We report the standard deviation in parenthesis. We highlight the smallest annualized average volatility amongst all estimators in bold.

Minimum variance portfolio

| | US | | | Japan | | | Europe | | |
|---|---|---|---|---|---|---|---|---|---|
| $N$ | 100 | 200 | 300 | 100 | 200 | 300 | 100 | 200 | 300 |
| RIE (IWs) | **12.1 (0.1)** | **11.0 (0.2)** | **10.4 (0.1)** | 28.7 (2.7) | 28.2 (2.7) | 27.8 (2.7) | 15.3 (0.2) | **13.5 (0.1)** | **13.4 (0.1)** |
| Clipping | 12.2 (0.2) | 11.0 (0.2) | 10.5 (0.1) | 28.7 (2.7) | 28.5 (2.7) | 28.1 (2.8) | **15.0 (0.2)** | 13.7 (0.1) | 13.8 (0.1) |
| Linear | 12.3 (0.2) | 11.3 (0.2) | 10.6 (0.1) | **28.6 (2.7)** | **28.0 (2.7)** | **27.7 (2.8)** | 15.4 (0.2) | 13.7 (0.1) | 13.5 (0.2) |
| Identity | 16.4 (0.3) | 15.7 (0.3) | 15.3 (0.3) | 31.3 (2.7) | 31.0 (2.7) | 31.0 (2.8) | 20.4 (0.3) | 20.1 (0.4) | 20.2 (0.4) |
| In sample | 14.6 (0.2) | 13.1 (0.2) | 12.3 (0.2) | 32.0 (2.8) | 31.3 (2.8) | 31.0 (2.8) | 18.2 (0.2) | 16.6 (0.2) | 18.2 (0.4) |

Mean reversion predictor

| | US | | | Japan | | | Europe | | |
|---|---|---|---|---|---|---|---|---|---|
| $N$ | 100 | 200 | 300 | 100 | 200 | 300 | 100 | 200 | 300 |
| RIE (IWs) | **21.9 (0.3)** | **11.8 (0.07)** | **10.0 (0.1)** | **24.5 (0.4)** | **13.8 (0.1)** | **12.5 (0.2)** | **26.4 (0.8)** | **15.4 (0.3)** | **10.0 (0.1)** |
| Clipping | 22.1 (0.3) | 11.9 (0.08) | 10.2 (0.1) | 25.2 (0.4) | 14.3 (0.1) | 13.2 (0.4) | 27.3 (0.9) | 15.9 (0.2) | 10.1 (0.1) |
| Linear | 22.6 (0.4) | 12.1 (0.08) | 10.3 (0.1) | 25.5 (0.5) | 14.2 (0.1) | 12.8 (0.3) | 27.3 (0.9) | 16.1 (0.3) | 10.3 (0.2) |
| Identity | 43.2 (2.5) | 27.3 (0.6) | 21.1 (0.3) | 64.0 (4.6) | 43.9 (3.9) | 41.3 (5.2) | 66.2 (2.5) | 42.2(1.7) | 31.2 (0.7) |
| In sample | 30.0 (0.6) | 15.7 (0.2) | 13.5 (0.2) | 31.7 (0.4) | 18.5 (0.3) | 15.8 (0.5) | 34.5 (1.2) | 20.0 (0.4) | 11.4 (0.1) |

Omniscient predictor

| | US | | | Japan | | | Europe | | |
|---|---|---|---|---|---|---|---|---|---|
| $N$ | 100 | 200 | 300 | 100 | 200 | 300 | 100 | 200 | 300 |
| RIE (IWs) | **13.6 (0.2)** | **11.1 (0.2)** | **11.7 (0.2)** | **12.1 (0.2)** | **11.2 (0.1)** | **12.2 (0.2)** | **10.2 (0.1)** | **9.9 (0.2)** | **9.82 (0.2)** |
| Clipping | 13.8 (0.2) | 11.2 (0.2) | 11.9 (0.2) | 12.3 (0.2) | 11.4 (0.1) | 12.7 (0.2) | 10.4 (0.1) | 11.3 (0.2) | 9.91 (0.2) |
| Linear | 13.9 (0.2) | 11.5 (0.2) | 12.0 (0.2) | 12.3 (0.2) | 11.4 (0.1) | 12.5 (0.2) | 10.6 (0.1) | 11.3 (0.2) | 9.87 (0.2) |
| Identity | 19.4 (0.5) | 16.4 (0.4) | 16.3 (0.3) | 20.7 (0.5) | 19.1 (0.3) | 22.6 (0.9) | 18.5 (0.3) | 18.4 (0.4) | 18.3 (0.5) |
| In sample | 16.7 (0.4) | 13.7 (0.3) | 14.6 (0.3) | 14.0 (0.3) | 14.7 (0.3) | 15.0 (0.3) | 11.0 (0.1) | 10.5 (0.2) | 11.4 (0.2) |

Uniform predictor

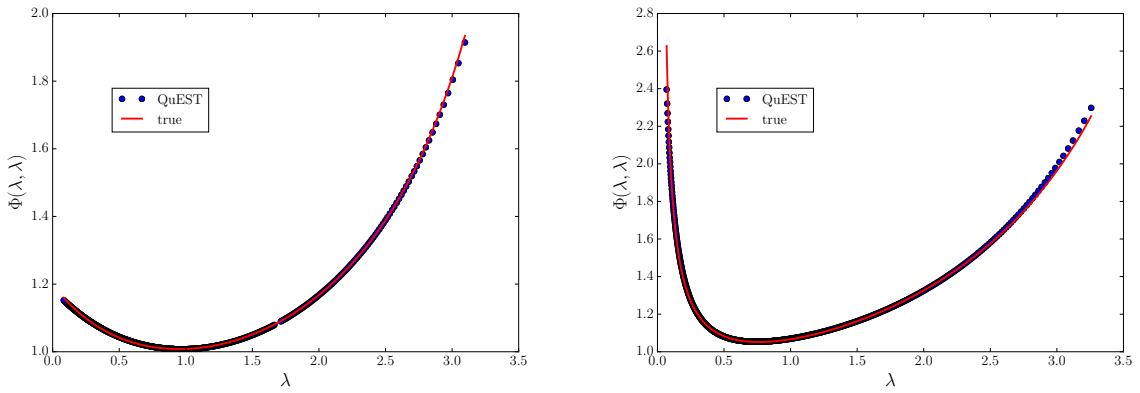
| | US | | | Japan | | | Europe | | |
|---|---|---|---|---|---|---|---|---|---|
| $N$ | 100 | 200 | 300 | 100 | 200 | 300 | 100 | 200 | 300 |
| RIE (IWs) | **2.72 (3e-3)** | **1.91 (2e-3)** | **1.57 (1e-3)** | **3.06 (4e-3)** | **2.16 (2e-3)** | **1.73 (1e-3)** | **2.85 (5e-3)** | **2.01 (4e-3)** | **1.58 (1e-3)** |
| Clipping | 2.77 (3e-3) | 1.94 (2e-3) | 1.59 (1e-3) | 3.19 (5e-3) | 2.2 (2e-3) | 1.80 (1e-3) | 2.96 (6e-3) | 2.16 (4e-3) | 1.63 (1e-3) |
| Linear | 2.74 (3e-3) | 1.93 (2e-3) | 1.61 (1e-3) | 3.07 (4e-3) | 2.18 (2e-3) | 1.75 (1e-3) | 2.90 (5e-3) | 2.03 (3e-3) | 1.6 (1e-3) |
| Identity | 3.25 (6e-3) | 2.36 (3e-3) | 1.85 (2e-3) | 4.82 (3e-2) | 3.23 (1e-2) | 3.13 (2e-2) | 3.71 (7e-3) | 3.01 (8e-3) | 2.3 (5e-3) |
| In sample | 3.71 (7e-3) | 2.56 (3e-3) | 2.12 (2e-3) | 4.11 (8e-3) | 3.0 (4e-3) | 2.38 (3e-2) | 3.69 (9e-3) | 3.13 (2e-2) | 2.33 (9e-3) |

overlaps as:

$$\langle \mathbf{u}_i \, , \tilde{\mathbf{u}}_i \rangle^2 = \frac{1}{Z_i} \sum_{j=1}^{N} \frac{\langle \mathbf{u}_i \, , \tilde{\mathbf{u}}_j \rangle^2}{(\lambda_i - \tilde{\lambda}_j)^2 + \eta^2}, \tag{9.4.1}$$

with $Z_i = \sum_{k=1}^{N} ((\lambda_i - \tilde{\lambda}_k)^2 + \eta^2)^{-1}$ the normalization constant and $\eta$ the width of the Cauchy kernel, that we choose to be $N^{-1/2}$ in such a way that $N^{-1} \ll \eta \ll 1$. We then average this quantity over all sample pairs for a given label $i$ to obtain $[\langle \mathbf{u}_i \, , \tilde{\mathbf{u}}_i \rangle^2]_e$, which should be a good approximation of Eq. (5.0.4) provided that we have enough data.

We consider two simple synthetic cases. Let us assume that $\mathbf{C}$ is an inverse Wishart with parameter $\kappa = 10$. We generate one sample of $\mathbf{E} \sim \text{Wishart}(N, T, C^{-1}/T)$ with $N = 500$, $T = 2N$ and we can compute the self-overlap (5.2.18) using the sample eigenvalues. We compare in Figure 9.4.1 the estimation that we get using QuEST algorithm (blue points) with the limiting "true" analytical solution (5.2.23) (red line) and we see that the fit is indeed excellent.




(A) Inverse Wishart ($\kappa = 10$). (B) $\mathbf{I}_N$+ GOE ($\sigma = 0.35$).

FIGURE 9.4.1. Evaluation of the self-overlap $\Phi(\lambda, \lambda)$ as a function of the sample eigenvalues $\lambda$ when $\mathbf{C}$ is an inverse Wishart of parameter $\kappa = 10$ (left) and $\mathbf{C}$ is a GOE centered around the identity with $\sigma = 0.35$ (right). In both cases, we compute the self-overlap (5.2.18) using analytical solution (red line) and the estimated from the sample eigenvalues using QuEST algorithm (blue points).

Next, we proceed to the same test using the power law distribution proxy (7.5.4) for $\rho_{\mathbf{C}}$ with $\lambda_0 = -0.6$ (see Eq. (4.2.41) for the precise definition of $\lambda_0$). Again, we emphasize that this model is quite complex since it naturally generates a finite number of outliers. The result is reported in Figure 9.4.2 where we plotted the self-overlap obtained by the limiting exact spectral density using Eq. (4.2.42) (red dashed line), the QuEST algorithm (blue plain line) and the empirical estimate (9.4.1) over 100 realizations of $\mathbf{E}$ (green points). Quite surprisingly, we see that the estimation obtained from the QuEST algorithm remains accurate for the outliers while the analytical solution becomes inaccurate for $\lambda \gtrsim 3.5$. This can be understood by the fact that the discrete approximation of the density (9.1.5) in QuEST yields a Dirac mass of weight of order $\mathcal{O}(N^{-1})$ (with $N$ finite numerically) while the limiting continuous density $\rho_{\mathbf{E}}(\lambda)$ becomes arbitrarily small for large eigenvalues.

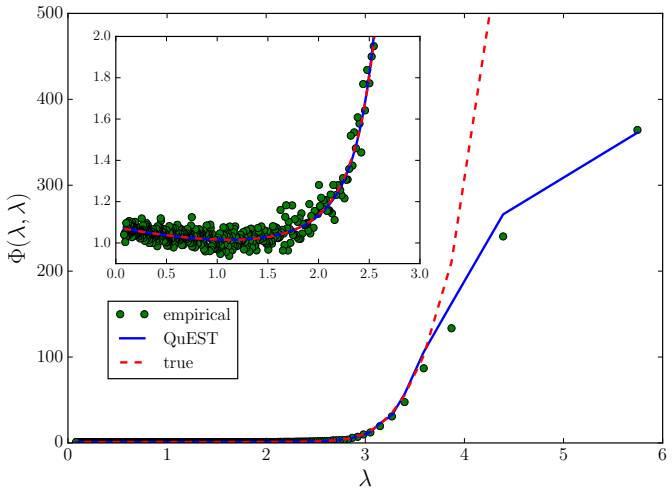



FIGURE 9.4.2. Main figure: Evaluation of the self-overlap $\Phi(\lambda,\lambda)$ as a function of the sample eigenvalues $\lambda$ when $\rho_{\mathbf{C}}$ is obtained from the power law proxy (7.5.4) with $\lambda_0 = 0.8$. We compare the analytical true solution using Eq. (4.2.42) (red dashed line) with the QuEST estimation (blue plain line) and also an empirical estimate over 100 realizations of $\mathbf{E}$ using Eq. (3.1.38) (green points). Inset: zoom in the bulk region of the main figure.

**9.4.2. Financial data.** We now investigate an application to real data, in the case of stock markets and using a bootstrap technique to generate different samples. Indeed, the difficulty here is to measure the empirical mean squared overlaps between the two sample correlation matrices $\mathbf{E}$ and $\mathbf{E}'$, as in Eq. (9.4.1), because we do not have enough data points to evaluate accurately an average over the noise as required in Eq. (5.0.4). To bypass this problem, we use a Bootstrap procedure to increase the size of the data.[3] Specifically, we take a total period of 2400 business days from 2004 to 2013 for the same three pools of assets that we split into two non-overlapping subsets of same size of 1200 days, corresponding to 2004 to 2008 and 2008 to 2013. Then, for each subset and of each Boostrap sample $b \in \{1,\dots,B\}$, we select randomly $T = 600$ distinct days for $N = 300$ stocks returns such that we construct two independent sample correlation matrices $\mathbf{E}_b$ and $\mathbf{E}'_b$, with $q = N/T = 0.5$. Note that we restrict to $N = 300$ stocks such that all of them are present throughout the whole period from 2004 to 2013. We then compute the empirical mean squared overlap (5.0.4) and also the theoretical limit (5.2.17) – using QuEST algorithm – from these $B$ bootstrap datasets.

For our simulations, we set $B = 100$ and plot in Figure 9.4.3 the resulting estimation of Eq. (5.0.4) we get from QuEST algorithm (blue dashed line) and the empirical bootstrap estimate (9.4.1) (green points) using US stocks. We also perform the estimation with an effective observation ratio $q_{\text{eff}}$ (red plain line) where use for each markets the values of $q_{\text{eff}}$ obtained above (see Figures 9.2.2-9.2.3b-9.2.4b). Note that the behaviour in bulk is quite well estimated by the asymptotic prediction Eq. (5.2.18) for both periods which is consistent with the estimation of Figure 9.2.2.

It is however clear from Figure 9.4.3 that the eigenvectors associated to large eigenvalues are

[3]This technique is especially useful in machine learning and we refer the reader to e.g. [79, Section 7.11] for a more detailed explanation.

not well described by the theory: we notice a discrepancy between the (estimated) theoretical curve and the empirical data even with an effective ratio $q_{\text{eff}}$. The difference is even worse for the market mode (data not shown). This is presumably related to the fact that the largest eigenvectors are expected to genuinely evolve with time, as argued in [2]. Note also the strong divergence at the left edge between the theoretical and empirical data in Figure 9.4.3, which can be partly corrected using the effective ratio $q_{\text{eff}}$, suggesting that we can still improve the estimation upon the Marčenko-Pastur framework by adding e.g. autocorrelation or heavy tailed entries which allows to widen the LSD of $\mathbf{E}$ (see e.g. [18, 48] for autocorrelation and [26, 47, 74] for heavy tailed entries) before invoking the need of some structural evolution of $\mathbf{C}$ with time.

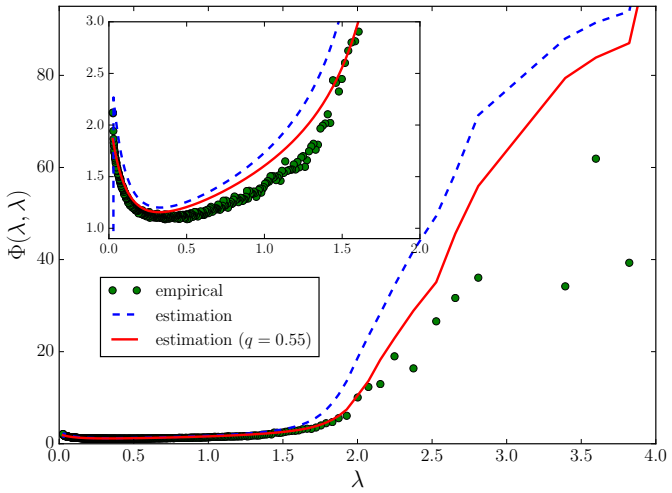
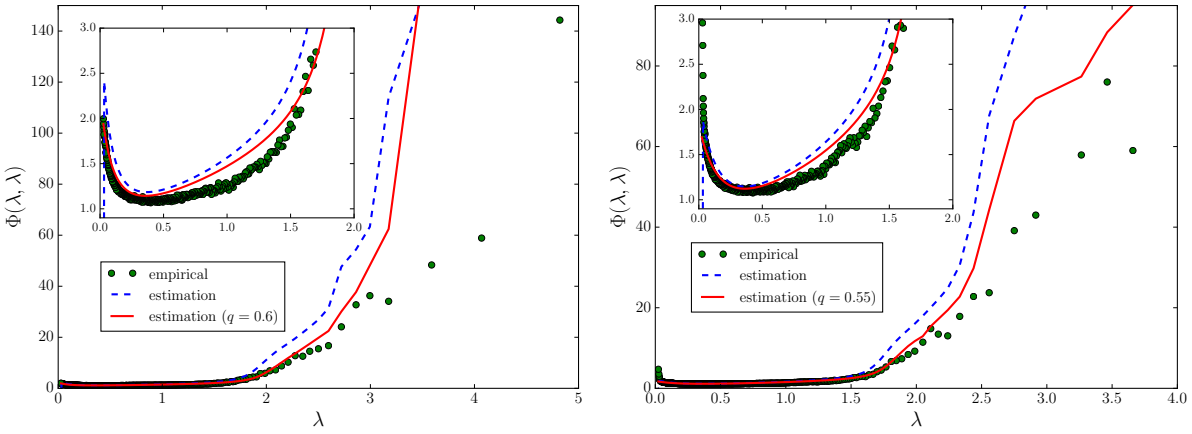



FIGURE 9.4.3. Evaluation of the self-overlap $\Phi(\lambda, \lambda)$ as a function of the sample eigenvalues $\lambda$ using the $N = 300$ most liquid US equities from 2004 to 2013. We split the data into two non-overlapping period with same sample size 1200 business days. For each period, we randomly select $T = 600$ days and we repeat $B = 100$ bootstraps of the original data. The empirical self-overlap is computed using Eq. (9.4.1) over these 100 bootstraps (green points) and the limiting formula (5.2.18) is estimated using QuEST algorithm with $q = 0.5$ (blue dashed line). We also provide the estimation we get using the same effective observation ratio $q_{\text{eff}}$ than in Figure 9.2.2. Inset: focus in the bulk of eigenvalues.

All the above results can be extended and qualitatively and quantitatively confirmed in the case of Japanese and European stocks, for which the results are plotted respectively in Figures 9.4.4a.

To conclude, these observations suggest further improvements upon the time independent framework of Marčenko and Pastur, that one allow one to account for some "true" dynamics of the underlying correlation matrix. That such dynamics exist for eigenvectors corresponding to the largest eigenvalues is intuitively reasonable, and empirically confirmed by the analysis of Ref. [2]. The full correlation matrix might also evolve and jump between different "market states", as suggested in various recent papers of the Guhr group (see e.g. [156, 185] and references therein). Extending the present framework to these cases is quite interesting and would shed light on the optimal value of the observation ratio $q_{\text{eff}}$ which was systematically found to be




(A) TOPIX (Japan)

(B) Bloomberg 500 (Europe)

FIGURE 9.4.4. Evaluation of the self-overlap $\Phi(\lambda, \lambda)$ as a function of the sample eigenvalues $\lambda$ using the $N = 300$ most liquid equities from the Japanese TOPIX (left) and the European Bloomberg 500 index (right) from 2004 to 2013. For each case, we split the data into two non-overlapping period with same sample size $T = 1200$ business days. For each period, we randomly select 600 realizations of the returns and we repeat $B = 100$ bootstraps of the original data. The empirical self-overlap is computed using Eq. (9.4.1) over these 100 bootstraps (green points) and the limiting formula (5.2.18) is estimated using QuEST algorithm with $q = 0.5$ (blue dashed line). We also provide the estimation we get using the same effective observation ratio $q$ than in Figure 9.2.2. Inset: focus in the bulk of eigenvalues.

larger than $q = N/T$. This could be an indication of non-stationnay effects. This is particularly apparent for the Japanese stocks (see e.g. Fig. 9.4.4a) where the theoretical prediction deviates significantly from the empirical one even if we calibrate the effective quality ratio $q_{\text{eff}}$. The case of eigenvectors associated to the small eigenvalues is particularly striking and probably need further scrutiny, in particular in the case of futures markets where the presence of very strongly correlated contracts (i.e. two different maturities for the same underlying) leads to very small "true" eigenvalues of the correlation matrix, for which the above IW-regularizing scheme is probably inadequate. We leave these issues, as well as several others alluded to in the following concluding chapter, for further investigations.

# Chapter 10

# Outroduction

We have discussed at length some of the most advanced techniques in RMT and their application to estimate large correlation matrices, in particular within a rotational invariant framework. Moreover, we showed through an extended empirical analysis that these estimators can be of great interest in real world situations. Instead of repeating the main messages emphasized in the previous sections, we want to end this whole part on covariance matrices with an (incomplete) list of potentially interesting open problems, that represent natural extensions of the results obtained above.

## 10.1 Extension to more general models of covariance matrices

One important assumption of the sample covariance matrix model (4.1.3) is the absence of temporal correlations and/or temporal structure in the data. As it is well known, this assumption is clearly not true in most real life applications (see e.g. Section 9.4). It is thus natural to extend this work to estimators that accounts for some temporal dependence. The simplest case is when some *autocorrelations* are present. A standard assumption is that of an exponential autocorrelation of the form [18, 47, 48]:

$$\mathbb{E}[Y_{it}Y_{jt'}] = C_{ij}\exp\big[-|t-t'|/\tau\big], \tag{10.1.1}$$

where $\tau$ controls the range of the time correlations.

Another frequent situation is when covariances are measured through an *Exponential Weighted Moving Average* (EWMA) [47, 142]:[1]

$$M_{ij}(\tau, T) = (1-\alpha)\sum_{t=0}^{T}\alpha^t Y_{i,\tau-t}Y_{j,\tau-t}, \tag{10.1.2}$$

where $\tau$ is the last estimation date available, $\alpha \in (0,1)$ is a constant and $T$ is the total size of the time series. Roughly, the idea of this estimator is that old data become gradually obsolete so that they should contribute less than more recent information. We see that the estimator (10.1.2) can be rewritten as

$$M_{ij}(\tau) \;=\; (1-\alpha)\sum_{t=0}^{T} H_{it}H_{jt}, \quad \text{with} \quad \mathbb{E}\big[H_{it}H_{it'}\big] = \delta_{tt'}(1-\alpha)\alpha^t, \tag{10.1.3}$$

[1] We denote in the following the different estimators of $\mathbf{C}$ by $\mathbf{M}$ to avoid confusion with Pearson's sample estimator $\mathbf{E} = \mathbf{X}\mathbf{X}^*/T$.

i.e. the variance of the random variables have an explicit time dependence.

Another interesting way to generalize the Marčenko-Pastur framework concerns the distribution of the entries. An important assumption for the Marčenko-Pastur equation to be valid is that each entry $Y_{it}$ possesses a finite fourth moment. Again, this assumption may not be satisfied in real dataset, especially in Finance [51]. As alluded to in Section 4.1.3, a more robust estimate of the covariance matrix in then needed [127]. Let us assume that we can rewrite the observations as $Y_{it} = \sigma_t \mathbf{C}^{1/2} X_{it}$ for any $i \in [\![1, N]\!]$ and $t \in [\![1, T]\!]$, where $\sigma_t$ is a fluctuating global volatility that sets the overall scale of the returns, and $\mathbf{X}$ are iid Gaussian variables. In that particular context, the sample covariance matrix is obtained as the solution of the fixed-point equation [127]:

$$\mathbf{M} := \frac{1}{T} \sum_{t=1}^{T} U\Big(\frac{1}{N} \mathbf{y}_t^* \mathbf{M}^{-1} \mathbf{y}_t\Big) \mathbf{y}_t \mathbf{y}_t^*,$$

where $U$ is a non-increasing function. As mentioned in Section 4.1.3, it is possible to show that for the $U(x) = x^{-1}$, one has $\mathbf{M} \to \mathbf{E}$ in the large $N$ limit [26, 58, 74, 197], where $\mathbf{E} = \mathbf{C}^{1/2} \boldsymbol{\mathcal{W}} \mathbf{C}^{1/2}$ and $\boldsymbol{\mathcal{W}}$ is a Wishart matrix. However, the asymptotic limit is more complex for general $U$'s and reads:

$$\mathbf{M} \to \mathbf{C}^{1/2} \mathbf{X} \mathbf{B} \mathbf{X}^* \mathbf{C}^{1/2}, \tag{10.1.4}$$

where $\mathbf{B}$ is a *deterministic* diagonal $T \times T$ matrix where each entry is a functional of the $\{\sigma_t\}_t$ and the function $U$ (see e.g. [58] for the exact expression of the matrix $\mathbf{B}$).

Interestingly, all the above models, (10.1.1), (10.1.3) and (10.1.4), can be wrapped into a general multiplicative framework that reads:

$$\mathbf{M} \;:=\; \mathbf{C}^{1/2} \mathbf{X} \mathbf{B} \mathbf{X}^* \mathbf{C}^{1/2}, \tag{10.1.5}$$

where $\mathbf{X} := (X_{it}) \in \mathbb{R}^{N \times T}$ is a random matrix with zero mean and variance $T^{-1}$ iid entries and $\mathbf{B} = (B_{tt'}) \in \mathbb{R}^{T \times T}$ is fixed matrix, independent from $\mathbf{C}$. Indeed, for (10.1.1), we have $B_{tt'} = \exp[-|t - t'|/\tau]$ while we set $B_{tt'} = \delta_{tt'}(1 - \alpha)\alpha^t$ for (10.1.3).

The optimal RIE for this model has been briefly mentioned in Section 7.6 give precise equation number and in more exquisite details in [40]. We saw that the oracle estimator associated to the model (10.1.5) converges – at least for bulk eigenvalues – to a limiting function that does not depend explicitly on the spectral density of $\mathbf{C}$ (see Eq. (7.6.3)). It is thus interesting to see whether one of the aforementioned models can be solved in full generality (see [48] for model (10.1.1)) and whether one can explain the appearance of an effective ratio $q_{\text{eff}} > q$, as encountered in Chapter 9. Furthermore, another important result would be to see whether the estimator (7.6.3) is also valid for outliers, as is the case for the time-independent sample covariance matrices.

## 10.2 Singular Value Decomposition

A natural extension of the work presented in this manuscript is to consider rectangular correlation matrices. This is particularly useful when one wishes to measure the correlation between $N$ *inputs* variables $\mathbf{x} := (x_1, \ldots, x_N)$ and $M$ *outputs* variables $\mathbf{y} := (y_1, \ldots, y_M)$. The vector $\mathbf{x}$ and the $\mathbf{y}$ may be completely different from one another (for example, $\mathbf{x}$ could be production indicators and $\mathbf{y}$ inflation indexes) or it also could be the same set of observables but observed

at different times (*lagged* correlation matrix [29]). The cross-correlations is thus characterized by a rectangular $N \times M$ matrix $\boldsymbol{\mathcal{C}}$ defined as:

$$\mathcal{C}_{ia} \;:=\; \mathbb{E}[x_i y_a], \tag{10.2.1}$$

where we assumed that both quantities have zero mean and variance unity.

What can be said about the structure of this rectangular and non symmetric correlation matrix (10.2.1)? The answer is obtained from the singular value decomposition (SVD) in the following sense: what is the (normalized) linear combination of $\mathbf{x}$'s on the one hand, and of $\mathbf{y}$'s on the other hand, that have the strongest mutual correlation? In other words, what is the best pair of predictor and predicted variables, given the data? The largest singular value – say $c_1 \in (0,1)$ and its corresponding left and right eigenvectors answer precisely this question: the eigenvectors tell us how to construct these optimal linear combinations, and the associated singular value gives us the strength of the cross-correlation. We may then repeat this operation on the $N-1$ and $M-1$ dimensional sub-spaces orthogonal to the two eigenvectors for both input and output variables. This yields a list of singular values $\{c_i\}_i$ that represent the prediction power of the corresponding linear combinations (in decreasing order). This is called *Canonical Correlation Analysis* (CCA) in the literature and has (see [91] or [101, 193] for more recent works).

In order to study the singular values and the associated left and right eigenvectors, we consider the $N \times N$ matrix $\boldsymbol{\mathcal{C}}\boldsymbol{\mathcal{C}}^*$, which is now symmetric and has $N$ non negative eigenvalues. Indeed, the trick behind this change of variable is that the eigenvalues of $\boldsymbol{\mathcal{C}}\boldsymbol{\mathcal{C}}^*$ are equal to the square of a singular value of $\boldsymbol{\mathcal{C}}$ itself. Then, the eigenvectors give us the weights of the linear combination of the $\mathbf{x}$'s that construct the *best* predictors in the above sense. In order to obtain the right eigenvectors of $\boldsymbol{\mathcal{C}}$, one forms the $M \times M$ matrix $\boldsymbol{\mathcal{C}}^*\boldsymbol{\mathcal{C}}$ that has exactly the same non zero eigenvalues as $\boldsymbol{\mathcal{C}}\boldsymbol{\mathcal{C}}^*$; the corresponding eigenvectors now give us the weights of the linear combination of the $\mathbf{y}$'s that construct the *best* predictees. If $M > N$, the matrix $\boldsymbol{\mathcal{C}}^*\boldsymbol{\mathcal{C}}$ has $M-N$ additional zero eigenvalues; whereas in the other case, it is $\boldsymbol{\mathcal{C}}\boldsymbol{\mathcal{C}}^*$ that has an excess of $N-M$ zero eigenvalues.

However, as for standard correlation matrices, the knowledge of the true population matrix Eq. (10.2.1) is unavailable. Hence, one resorts to an empirical determination of $\boldsymbol{\mathcal{C}}$ that is strewn with measurement noise, as above. We expect to be able to use tools from RMT to understand the how the true singular values are dressed by the measurement noise. To that end, suppose that we have a total of $T$ observations of both quantities that we denote by $[X_{it}]_t$ and $[Y_{at}]_t$. Then, the empirical estimate of $\boldsymbol{\mathcal{C}}$ is given by

$$\mathcal{E}_{ia} \;:=\; \frac{1}{T}\sum_{t=1}^{T} X_{it} Y_{at}\,, \tag{10.2.2}$$

and the aim is to study the singular values of this matrix. Indeed, as in Chapter 4, we expect the measurement noise to affect the accuracy of the estimation in the limit $N, M, T \to \infty$ with $n = N/T$ and $m = M/T$ finite, which we will assume to be both smaller than unity in the following. As explained in the previous paragraph, a convenient way to perform this analysis is to consider the eigenvalues of $\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^*$ (or $\boldsymbol{\mathcal{E}}^*\boldsymbol{\mathcal{E}}$). Using tools from Appendix B, especially Eq. (B.2.3), we see that

$$\det(\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^* - z\mathbf{I}_N) = \det\Big(\mathbf{S}_\mathbf{X}\mathbf{S}_\mathbf{Y} - z\mathbf{I}_T\Big)\,, \qquad \mathbf{S}_\mathbf{X} \;:=\; \frac{\mathbf{X}^*\mathbf{X}}{T}, \quad \mathbf{S}_\mathbf{Y} \;:=\; \frac{\mathbf{Y}^*\mathbf{Y}}{T}$$

so that $\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^*$ shares the same non-zero eigenvalues than the product of the dual $T \times T$ samples covariance matrix $\mathbf{S_X}$ and $\mathbf{S_Y}$.

It is easy to see that when $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated, i.e. $\boldsymbol{\mathcal{C}} = \mathbf{0}$, one can compute the spectral density of $\mathbf{S_X S_Y}$ using the free multiplication formula (3.1.81). However, the result depends in general on the correlation structure of the input variables, $\mathbf{C}_X$, and of the output variables $\mathbf{C}_Y$. A way to obtain a universal result is to consider the exact normalized PCA's of the $\mathbf{X}$ and of the $\mathbf{Y}$, that we call $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$, such that $\mathbf{S}_{\hat{\mathbf{X}}}$ has $N$ eigenvalues equal to 1 and $T-N$ eigenvalues equal to zero, while $\mathbf{S}_{\hat{\mathbf{Y}}}$ has $M$ eigenvalues equal to 1 and $T-M$ eigenvalues equal to zero. In this case, the limiting spectrum of singular values can be found explicitly (see [28] and [184] for an early derivation without using free probability methods), and is given by:

$$\rho(c) = \max(m+n-1, 0)\delta(c-1) + \mathrm{Re}\frac{\sqrt{(c^2-\gamma_-)(\gamma_+-c^2)}}{\pi c(1-c^2)}, \tag{10.2.3}$$

where $\gamma_\pm$ are given by:

$$\gamma_\pm = n + m - 2mn \pm 2\sqrt{mn(1-n)(1-m)}, \quad 0 \leqslant \gamma_\pm \leqslant 1 \tag{10.2.4}$$

The allowed $c$'s are all between 0 and 1, as they should since these singular values can be interpreted as correlation coefficients. In the limit $T \to \infty$ at fixed $N$, $M$, all singular values collapse to zero, as they should since there is no true correlations between $X$ and $Y$. the allowed band in the limit $n, m \to 0$ becomes:

$$c \in \left[|\sqrt{m} - \sqrt{n}|, \sqrt{m} + \sqrt{n}\right],$$

showing that for fixed $N, M$, the order of magnitude of allowed singular values decays as $T^{-1/2}$. The above result allows one devise precise statistical tests, see [28, 101, 193].

The general case where when $\mathbf{X}$ and $\mathbf{Y}$ are correlated, i.e. $\boldsymbol{\mathcal{C}} \neq \mathbf{0}$, is, to our knowledge, unknown. This is particularly relevant for practical cases since one might expect some true correlations between the input and output variables. It would be interesting to characterize how the noise distorts the "true" cross-correlations between $\mathbf{X}$ and $\mathbf{Y}$, as the analogue of the Marčenko-Pastur equation (4.2.1). Moreover, an analysis of the left and right eigenvectors like in Chapter 5 would certainly be of interest in many real life problems (see e.g. [6, 79, 80, 108] for standard applications). Note that the case of outlier singular values and vectors of rectangular random matrices subject to a low rank perturbation has been considered [24].

## 10.3 Estimating the eigenvectors

As indicated by its name, the optimal RIE is optimal under the assumption that we have no prior insights on the true components, i.e. the eigenvectors of the population covariance matrix $\mathbf{C}$. However, in some problems we expect these eigenvectors to have some specific, non isotropic structure. One possible solution to this problem is to formulate prior structures for these eigenvectors through factor models [52, 77], ultrametric tree models (*eigenvector clustering*) [67, 177], or constraints on the participation ratios [136].

Very recently, an attempt to "clean" empirical outlier eigenvectors was formulated in [136]. Let us focus for example on the top eigenvector; the prior is then defined as a weighted sum of the sample eigenvectors:

$$\widehat{\mathbf{v}}_1 = \sqrt{\Phi(\mu_1, \lambda_1)}\,\mathbf{u}_1 + \sum_{j=2}^{N} \varepsilon_j \sqrt{\Phi(\mu_1, \lambda_j)}\mathbf{u}_j\,, \tag{10.3.1}$$

where the bivariate mean squared overlap $\Phi$ is defined in Eq. (5.0.3) and the $\{\varepsilon_j\}_{j\geqslant 2}$ is a set of i.i.d. Gaussian random variables with zero mean and unit variance, that must be determined in such a way that $\widehat{\mathbf{v}}_1$ is, for example, as "localized" as possible. One notices that the first term in the RHS of Eq. (10.3.1) can be computed using Eq. (5.1.10) and the second one can be inferred from Eq. (5.1.12). On average, we see that $\langle \widehat{\mathbf{v}}_1 \rangle_\varepsilon \cdot \mathbf{u}_1 = \sqrt{\Phi(\mu_1, \lambda_1)}$, as it should. While this prior requires some knowledge about the number of outliers – which is still an open question – it is shown in [136] that this method improves the accuracy of the estimation on synthetic data. It would be interesting to make use of some of these ideas in the context financial data.

## 10.4 Cleaning recipe for $q > 1$

As observed in Chapter 9, the optimal RIE (9.1.4) returns very satisfactory result in terms of estimating the oracle estimator either with synthetic data or real data when the sample size is greater than the number of variables. However, it may happen in practice that one is confronted to the case where $N > T$ in which the sample covariance matrix $\mathbf{E}$ has generically $N - T$ zero eigenvalues. The main difficulty is to interpret these null eigenvalues since they could either be due to the fact we do not have enough data points, or else that $\mathbf{C}$ has some exact zero modes. It is therefore not surprising that both regularizations schemes of Chapter 9 fail to estimate correctly the small eigenvalues in this case (see Figure 10.4.1). However, they fail in different ways: the IWs-regularization leaves zero eigenvalues unaltered while the QuEST algorithm shrinks the small eigenvalues upwards too much.

A naive and ad-hoc approach to this problem when $\mathbf{C}$ has *no* zero mode is to rescale the $N-T$ zeros eigenvalues of the IWs-regularization by a constant so that the trace of the estimator is equal to $N$, as it should be. This is similar to the clipping procedure of Section 8.2. We see that the main problem with this simple recipe is that when $\mathbf{C}$ has some exact zero modes, then we will always overestimate the volatility of these zero risk modes. Hence, at this stage, it seems that there are no satisfactory systematic cleaning recipe when $q > 1$, in the absence of some information about the possibility of true zero modes.

## 10.5 A Brownian Motion model for correlated Wishart matrices

We present in Appendix 11 that Dyson's Brownian Motion that offers a nice physical interpretation of dynamics of the sample eigenvalues and eigenvectors in the case of an *additive* noise. It also provides a straightforward tool to compute the dynamics of the resolvent of the sample matrix; Eq. (12.2.16) is quite remarkable in that eigenvectors' overlaps may be easily inferred.

We are not aware of a similar result in the multiplicative case, with sample covariance matrices in mind, although Eq. (4.2.4) suggest that such a process should exist. In the case where $\mathbf{C} = \mathbf{I}_N$, Bru's Wishart process [37] allows one to obtain many interesting properties about both the eigenvalues and eigenvectors – see [4, 32], but time in this case is not related to the quality parameter $q$, as one would like it to be. This question is quite fundamental and also has practical applications, as it would for example allow to understand the overlap of the eigenvectors of $\mathbf{E}$ at different "times" (see e.g. [2,3] for a related question in the additive model). An attempt to construct such process can be found in [2, 5] but the standard Ito calculus, used in Appendix 12.2.2, cannot be used because the noise has a strongly non-Gaussian structure in this case. Progress seems however within reach.

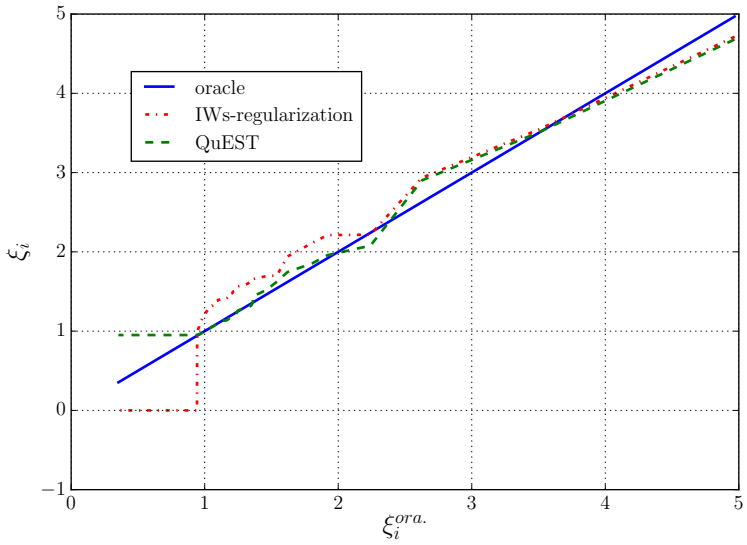



FIGURE 10.4.1. We apply the IWs (red dash-dotted line) and QuEST (green dashed line) regularization of Chapter 9 as a function of the oracle estimator (7.1.2) with $\rho_{\mathbf{C}}$ given by Eq. (7.5.4) with $\lambda_0 = 0.8$ and $N = 1000$. The sample covariance matrix $\mathbf{E}$ is a Wishart matrix with $q = 2$. We see that both regularizations provide results that are far from the optimal solution (blue plain line).

# Part II

# Contributions to additive random matrix models

# Chapter 11

# Introduction

## 11.1 Setting the stage

The previous part, which was the bulk of the thesis, dealt with the estimation of large covariance matrices. Precisely, we were mainly focused on the sample covariance matrix model and discussed about some possible generalizations. Nonetheless, as mentioned in Section 7.6, the theory of rotational invariant estimators holds in a much broader context. Even though the practical applications within the field of high dimensional statistics are not as concrete as the sample covariance matrix model, we believe that the analytical techniques of this part on additive models can be of particular interest for further works either in physics or statistics.

The model we consider in this part is the case where a fixed "pure" matrix $\mathbf{C}$ is corrupted by an additive independent noise. This model finds many applications in several situations in physics, in particular in quantum chaos and quantum transport [19], with renewed interest coming from problems of quantum ergodicity ("eigenstate thermalisation") [64,94], entanglement and dissipation (for recent reviews see [73,139]). We consider that the $N \times N$ matrix $\mathbf{C}$ has real eigenvalues. The model we shall study is of the form:

$$\mathbf{M} = \mathbf{C} + \mathbf{\Omega B \Omega}^*, \tag{11.1.1}$$

where $\mathbf{B}$ is a fixed matrix with a well defined spectral density $\rho_{\mathbf{B}}$ and $\mathbf{\Omega}$ is a random matrix chosen in the Orthogonal group $\mathbf{O}(N)$ according to the Haar measure. Clearly, the noise term is invariant under rotation so that we expect the resolvent of $\mathbf{M}$ to be (for large $N$) in the same basis as $\mathbf{C}$. We shall therefore posit without loss of generality that $\mathbf{C}$ is diagonal.

The most natural application of this general model (11.1.1) is obviously when the external noise $\mathbf{B}$ belongs to the Gaussian Orthogonal Ensemble, introduced in Chapter 3. We recall that the $N \times N$ matrix $\mathbf{B}$ is a GOE if it is a real symmetric matrix with Gaussian entries that satisfies

$$\mathbb{E}[B_{ij}] = 0 \qquad \mathbb{E}[B_{ij}^2] = \begin{cases} 2\sigma^2/N & \text{if } i = j\,, \\ \sigma^2/N & \text{otherwise.} \end{cases} \tag{11.1.2}$$

In the mathematical literature, we say that $\mathbf{M}$ is a *deformed GOE* matrix. In physics, this model is rather known as "deterministic plus noise" model [35,196]. In order to lighten the notations, we shall rather use the "deformed GOE " terminology to refer to this model throughout the following.

## 11.2 Outline and main contributions

In the following chapter, we focus on the deformed GOE. In particular, as for sample covariance matrices, we will show that closed formulas for the mean squared overlaps are available. For the bulk component, this observation is not new [3, 150] but the method we propose here allows to have a new dynamical interpretation of these mean squared overlaps. The main tool will be Dyson's Brownian Motion that provides a diffusion process interpretation of the deformed GOE [70]. The main advantage of using this dynamical framework is that we are also able to deal with outlier eigenvectors. This work lead to the article [5] with Romain Allez and Jean-Philippe Bouchaud. In addition, we also consider the mean squared overlaps between correlated deformed GOE – say $\mathbf{M}$ and $\tilde{\mathbf{M}}$ – as done in Section 5.2. Using the additivity of the model, we may show that contrary to sample covariance matrices, the case where the two realizations $\mathbf{M}$ and $\tilde{\mathbf{M}}$ have correlated noise can be solved in full generality. The overlaps between correlated deformed GOE have been investigated in the work [41] written with Jean-Philippe Bouchaud and Marc Potters.

The last chapter of this part is dedicated to the general model (11.1.1) without assuming a Gaussian structure. We will see that we are still able to study in details the mean squared overlaps (5.0.3) which are related to the free addition formula in the large $N$ limit. Hence, the first section of this chapter is actually dedicated to an elementary derivation of Voiculescu's free addition formula (3.1.66). We then use this result to derive the limiting value of the resolvent of the model (11.1.1). Then, thanks to the inversion formula (5.1.5), we may compute the mean squared overlaps and the optimal RIE for this model. Chapter 13 is based on the article [40] written in collaboration with Romain Allez, Jean-Philippe Bouchaud and Marc Potters. Note that contrary to the deformed GOE, the analysis of the general model (11.1.1) with the presence of a finite number of outliers is still an open question.

# Chapter 12

# Eigenvectors statistics of the deformed GOE

*This chapter is based on [5] and [41].*

In this section, we consider the deformed GOE presented in the previous chapter as a diffusion process $(\mathbf{M}(t))_{t\geq 0}$ in the space of $N \times N$ real symmetric or Hermitian matrices starting from a given deterministic matrix $\mathbf{C}$ and evolving with time according to a Hermitian Brownian motion. The matrix $\mathbf{M}(t)$ at time $t$ is given by

$$\mathbf{M}(t) \ := \ \mathbf{C} + \mathbf{B}(t) \tag{12.0.1}$$

where $(\mathbf{B}(t))_{t\geq 0}$ is a Hermitian Brownian motion, i.e. a diffusive matrix process such that $\mathbf{B}_0 = 0$ and whose entries $\{B_{ij}(t), i \leq j\}$ are given by

$$B_{ij}(t) := \frac{1}{\sqrt{N}} W_{ij}(t) \quad \text{if} \quad i \neq j, \quad B_{ii}(t) := \frac{\sqrt{2}}{\sqrt{N}} W_{ii}(t) \tag{12.0.2}$$

where the $W_{ij}(t), i \leqslant j$ are independent and identically distributed real or complex (real if $i = j$) Brownian motions.

The matrix $\mathbf{C}$ is the fixed external source and can be seen as a signal that one would like to estimate from the observation of the noisy matrix $\mathbf{M}(t)$.

The aim of this section is to investigate the effect of the addition of the noisy perturbation matrix $\mathbf{B}(t)$ in the limit of large dimension $N \to +\infty$. More precisely, we investigate the relationship between the eigenvectors of the perturbed matrix $\mathbf{M}(t)$ with those of the initial matrix $\mathbf{C}$ for some given $t > 0$. We emphasize that although we will focus here in the case where $t$ is independent from $N$, it is also possible to consider that $t$ scales with the dimension $N$ of the matrices (see [5] for details).

The evolution as $t$ grows of the eigenvalues $\lambda_1(t) \geq \lambda_2(t) \geq \cdots \geq \lambda_N(t)$ of the symmetric matrix $\mathbf{M}(t)$ has been investigated in tremendous details in random matrix theory (see [8, section 4.3] for a review). It was first shown by Dyson [70] in 1962 that the eigenvalues of the matrix $\mathbf{M}(t)$ evolve according to the Dyson Brownian motion which describes the evolution of $N$ positively charged particles (Coulomb gas) subject to electrostatic repulsion and to independent thermal noises. The dynamics of the Dyson Brownian motion were studied in many details for different purposes. The most striking applications of the Dyson Brownian motion are perhaps

the proofs of the universality conjectures for Wigner matrices (see e.g. [31,75,76] and references therein). The Dyson Brownian motion was also used in theoretical physics as a model to study disordered metals and chaotic billiards [19] (see also [78]). In this context, the authors compute the correlations between the positions of the eigenvalues in the bulk at a given time $s$ with those at a later time $t > s$. The asymptotic correlation functions are described in terms of the extended Hermite kernel. The correlations between the positions of the eigenvalues near the edge of the spectrum at different times were later computed in [119] in terms of the extended Airy kernel.

The study of the associated eigenvectors denoted respectively[1] by $|\mathbf{u}_1^t\rangle, |\mathbf{u}_2^t\rangle, \dots, |\mathbf{u}_N^t\rangle$ is comparatively much poorer. A few authors were interested in some aspects of eigenvector fluctuations (see e.g. in [21, 32, 68, 173] on the statistics of Haar matrices, [23, 113, 144] for eigenvectors of covariance matrices and [2] for applications in finance) but yet very little is known about the cross correlation of the eigenvectors at different times $s$ and $t > s$. It is a natural question to extend the results known for the eigenvalues [19, 119] by investigating the relation between the eigenvectors of the matrix $\mathbf{M}(s)$ with those at a later time $t > s$ (with possibly $s = 0$). This question was initiated in [190] and recently reconsidered in [3] where the authors investigated the projections of a given eigenvector $|\mathbf{u}_i^0\rangle$ at time 0 on the orthonormal basis of the perturbed eigenvectors at time $t$. Specifically, we consider the case where the associated eigenvalue $\lambda_i(0)$ lies in the continuous part of the spectrum and use Stieltjes transform methods to compute the asymptotic (mean squared) projections of this vector on the orthonormal basis at time $s = 0$.

As highlighted in Section 5 above, the information about the mean squared overlaps (5.0.3) can be studied through the resolvent. We will present two different ways to evaluate the limiting value of the resolvent of $\mathbf{M}(t)$. The first one is to use the dynamics of the eigenvalues and eigenvectors as done in [5] while the second approach consists in applying Itô's lemma directly on the entries on the resolvent of $\mathbf{M}(t)$. The main advantage of the first approach is that it provides a nice physical interpretation of possible outliers in the spectrum of $\mathbf{M}(t)$, and we shall study these outlier eigenvectors at the end of this chapter.

## 12.1 Eigenvalues and eigenvectors trajectories

**12.1.1. Eigenvalues and eigenvectors diffusion processes.** It is well known [8] that the eigenvalues $\lambda_1(t) \geq \lambda_2(t) \geq \cdots \geq \lambda_N(t)$ of the matrix $\mathbf{M}(t)$ evolve according to the Dyson Brownian motion

$$\mathrm{d}\lambda_i(t) = \sqrt{\frac{2}{\beta N}}\,\mathrm{d}b_i(t) + \frac{1}{N}\sum_{k\neq i}\frac{\mathrm{d}t}{\lambda_i(t)-\lambda_k(t)}\,, \quad i = 1,\dots,N\,, \tag{12.1.1}$$

where the $b_i(t)$ are independent real Brownian motions, and satisfy the initial conditions

$$\lambda_i(0) = \mu_i\,, \quad i = 1,\dots,N\,.$$

The eigenvalues of $\mathbf{M}(t)$ may be seen as positively charged particles in a one-dimensional Coulomb gas with electrostatic repulsion between them and subject to a thermal noise $db_i(t)$.

Conditionally on the eigenvalues paths, the trajectories of the associated eigenvectors $|\mathbf{u}_1^t\rangle$, $|\mathbf{u}_2^t\rangle, \dots, |\mathbf{u}_N^t\rangle$ can be realized continuously as a function of $t$. This eigenvector flow was first

[1] For the sake of clearness, especially when applying Itô's formula, we shall use Dirac bra-ket notations in this section.

exhibited in [37] for Wishart processes. Those continuous paths are determined using standard perturbation theory or stochastic analysis tools (see again [8]): in our case, we have, for all $i = 1, \dots, N$,

$$\mathrm{d}|\mathbf{u}_i^t\rangle = -\frac{1}{2N}\sum_{k\neq i}\frac{\mathrm{d}t}{(\lambda_i(t)-\lambda_k(t))^2}|\mathbf{u}_i^t\rangle + \frac{1}{\sqrt{N}}\sum_{k\neq i}\frac{\mathrm{d}w_{ik}(t)}{\lambda_i(t)-\lambda_k(t)}|\mathbf{u}_k^t\rangle\,, \tag{12.1.2}$$

$$\text{with} \quad |\mathbf{u}_i^0\rangle = |\mathbf{v}_i\rangle\,, \tag{12.1.3}$$

where the family of independent (up to symmetry) real Brownian motions $\{w_{ij} : i \neq j\}$ is independent of the eigenvalues trajectories (i.e. independent of the driving Brownian motions $b_i$ in (12.1.1)). We can therefore freeze the eigenvalues trajectories and then, conditionally on this eigenvalues path, study the eigenvectors evolution. The eigenvector process can thus be regarded as a diffusion process in a random environment which depends on the realized trajectories of the eigenvalues. This is an important fact that will be used several times throughout this chapter. Most of the results derived in this chapter concern the large dimensional statistics of the eigenvectors and hold almost surely with respect to the eigenvalues trajectories.

The evolution equation (12.1.2) for the $i$-th eigenvector contains two orthogonal terms. The first term, collinear to $|\mathbf{u}_i^t\rangle$, pulls back $|\mathbf{u}_i^t\rangle$ towards 0 in such a way that the eigenvectors remain normalized $\langle\mathbf{u}_i^t|\mathbf{u}_i^t\rangle = 1$. The randomness comes in the second interaction and transverse term. We see that the $i$-th eigenvector $|\mathbf{u}_i^t\rangle$ trades more information with the eigenvectors $|\mathbf{u}_j^t\rangle, j \neq i$ that are associated to the closest neighboring eigenvalues $\lambda_j(t) \sim \lambda_i(t)$. If the neighboring eigenvalues $\lambda_j(t)$ are very close to $\lambda_i(t)$ (typically at a distance of order $1/N$ in the continuous part of the spectrum for large $N$), we shall see that this singular interaction leads to unstable (discontinuous) eigenstates trajectories with respect to time $t$, in the large $N$ limit (see below).

**12.1.2. Evolution of the mean squared overlaps at finite $N$.** In order to quantify the relationship between the perturbed eigenstates at time $t$ and the eigenstates at the initial time, we consider the scalar products or *overlaps* $\langle\mathbf{u}_i^t|\mathbf{u}_j^0\rangle$ for $i, j = 1, \dots, N$. Specifically, we investigate the mean square overlaps $\left[\langle\mathbf{u}_i^t|\mathbf{u}_j^0\rangle^2\right]_w$ where we use the notation $[\dots]_w$ for the expectation over the Brownian motions $w_{ij}, i \neq j \in \{1, \dots, N\}$ which appear in the eigenvectors evolution equation (12.1.2). Recall that those Brownian motions are *independent* of the eigenvalues so that this conditioning does not modify the law of the eigenvalue process. Note also that the variables $\left[\langle\mathbf{u}_i^t|\mathbf{u}_j^0\rangle^2\right]_w, 1 \leq i, j \leq N$ are still random, measurable with respect to the sigma field generated by the Brownian trajectories $\{(W_i(s)), 0 \leq s \leq t, i = 1, \cdots, N\}$.

For $j$ fixed, we find using Itô's formula (see e.g. [154] for a reminder):

$$\begin{aligned}
\mathrm{d}\langle\mathbf{u}_i^t|\mathbf{u}_j^0\rangle^2 &= 2\langle\mathbf{u}_i^t|\mathbf{u}_j^0\rangle\mathrm{d}\langle\mathbf{u}_i^t|\mathbf{u}_j^0\rangle + \frac{1}{N}\sum_{k=1}^{N}\frac{\langle\mathbf{u}_k^t|\mathbf{u}_j^0\rangle^2}{(\lambda_i(t)-\lambda_k(t))^2}\mathrm{d}t \\
&= \frac{1}{N}\sum_{k\neq i}\frac{\langle\mathbf{u}_k^t|\mathbf{u}_j^0\rangle^2 - \langle\mathbf{u}_i^t|\mathbf{u}_j^0\rangle^2}{(\lambda_i(t)-\lambda_k(t))^2}\mathrm{d}t + \frac{2}{\sqrt{N}}\sum_{k\neq i}\frac{\langle\mathbf{u}_i^t|\mathbf{u}_j^0\rangle\langle\mathbf{u}_k^t|\mathbf{u}_j^0\rangle}{\lambda_i(t)-\lambda_k(t)}\mathrm{d}w_{ik}(t).
\end{aligned}$$

We introduce the following short hand notation for the rescaled mean squared overlaps between the non-perturbed eigenstate $|\mathbf{u}_j^0\rangle$ and the perturbed eigenvectors $|\mathbf{u}_i^t\rangle$

$$\Phi_{i|j}(t) \equiv \Phi(\lambda_i(t), \mu_j) := N\left[\langle\mathbf{u}_i^t|\mathbf{u}_j^0\rangle^2\right]_w\,, \tag{12.1.4}$$

for $i = 1, \dots, N$. We see that this quantity is the analogue of Eq. (5.0.3) for the additive model and one notices that it is now a function of the time $t$. We can deduce from (12.1.4) that the mean squared overlaps satisfies an *autonomous* evolution equation,

$$\partial_t \Phi_{i|j}(t) = \frac{1}{N} \sum_{k \neq i} \frac{\Phi_{k|j}(t) - \Phi_{i|j}(t)}{(\lambda_k(t) - \lambda_i(t))^2} \quad \text{with} \quad \Phi_{i|j}(0) = N\delta_{ij} \,. \tag{12.1.5}$$

This evolution equation was discovered in 1995 by Wilkinson and Walker (see [190, Equation (4.7)]). It was also used to analyze the large dimensional statistics of Haar matrices in [32]. The equation (12.1.5) is the main tool used in the forthcoming sections to analyze the asymptotics of the overlaps in the large $N$-limit. Let us re emphasize the fact that the evolution equation (12.1.5) for $\Phi_{i|j}(t)$ depends only on the projections of the $j$-th eigenvector $|\mathbf{u}_j^0\rangle$ on the perturbed eigenstates $|\mathbf{u}_i^t\rangle$ and does not involve any other non-perturbed eigenvector $|\mathbf{u}_\ell^0\rangle, \ell \neq j$. This is a very convenient fact as we can fix a given non-perturbed eigenstate $|\mathbf{u}_j^0\rangle$ and work out the system of closed equations (12.1.5) satisfied by its $N$ projections on the perturbed eigenvectors $|\mathbf{u}_i^t\rangle, i = 1, \cdots, N$.

**12.1.3. Spectral density and spikes trajectories in the large $N$ limit.** In this section, we describe the evolution of the limiting eigenvalues density when $N \to \infty$. To simplify the notations, we introduce for any $k \in [\![1, N]\!]$:

$$p_k := \frac{k}{N} \in [0, 1] \,. \tag{12.1.6}$$

Then, in the asymptotic regime $N \to \infty$, we will always assume that the eigenvalues are smoothly allocated according to their "classical" position, that is to say

$$\mu_k = \mu(p_k) \quad \text{with} \quad p_k = \int_{\mu(p_k)}^{\infty} \rho_{\mathbf{C}}(x) \mathrm{d}x, \tag{12.1.7}$$

for any $k = 1, \dots, N$. Similarly, we make the same assumption regarding the eigenvalues $\lambda_i(t)$ of $\mathbf{M}(t)$ at any time $t$ fixed, that is to say

$$\lambda_i(t) = \lambda(p_i, t) \quad \text{with} \quad p_i = \int_{\lambda(p_i,t)}^{\infty} \rho_{\mathbf{M}}(x, t) \mathrm{d}x, \tag{12.1.8}$$

with $\lambda_i(0) = \mu_i$ for any $i = 1, \dots, N$. These technical assumptions is useful if $i := (i_N)_{N \in \mathbb{N}}$ is a sequence such that $i_N/N \to p \in (0, 1)$, then the $i$-th eigenvalue $\lambda_i(t) := (\lambda_{i_N}(t))$ converges (almost surely) towards $\lambda(p, t)$ as $N \to \infty$. This remarks also holds for the function $\mu$ defined in (12.1.7).

Let us study the infinitesimal increments over time of the empirical Stieltjes transform ,

$$\mathfrak{g}_N(z, t) := \int_{\mathbb{R}} \frac{\mu_t^N(d\lambda)}{z - \lambda} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{z - \lambda_i(t)}, \qquad \mu_t^N(\mathrm{d}\lambda) := \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i(t)}(\mathrm{d}\lambda) \,, \tag{12.1.9}$$

and using Itô's formula (see for instance [8, Subsection 4.3 page 248], [155] or more recently [4] in a slightly wider context), one obtains the following Burgers evolution equation for the Stieltjes transform ,

$$\mathrm{d}\mathfrak{g}_N(z, t) = -\mathfrak{g}_N(z, t) \partial_z \mathfrak{g}_N(z, t) \mathrm{d}t + \sqrt{\frac{2}{\beta N}} \sum_{i=1}^{N} \frac{1}{(z - \lambda_i)^2} dB_i + \frac{1}{2N} (\frac{2}{\beta} - 1) \partial_z^2 \mathfrak{g}_N(z, t) \mathrm{d}t \,. \tag{12.1.10}$$

Following (12.1.8), this evolution equation becomes deterministic in the large $N$ limit and the solution is the limiting Stieltjes transform $\mathfrak{g}$ of the limiting eigenvalues density $\rho(\cdot,t)$ of the matrix $\mathbf{M}(t)$ such that $\mu_t^N(\mathrm{d}\lambda) \to \rho(\lambda,t)\mathrm{d}\lambda$ when $N \to \infty$. The Stieltjes transform of the density $\rho(\lambda,t)$ is defined for $z \in \mathbb{C} \setminus \mathrm{supp}[\rho]$ as

$$\mathfrak{g}(z,t) = \int_{\mathbb{R}} \frac{\rho_{\mathbf{M}}(\lambda,t)}{z-\lambda}\mathrm{d}\lambda\,.$$

This analytic function characterizes the probability density $\rho(\cdot,t)$ that one can compute from the imaginary part of $\mathfrak{g}$ near the real axis thanks to the *Stieltjes inversion formula* $\mathrm{Im}\,\mathfrak{g}(\lambda - \mathrm{i}\varepsilon,t) \to_{\varepsilon\to 0} \pi\rho_{\mathbf{M}}(\lambda,t)$. Hence, we see by sending $N \to \infty$ in (12.1.10) that the dynamics of the Stieltjes transform $\mathfrak{g}$ is governed by a *Burgers* evolution equation

$$\partial_t \mathfrak{g}(z,t) = -\mathfrak{g}(z,t)\partial_z \mathfrak{g}(z,t), \qquad \text{with} \qquad \mathfrak{g}(z,0) = \int_{\mathbb{R}} \frac{\rho_{\mathbf{C}}(\lambda)}{z-\lambda}\mathrm{d}\lambda\,. \tag{12.1.11}$$

Interestingly, the solution of (12.1.11) is known [158] to satisfy the fixed point equation (see also [3, Proposition 4.1])

$$\mathfrak{g}(z,t) = \int_0^1 \frac{\mathrm{d}p}{z - \mu(p) - t\mathfrak{g}(z,t)} \tag{12.1.12}$$

where $\mu : [0,1] \to \mathbb{R}$ is the continuous function introduced in (12.1.7) mapping the index $x \in [0,1]$ to the eigenvalue $a(x)$ of the matrix $\mathbf{C}$ in the continuous limit $N \to \infty$. In the special case $\mathbf{C} = 0$, all eigenvalues start from the origin, i.e. $\mu(p) = 0$ for any $p \in [0,1]$ and the solution $\mathfrak{g}$ is fully explicit corresponding to the Wigner semi-circle density $\rho_{\mathbf{M}}(\lambda,t) = \frac{1}{2\pi t}\sqrt{4t-\lambda^2}$ with radius $2\sqrt{t}$.

One can also write the evolution equation directly in terms of the density $\rho(\lambda,t)$ itself by projecting the Burgers equation (12.1.11) on the real line thanks to the Stieltjes inversion formula: for $\lambda \in \mathbb{R}$ and $t \geq 0$,

$$\partial_t \rho_{\mathbf{M}}(\lambda,t) + \partial_\lambda\left(v_{\mathbf{M}}(\lambda,t)\rho_{\mathbf{M}}(\lambda,t)\right) = 0 \qquad \text{where} \qquad v_{\mathbf{M}}(\lambda,t) = \fint_{\mathbb{R}} \frac{\rho_{\mathbf{M}}(\lambda',t)}{\lambda-\lambda'}\mathrm{d}\lambda' \tag{12.1.13}$$
$$\text{and with the initial condition} \qquad \rho_{\mathbf{M}}(\lambda,0) = \rho_{\mathbf{C}}(\lambda)\,.$$

From (12.1.8), we obtain $\partial_p \lambda(p,t) = -1/\rho_{\mathbf{M}}(\lambda(p,t),t)$ and using (12.1.13), we find

$$\partial_t \lambda(p,t) = v_{\mathbf{M}}(\lambda(p,t),t)\,. \tag{12.1.14}$$

This gives a clear physical interpretation of the function $v_{\mathbf{M}}(\lambda(p,t),t)$ as the speed of the particles in the scaling limit.

The spikes trajectories are also expected to become deterministic in the large $N$ limit. Indeed, we can compute them by sending $N \to +\infty$ directly in the Dyson Brownian motion equation (12.1.1) for $j = 1,\dots,r$ to find

$$\dot{\lambda}_j(t) = \int_{\mathbb{R}} \frac{\rho_{\mathbf{M}}(\lambda,t)}{\lambda_j(t)-\lambda}d\lambda \qquad \text{with} \qquad \lambda_j(0) = \mu_j\,. \tag{12.1.15}$$

The limiting path of the spike $\lambda_j(t)$ for $j = 1,\dots,r$ is thus driven by the density $\rho_{\mathbf{M}}(\cdot,t)$ satisfying (12.1.13). Notice that we use the same notation for the spike trajectories $\lambda_j(t)$ for

both the limiting case $N \to \infty$ and the finite dimensional case $N < \infty$ [2]. At the initial time $t = 0$, the spike $\lambda_j(t)$ starts from a position $\mu_j$ outside the bulk of the spectrum of $\mathbf{C}$. As $t$ increases, the spike $\lambda_j(t)$ is pushed away with an electrostatic force exerted by the other particles. Each particle inside the bulk of the spectrum exerts a force which is proportional to the inverse of its distance to the spike. In such a way, the spike remains at a non-negative distance to the bulk at any time $t \geq 0$. Nonetheless, as illustrated in the next subsection, we shall notice that the spike may eventually by caught back by the continuous part of the spectrum.

**12.1.4. Factor model.** In this subsection, we illustrate the results of the previous subsection by analyzing explicitly the special case where the matrix $\mathbf{C}$ is of low rank $r$ compared to the dimension, $r \ll N$. Such factor models are used in applications in biology to study population dynamics [192] or in finance where the setting is nevertheless slightly different, as explained in Section 4.3.2. The matrix $\mathbf{C}$ has $r$ spikes $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$ and 0 is an eigenvalue of $\mathbf{C}$ with multiplicity $N - r \sim O(N)$. The structure of the matrix $\mathbf{C}$ is therefore very simple with only a few relevant factors that one wants to estimate.

The few spikes do not bring any macroscopic contribution to the empirical density $\rho_{\mathbf{M}}(\cdot, t)$ of the particles and in the large $N$ limit, we recover the Wigner semicircle density centered at 0 with radius $2\sqrt{t}$,

$$\rho_{\mathbf{M}}(\lambda, t) = \frac{1}{2\pi t}\sqrt{4t - \lambda^2}\,, \quad -2\sqrt{t} \leq \lambda \leq 2\sqrt{t}\,. \tag{12.1.16}$$

The speed of the particles inside the spectrum can be computed explicitly as well: it is linear given for $t > 0, |\lambda| \leq 2\sqrt{t}$ by

$$v_{\mathbf{M}}(\lambda, t) = \frac{\lambda}{2t}\,.$$

With such a simple form (12.1.16) for the limiting density of particles, it turns out that the ordinary differential equation (12.1.15) can be solved explicitly and we obtain for any $j = 1, \ldots, r$,

$$\lambda_j(t) = \mu_j + \frac{t}{\mu_j}\,.$$

Comparing this value of the $j$-th spike with the value of the edges of the spectrum at time $t$, we easily check that for any $\mu_j \neq 0$, the bulk eventually catches up the isolated particle $\lambda_j(t)$ at the critical time $t_c^j = \mu_j^2$ at which $\lambda_j(t_c^j) = 2\sqrt{t_c^j}$. Beyond this critical time, the spike is "swallowed" by the Wigner sea and disappears. See Fig. 12.1.1 for an illustration of a sample path of the eigenvalues of $\mathbf{M}(t)$ when the initial matrix $\mathbf{C}$ has rank one.

## 12.2 Eigenvector in the bulk of the spectrum

In the bulk of the spectrum, the mean spacings $\delta\mu$ between the eigenvalues of the matrix $\mathbf{C}$ is approximately of order $1/N$ and depends on the position $\mu$ in the spectrum and the local density $\rho_{\mathbf{C}}(\mu)$ of particles near $\mu$ as $\delta\mu \sim 1/(N\rho_{\mathbf{C}}(\mu))$.

[2] In order to avoid heavy notations, we omit to use an additional super script $N$ for the eigenvalues $\lambda_i^N(t)$ at finite $N$.

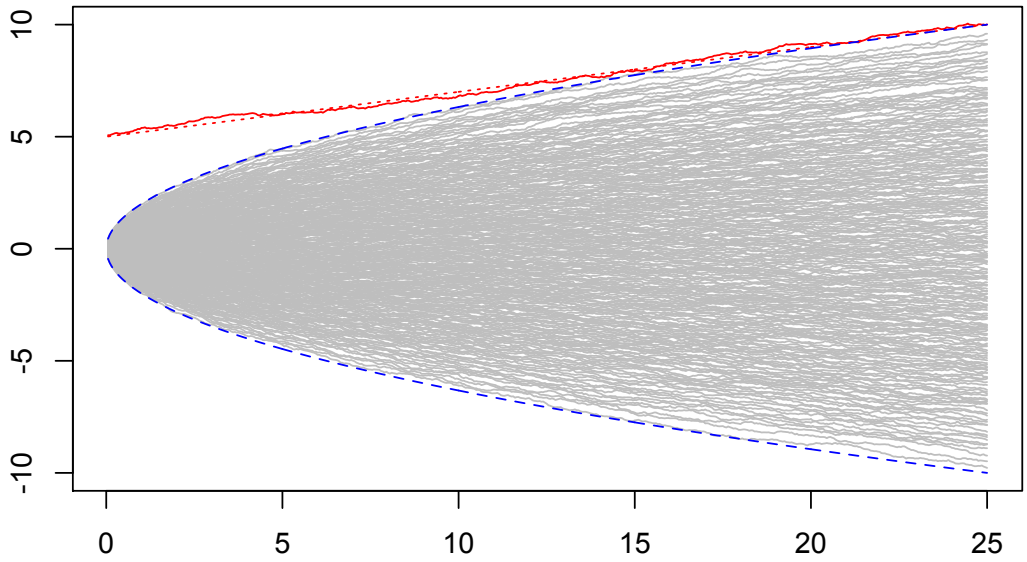



FIGURE 12.1.1. (Color online). Sample trajectories of the eigenvalues of the matrix $\mathbf{M}(t)$ defined in (12.0.1) where $\mathbf{C}$ has only one non zero eigenvalue $\mu_1 = 5$, as a function of time $t > 0$. The grey lines represent the eigenvalues trajectories in the bulk. The blue dashed lines represent the trajectories of the edges $\pm 2\sqrt{t}$. The red plain line is the sample path of the spike $\lambda_1(t)$ and the red dashed line is $5 + t/5$. Beyond $t_c = 25$, the spike eigenvalue $\lambda_1(t)$ is "swallowed" by the Wigner sea and disappears.

When one perturbs the initial matrix $\mathbf{C}$ by adding the matrix $\mathbf{B}(t)$, it is well known that one should compare the magnitude of the entries of the perturbation $B_{ij}(t) \sim \sqrt{t}/\sqrt{N}$ with the mean level spacing $\delta a$ of the non-perturbed matrix $\mathbf{C}$.

There are therefore three distinct regimes of perturbation which lead to different asymptotics for the relation between the perturbed and non-perturbed eigenstates:

(i) The microscopic or perturbative regime corresponds to values of $t := t_N$ depending on $N$ such that

$$t_N \ll \frac{1}{N}\,.$$

For such values of $t := t_N$, the perturbation matrix $\mathbf{B}(t)$ is in fact asymptotically small compared to $\mathbf{C}$ and for any fixed $i$, the eigenvector $|\mathbf{u}_j^t\rangle$ of $\mathbf{M}(t)$ converge to those of $\mathbf{C}$ when $N \to \infty$ in the $L^2$ norm,

$$||\mathbf{u}_j^{t_N} - \mathbf{u}_j^0||_2 \longrightarrow 0\,. \tag{12.2.1}$$

One can even obtain an asymptotic expansion for $|\mathbf{u}_j^t\rangle$ around $|\mathbf{u}_j^0\rangle$ using (12.1.2). This regime is rather trivial and will not be further considered here.

(ii) The second mesoscopic regime establishes a smooth crossover between the microscopic and macroscopic regimes. It corresponds to values of $t := t_N$ which are inversely proportional to $N$ i.e. such that there exists $\tau > 0$ fixed such that

$$t_N = \frac{\tau}{N\rho_{\mathbf{C}}(\mu_j)}\,.$$

Although the operator norm of the matrix $\mathbf{B}(t_N)$ tends to 0 when $N \to \infty$, this regime is non-perturbative in the sense that we do not have the convergence (12.2.1) of $|\mathbf{u}_i^t\rangle$ towards $|\mathbf{u}_i^0\rangle$. This non trivial rotation of the eigenvectors may appear surprising at first sight (it is generated by the addition of a microscopic perturbation) but is in fact simply due to the small spacings $\delta a$ between the eigenvalues of $\mathbf{C}$ in the bulk of the spectrum. This regime is studied in details in [5].

(iii) The macroscopic regime corresponds to values of $t = \mathcal{O}(1)$ that do not depend on $N$ and this is the one we shall focus on in the following. Even though the perturbation is macroscopic, we are still able to extract information on the non-perturbed eigenstate $|\mathbf{u}_j^0\rangle = |\phi_j\rangle$ from the observation of the perturbed eigenstates $|\mathbf{u}_i^t\rangle$ for general $i, j$. Indeed we compute explicitly the asymptotic mean overlaps $\Phi_{i|j}(t)$, defined in (5.0.4), which are proportional to $1/N$ in the large $N$ limit, using again the overlap equation (12.1.5). If $i := (i_N)_{n\in\mathbb{N}}$ and $j := (i_N)_{n\in\mathbb{N}}$ are sequences such that $i_N/N \to p \in (0,1)$ and $j_N/N \to p' \in (0,1)$ when $N \to \infty$, our result reads (see below)

$$\Phi_{i|j}(t) \underset{N\to\infty}{\sim} \Phi(\lambda(p,t), \mu(p'))$$

where the function $\Phi(\lambda(p,t), \mu(p')) \sim \mathcal{O}(1)$ is determined explicitly for any matrix $\mathbf{C}$ in terms of the trajectory of the limiting density $(\rho_{\mathbf{M}}(\cdot, s))_{0\le s\le t}$ described in (12.1.13).

Throughout the following, we consider the regime (iii), i.e. the case where $t > 0$ is fixed independently of $N$. In this regime, we expect the distribution of the overlaps to be much more spread out compared to the other regimes: the non-perturbed eigenstates are delocalized in the basis of the perturbed eigenvectors. All the mean squared overlaps have the same order of magnitude of order $1/N$ for large $N$. We start by deriving the evolution equation of the local density of the state $|\mathbf{u}_j^0\rangle$ for a fixed index $j$.

**12.2.1. Local density of state.** The local density of the state $|\mathbf{u}_j^0\rangle$ describes the allocation of the mean squared projections of the non-perturbed state $|\mathbf{u}_j^0\rangle$ on the basis of the perturbed eigenvectors $|\mathbf{u}_i^t\rangle$. It is a probability measure defined as

$$\nu_N^{(j,t)}(d\lambda) := \frac{1}{N}\sum_{i=1}^{N} \Phi_{i|j}(t)\,\delta_{\lambda_i(t)}(d\lambda)$$

Let us denote by $U_N(z,t)$ the empirical Stieltjes transform of this probability measure

$$U_N^{(j)}(z,t) \;:=\; \frac{1}{N}\sum_{i=1}^{N} \frac{\Phi_{i|j}(t)}{z-\lambda_i(t)}\,.$$

It we define the resolvent of $\mathbf{M}(t)$ by $\mathbf{G}(z,t) := (z\mathbf{I}_N - \mathbf{M}(t))^{-1}$, we have that $U_N^{(j)}(z,t)$ is equal, for any $j$ and $z \in \mathbb{C}\setminus \mathrm{supp}[\rho_{\mathbf{M}}]$, to:

$$U_N^{(j)}(z,t) = \langle \mathbf{u}_j^0 | G_{ii}(z,t) | \mathbf{u}_j^0 \rangle\,. \tag{12.2.2}$$

Hence, in the basis where $\mathbf{C}$ is diagonal, $U_N^{(j)}(z,t)$ corresponds to the diagonal entries of $\mathbf{G}_(z,t)$. Now, using Dyson equation for the eigenvalues (12.1.1) and the overlap evolution equation (12.1.5), we obtain the following evolution equation for the local resolvent:

$$\partial_t U_N^{(j)}(z,t) = -\mathfrak{g}_N(z,t)\partial_z U_N^{(j)}(z,t) + \sqrt{\frac{2}{\beta N}}\sum_{i=1}^{N}\frac{\Phi_{i|j}(t)}{(z-\lambda_i)^2}\frac{\mathrm{d}b_i}{\mathrm{d}t} + \frac{1}{2N}(\frac{2}{\beta}-1)\partial_z^2 U_N(z,t), \tag{12.2.3}$$

$$U_N^{(j)}(z,0) = \frac{1}{z-\mu_j}$$

where $\mathfrak{g}_N(z,t)$ is defined in Eq. (12.1.9) and satisfies the Burgers equation (12.1.10). We remark that summing over $j=1,\dots,N$ in (12.2.3) actually yields the Burgers equation (12.1.10). The derivation of Eq. (12.2.3) is given at the end of this section.

By invoking the same arguments as above, we expect the stochastic partial differential equation (12.2.3) to become deterministic in the large $N$-limit. We denote by $U(z,\mu(p),t)$, with $p\in[0,1]$, the limiting value of $U_N^{(j)}(z,t)$ when $N\to+\infty$. The equation on the limiting local resolvent $U(z,\mu(p),t)$ reads

$$\partial_t U(z,\mu(p),t) = -\mathfrak{g}(z,t)\partial_z U(z,\mu(p),t), \qquad \text{with} \qquad U(z,\mu(p),0) = \frac{1}{z-\mu(p)}\,, \tag{12.2.4}$$

where $\mathfrak{g}$ satisfies the limiting Burgers equation (12.1.11). Recalling the fixed point equation Eq. (12.1.12) satisfied by $\mathfrak{g}(z,t)$, the solution $U(z,\mu(p),t)$ such that

$$\mathfrak{g}(z,t) = \int_0^1 U(z,\mu(p),t)\mathrm{d}x$$

is actually given, for any $p\in(0,1), z\in\mathbb{C}\setminus\mathrm{supp}[\rho_{\mathbf{M}}], t\geq 0$ by

$$U(z,\mu(p),t) = \frac{1}{z-\mu(p)-t\mathfrak{g}(z,t)}\,. \tag{12.2.5}$$

This explicit solution of (12.2.4) is quite remarkable. This limiting result was already obtained by Shlyakhtenko in [158] using Free probability theory. We think the Dyson style approach developed here is very intuitive, shedding new lights on this result.

Using the result (12.2.5), we may now easily derive the mean squared overlap. Indeed, we recall that the function $U$ satisfies Eq. (12.2.2) in the large $N$ limit. Then, we invoke the inversion formula (5.1.5) to find for $N\to\infty$:

$$\Phi(\lambda_i(t),\mu_j) = \frac{t}{\left|\lambda_i(t)-t\mathfrak{g}_{\mathbf{M}}(z,t)-\mu_j\right|^2}\,, \tag{12.2.6}$$

where $\mathfrak{g}_{\mathbf{M}}$ is the Stieltjes transform of $\mathbf{M}$ that satisfies Eq. (12.1.11). This result is in perfect adequation with [3,150] with $t$ that plays the role of the variance.

Let us now derive Eq. (12.2.3). Thanks to Itô's formula, we get

$$\begin{aligned}
\partial_t U_N(z,t) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z-\lambda_i(t)}\sum_{k\neq i}\frac{\Phi_{k|j}(t)-\Phi_{i|j}(t)}{(\lambda_k-\lambda_i)^2} + \sum_{i=1}^{N}\frac{\Phi_{i|j}(t)}{(z-\lambda_i(t))^2}\frac{d\lambda_i}{dt} + \frac{2}{\beta N}\sum_{i=1}^{N}\frac{\Phi_{i|j}(t)}{(z-\lambda_i(t))^3}\\
&= \frac{1}{2N}\sum_{i\neq k}\frac{\Phi_{k|j}(t)-\Phi_{i|j}(t)}{\lambda_i-\lambda_k}\frac{1}{(z-\lambda_i)(z-\lambda_k)}\\
&+ \sum_{i=1}^{N}\frac{\Phi_{i|j}(t)}{(z-\lambda_i(t))^2}\left(\sqrt{\frac{2}{\beta N}}\frac{\mathrm{d}b_i}{dt} + \frac{1}{N}\sum_{k\neq i}\frac{1}{\lambda_i-\lambda_k}\right) + \frac{1}{\beta N}\partial_z^2 U_N(z,t)
\end{aligned}$$

where we have used the classical symmetrization trick to obtain the second line.

Now, the trick is to rewrite the first term as

$$\frac{1}{2N}\sum_{i\neq k}\frac{\Phi_{k|j}(t)-\Phi_{i|j}(t)}{\lambda_i-\lambda_k}\frac{1}{(z-\lambda_i)(z-\lambda_k)}$$
$$=\frac{1}{2N}\sum_{k=1}^{N}\frac{\Phi_{k|j}(t)}{z-\lambda_k}\sum_{i\neq k}\frac{1}{(\lambda_i-\lambda_k)(z-\lambda_i)}-\frac{1}{2N}\sum_{i=1}^{N}\frac{\Phi_{i|j}(t)}{z-\lambda_i}\sum_{k\neq i}\frac{1}{(\lambda_i-\lambda_k)(z-\lambda_k)}\,.$$

We notice that

$$\sum_{i\neq k}\frac{1}{(\lambda_i-\lambda_k)(z-\lambda_i)}=\frac{1}{z-\lambda_k}\sum_{i\neq k}\frac{1}{z-\lambda_i}+\frac{1}{\lambda_i-\lambda_k}$$

and

$$\sum_{k\neq i}\frac{1}{(\lambda_i-\lambda_k)(z-\lambda_k)}=-\frac{1}{z-\lambda_i}\sum_{k\neq i}\frac{1}{z-\lambda_k}-\frac{1}{\lambda_i-\lambda_k}\,.$$

Therefore we deduce that

$$\frac{1}{2N}\sum_{i\neq k}\frac{\Phi_{k|j}(t)-\Phi_{i|j}(t)}{\lambda_i-\lambda_k}\frac{1}{(z-\lambda_i)(z-\lambda_k)}=\sum_{k=1}^{N}\frac{\Phi_{k|j}(t)}{(z-\lambda_k)^2}\frac{1}{N}\sum_{i\neq k}\frac{1}{z-\lambda_i}+\sum_{k=1}^{N}\frac{\Phi_{k|j}(t)}{(z-\lambda_k)^2}\frac{1}{N}\sum_{i\neq k}\frac{1}{\lambda_i-\lambda_k}$$
$$=\sum_{k=1}^{N}\frac{\Phi_{k|j}(t)}{(z-\lambda_k)^2}\left(\mathfrak{g}_N(z,t)-\frac{1}{N}\frac{1}{z-\lambda_k}\right)+\sum_{k=1}^{N}\frac{\Phi_{k|j}(t)}{(z-\lambda_k)^2}\frac{1}{N}\sum_{i\neq k}\frac{1}{\lambda_i-\lambda_k}$$

where $\mathfrak{g}_N(z,t):=\frac{1}{N}\sum_{i=1}\frac{1}{z-\lambda_i(t)}$. As a conclusion, we deduce that

$$\partial_t U_N(z,t)=-\mathfrak{g}_N(z,t)\partial_z U_N(z,t)+\sqrt{\frac{2}{\beta N}}\sum_{i=1}^{N}\frac{\Phi_{i|j}(t)}{(z-\lambda_i)^2}\frac{\mathrm{d}b_i}{dt}+\frac{1}{2N}(\frac{2}{\beta}-1)\partial_z^2 U_N(z,t)\,.$$

**12.2.2. An alternative derivation of Eq. (12.2.5).** In this subsection, we present an alternative approach that considers directly the time evolution of the full matrix (12.0.2), which we have not seen in the literature before. To that end, we define the time dependent resolvent

$$\mathbf{G}(z,t)\;:=\;\mathbf{H}^{-1}(z,t),\qquad \mathbf{H}(z,t)\;:=\;z\mathbf{I}_N-\mathbf{M}(t). \tag{12.2.7}$$

Using Itô formula and the fact that $\mathrm{d}M_{kl}=\mathrm{d}B_{kl}$, one has

$$\mathrm{d}G_{ij}(z,t)\;\;=\;\;\sum_{k,l=1}^{N}\frac{\partial G_{ij}}{\partial M_{kl}}\mathrm{d}B_{kl}+\frac{1}{2}\sum_{k,l,m,n=1}^{N}\sum_{m,n=1}^{N}\frac{\partial^2 G_{ij}}{\partial M_{kl}\partial M_{mn}}\mathrm{d}\big[B_{kl}B_{mn}\big], \tag{12.2.8}$$

Next, we compute the derivatives:

$$\frac{\partial G_{ij}}{\partial M_{kl}}=\frac{1}{2}\left[G_{ik}G_{jl}+G_{jk}G_{il}\right], \tag{12.2.9}$$

from which we deduce the second derivatives

$$\frac{\partial^2 G_{ij}}{\partial M_{kl}\partial M_{mn}}=\frac{1}{4}\left[\left(G_{im}G_{kn}+G_{im}G_{kn}\right)G_{jl}+...\right], \tag{12.2.10}$$

where we have not written the other 6 $GGG$ products. Now, using Eqs. (12.0.2) and (11.1.2), the quadratic covariation reads

$$\mathrm{d}\big[B_{kl}B_{mn}\big]=\frac{\mathrm{d}t}{N}\left(2\delta_{k=l=m=n}+\delta_{k=m}\delta_{l=n}+\delta_{k=n}\delta_{l=m}\right) \tag{12.2.11}$$

so that we get from (12.2.8) and taking into account symmetries:

$$\mathrm{d}G_{ij}(z,t) = \sum_{k,l=1}^{N} G_{ik}G_{jl}\mathrm{d}B_{kl} + \frac{1}{N}\sum_{k,l=1}^{N}\Big(G_{ik}G_{lk}G_{lj} + G_{ik}G_{kj}G_{ll}\Big)\mathrm{d}t\,. \tag{12.2.12}$$

If we now take the average over with respect to the Brownian motion $W_{kl}$ defined in Eq. (12.0.2), we find the following evolution for the average resolvent:

$$\partial_t \mathbb{E}[\mathbf{G}(z,t)] \;=\; \mathfrak{g}_{\mathbf{M}}(z,t)\,\mathbb{E}[\mathbf{G}^2(z,t)] + \frac{1}{N}\mathbb{E}[\mathbf{G}^3(z,t)]. \tag{12.2.13}$$

Now, one can notice that:

$$\mathbf{G}^2(z,t) = -\partial_z \mathbf{G}(z,t); \qquad \mathbf{G}^3(z,t) = \partial_z^2 \mathbf{G}(z,t), \tag{12.2.14}$$

which hold even before averaging. By sending $N \to \infty$, we obtain the following matrix PDE for the resolvent:

$$\partial_t \mathbb{E}[\mathbf{G}(z,t)] \;=\; -\mathfrak{g}_{\mathbf{M}}(z,t)\,\partial_z \mathbb{E}[\mathbf{G}(z,t)]\,, \quad \text{with} \quad \mathbf{G}(z,0) \;=\; \mathbf{G}_{\mathbf{C}}(z)\,. \tag{12.2.15}$$

The solution of Eq. (12.2.15) reads [5, 158]:

$$\mathbf{G}_{\mathbf{M}}(z,t) = \mathbf{G}_{\mathbf{C}}(Z(z,t)), \tag{12.2.16}$$

and this is exactly equivalent to (12.2.5) if we place ourselves in the basis where $\mathbf{C}$ is diagonal. We see by taking the normalized trace into this latter equation that we retrieve the standard Burgers equation for the Stieltjes transform (12.1.11), as it should. Note that the result (12.2.16) holds is universal in the sense that it also works when the entries are not Gaussian (see [109] for details).

## 12.3 Isolated eigenvectors

In this section, we study the projections of a given initial eigenstate $|\mathbf{u}_j^0\rangle$, associated to an eigenvalue $\mu_j$ lying outside the bulk of the spectrum of the initial matrix $\mathbf{C}$, on the perturbed eigenvectors in the limit of large dimension $N$. We recall that the eigenvalues $\mu_i$ are indexed in non-increasing order and we denote by $r$ the number of outliers (which does not depend on $N$). To fix ideas and simplify notations, we will suppose that $r = 1$ so that the eigenvalue $\mu_1$ is the largest spike (see Fig. 12.3.1) of the matrix $\mathbf{C}$.

**12.3.1. Principal component.** From the eigenvector evolution equation (12.1.2) and the definition (5.0.4), we easily check that

$$\mathrm{d}\left[\langle \mathbf{u}_1^t | \mathbf{u}_1^0\rangle\right]_w = -\frac{1}{2N}\sum_{k\neq 1}\frac{\left[\langle \mathbf{u}_1^t | \mathbf{u}_1^0\rangle\right]_w \mathrm{d}t}{(\lambda_1(t)-\lambda_k(t))^2}\,,$$

where we recall that $[\cdots]_w$ denotes the average over the Brownian motion $\{w_{ij}\}_{i,j\in[\![1,N]\!]}$. This ordinary differential equation can be solved explicitly and, using the initial condition, we obtain the following equality, valid for any finite $N$,

$$\left[\langle \mathbf{u}_1^t | \mathbf{u}_1^0\rangle\right]_w = \exp\left(-\frac{1}{2N}\int_0^t \sum_{k\neq 1}\frac{\mathrm{d}s}{(\lambda_1(s)-\lambda_k(s))^2}\right)\,.$$

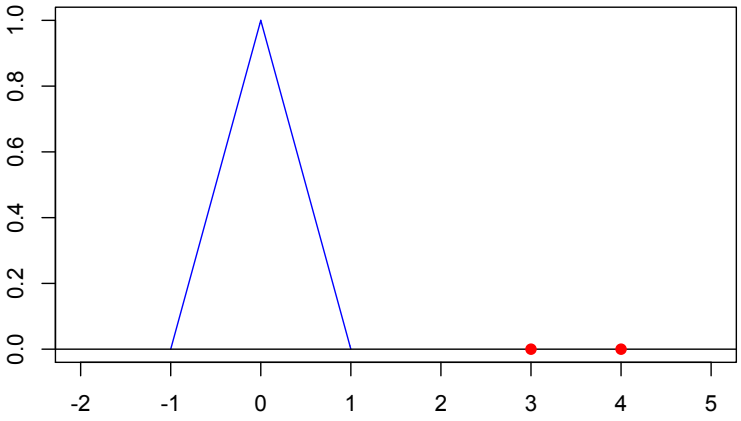



FIGURE 12.3.1. Spectrum of a matrix $\mathbf{C}$ satisfying our hypothesis with a continuous triangular density and $r = 2$ spikes $\mu_1 = 4, \mu_2 = 3$.

Sending $N \to \infty$ and using the results explained in subsection 12.1.3, we get,

$$\left[\langle \mathbf{u}_1^t | \mathbf{u}_1^0 \rangle\right]_w \sim \exp\left(-\frac{1}{2}\int_0^t ds \int_{\mathbb{R}} \frac{\rho_{\mathbf{M}}(\lambda, s)}{(\lambda_1(s) - \lambda)^2}\, \mathrm{d}\lambda\right) \tag{12.3.1}$$

where $(\lambda_1(s))_{0 \le s \le t}$ is the limiting trajectory of the first spike (already described in (12.1.15)) such that

$$\dot{\lambda}_1(s) = \int_{\mathbb{R}} \frac{\rho(\lambda, s)}{\lambda_1(s) - \lambda} \mathrm{d}\lambda\,, \quad \lambda_1(0) = \mu_1\,. \tag{12.3.2}$$

The convergence (12.3.1) can of course be extended to any finite fixed value of $r$, with similar asymptotic formulas.

We see that if $t$ is small enough so that the spike is still isolated from the bulk at time $t$, then the overlap between the initial top eigenvector and its perturbed version does not vanish in the large $N$ limit even though $t$ and $\mathbf{B}_t$ have macroscopic sizes, in contrast with the bulk overlaps which were of order $1/N$.

**12.3.2. Transverse components.** We now consider the overlaps between the initial top eigenvector $|\mathbf{u}_1^0\rangle$ and the perturbed eigenvectors $|\mathbf{u}_i^t\rangle$ for $i \neq 1$. In order to lighten the notation, we define

$$\Phi(p, t) \;\equiv\; \Phi(\lambda(p, t), t)\,, \tag{12.3.3}$$

which is the asymptotic limit of $\Phi_{i|1}(t)$.

Since the eigenvalue $\lambda_1(t)$ is isolated from the other eigenvalues, we expect the overlaps between the corresponding perturbed and non-perturbed eigenvectors to be microscopic of order $1/N$.

To prove this, we start again from the overlap equation (12.1.5), and we denote by $f(t)$ the limit of $[\langle \mathbf{u}_1^0 | \mathbf{u}_1^t \rangle^2]_w$ when $N \to +\infty$ which is easily deduce from the RHS of Eq. (12.3.1):

$$f(t) := \exp\left( - \int_0^t ds \int_{\mathbb{R}} \frac{\rho_{\mathbf{M}}(\lambda, s)}{(\lambda_1(s) - \lambda)^2} \, d\lambda \right) . \tag{12.3.4}$$

Then, from (12.1.5), we may now derive the Cauchy problem satisfied by the limiting family of overlaps $\Phi_{i|1}(t)$ for $i > 1$ and $N \to \infty$,

$$\partial_t \, \Phi_{i|1}(t) \sim \int_0^1 \frac{\Phi(p,t) - \Phi(p_i,t)}{(\lambda(p,t) - \lambda(p_i,t))^2} \, \mathrm{d}p + \frac{f(t)}{(\lambda_1(t) - \lambda(p_i,t))^2} , \quad \Phi(p_i, 0) = 0 . \tag{12.3.5}$$

Note that the solution of (12.3.5) satisfies $\Phi(o,t) \geq 0$ for all $t \geq 0$ and any $p$ in the bulk of the spectrum, as it should be for a mean squared overlap. We notice that unlike the principal component (12.3.1), the solution $\Phi(p_i, t)$ is delocalized for any $i \neq 1$ in the large $N$ limit (see (12.1.4) for the definition of $\Phi$).

**12.3.3. Gaussian fluctuations of the principal component.** Using the convergence (12.3.5) of the transverse overlaps, we can compute the higher order moments of the principal component and deduce that the random variable $\langle \mathbf{u}_1^t | \mathbf{u}_1^0 \rangle$ is asymptotically a Gaussian variable with mean value $\sqrt{f(t)}$ as defined in (12.3.4) and variance of order $1/N$ that we are able to compute explicitly.

In this subsection , we work with a time $t > 0$ small enough so that the spike $(\lambda_i(s))_{0 \leq s \leq t}$ has not yet been swallowed by the limiting bulk density $(\rho_{\mathbf{M}}(\lambda, s))_{0 \leq s \leq t}$ of the Gaussian matrix process $(\mathbf{M}(s))_{0 \leq s \leq t}$. This critical time $t_c$ was explicitly computed in section 12.1.4 in the case of a small initial rank for the matrix $A$.

For such a time $t < t_c$, we shall now prove that conditionally to the eigenvalues path $(\lambda_i(s))_{s<t}, i = 1, \ldots, N$, the random variable

$$\eta_1(t) \; := \; \sqrt{N} \, \left( \langle \mathbf{u}_1^t | \mathbf{u}_1^0 \rangle - \left[ \langle \mathbf{u}_1^t | \mathbf{u}_1^0 \rangle \right]_w \right)$$

converges weakly towards a centered Gaussian distribution with variance

$$\varsigma_1^2(t) := \int_0^t \mathrm{d}s \exp\left( - \int_s^t \int_{\mathbb{R}} \frac{\rho_{\mathbf{M}}(\lambda, u)}{(\lambda_1(u) - \lambda)^2} \, \mathrm{d}\lambda \, \mathrm{d}u \right) \int_{\mathbb{R}} \frac{w(x,s)}{(\lambda_1(s) - x)^2} \rho(x,s) \mathrm{d}x ,$$

where

- $(\lambda_1(s))_{0 \leq s \leq t}$ is the limiting trajectory of the largest eigenvalue satisfying (12.1.15);
- $(\rho(\lambda, s))_{0 \leq s \leq t, \lambda \in \mathbb{R}}$ is the limiting bulk density trajectory satisfying (12.1.11);
- $(w(\lambda, t)_{0 \leq s \leq t, \lambda \in \mathbb{R}}) = (\Phi(p,s))_{0 \leq s \leq t, p \in [0,1]}$ is the function describing the limiting transverse overlaps satisfying the evolution equation (12.3.5).

Indeed, to prove this claim, we introduce the characteristic function

$$F_N(\xi, t) \; := \; \mathbb{E}\left[ \exp\left( \mathrm{i} \, \xi \, \eta_1(t) \right) \right] ,$$

and we obtain thanks to Itô's formula that $F_N$ satisfies the partial differential equation

$$\frac{\partial}{\partial t} F_N(\xi, t) = - \frac{\xi}{2N} \frac{\partial}{\partial \xi} F_N(\xi, t) \sum_{k \neq 1} \frac{1}{(\lambda_1 - \lambda_k)^2} - \frac{\xi^2}{2} h_N(t) F_N(\xi, t) .$$

In the scaling limit $N \to \infty$, this equation becomes

$$\frac{\partial}{\partial t}F(\xi,t) = -\frac{\xi}{2}\frac{\partial}{\partial \xi}F(\xi,t)\int_{\mathbb{R}}\frac{\rho(\lambda,t)}{(\lambda_1(t)-\lambda)^2}d\lambda - \frac{\xi^2}{2}h(t)F(\xi,t)$$

which is satisfied by the Gaussian characteristic function $F(\xi,t) = \exp(-\frac{\xi^2}{2}\varsigma_1^2(t))$. A more explicit proof of this result using the moment method can be found in [5].

**12.3.4. Estimation of the main factors.** As an illustration of the results obtained in the previous subsection, we come back on the factor model. In section 12.1.4, we have seen that the limiting density of eigenvalues is the Wigner semicircle with radius $2\sqrt{t}$ at time $t$ and that the limiting trajectories of the spikes are $\lambda_j(t) = \mu_j + t/\mu_j$ for any $j \leqslant r$.

It turns out that the limiting mean square overlap between the first non-perturbed and perturbed eigenvectors (respectively $|\mathbf{u}_1^0\rangle$ and $|\mathbf{u}_1^t\rangle$) can also be computed analytically.

From (12.3.1), we obtain:

$$\left[\langle \mathbf{u}_1^t|\mathbf{u}_1^0\rangle\right]_w \to \exp\left(-\frac{1}{2}\int_0^t \frac{\mathrm{d}s}{2\pi s}\int_{\mathbb{R}}\frac{\sqrt{4s-\lambda^2}}{(\mu_1+\frac{s}{\mu_1}-\lambda)^2}\,d\lambda\right) = \sqrt{\max\left(1-\frac{t}{\mu_1^2},0\right)}\,.$$

We see that the information contained in the perturbed eigenvector is completely lost at the time $t_c = \mu_1^2$, i.e. when the spike $\lambda_1(t_c)$ is swallowed by the Wigner sea. Moreover, from the results of subsection 12.3.3, we conclude that the random variable $\sqrt{N}\langle \mathbf{u}_1^0|\mathbf{u}_1^t\rangle$ has Gaussian fluctuations in the large $N$-limit around its mean asymptotic value $1 - t/\mu_1^2$ for $t \leq \mu_1^2$.

## 12.4 Eigenvectors between correlated deformed GOE

As in Section 5.2, we can derive the overlaps between two independent deformed GOEs and the result is very similar to sample covariance matrices. Hence, we shall omit most details that can be obtained by following the arguments of Section 5.2 (see also the appendix of [41]). Note that we shall rename the parameter $t \equiv \sigma^2$ in the following as it makes more sense to see this as a variance instead of a time parameter.

We define two $N \times N$ matrices $\mathbf{M} = \mathbf{C} + \mathbf{B}$ and $\tilde{\mathbf{M}} = \mathbf{C} + \tilde{\mathbf{B}}$ where the noises $\mathbf{B}$ and $\tilde{\mathbf{B}}$ are independent with possibly different variance $\sigma^2$ and $\tilde{\sigma}^2$. We denote by $\mathfrak{g}$ and $\tilde{\mathfrak{g}}$ the Stieltjes transform of $\mathbf{M}$ and $\tilde{\mathbf{M}}$ and we introduce the function,

$$\xi(z) = z - \sigma^2\mathfrak{g}(z), \qquad \tilde{\xi}(\tilde{z}) = \tilde{z} - \tilde{\sigma}^2\tilde{\mathfrak{g}}(\tilde{z})\,. \tag{12.4.1}$$

Note that we shall use the convention $\xi_0(\lambda) = \lim_{\eta\downarrow 0}\xi(\lambda - \mathrm{i}\eta) \equiv \xi_R + \mathrm{i}\xi_I$ and $\tilde{\xi}_0(\lambda) = \lim_{\eta\downarrow 0}\xi(\tilde{\lambda} - \mathrm{i}\eta) \equiv \tilde{\xi}_R + \mathrm{i}\tilde{\xi}_I$. If we use the decomposition $\mathfrak{g}(z) = \mathfrak{g}_R(z) + \mathrm{i}\mathfrak{g}_I(z)$, it is easy to see that $\xi_R = \lambda - \sigma^2\mathfrak{g}_R$ and $\xi_I = -\sigma^2\mathfrak{g}_I$.

Using the result (12.2.16), one has for $N \to \infty$:

$$\left\langle (z-\mathbf{M})^{-1}_{kl}\right\rangle_{\mathcal{P}(\mathbf{M})} \sim \Big(\xi(z) - \mathbf{C}\Big)^{-1}_{kl} \tag{12.4.2}$$

and $\left\langle (\tilde{z}-\tilde{\mathbf{M}})^{-1}_{kl}\right\rangle_{\mathcal{P}(\mathbf{M})}$ is obtained from Eq. (12.4.2) by replacing $\xi$ by $\tilde{\xi}$. Since the noises are independent, it suffices to plug these values into (5.2.2) to obtain, after some algebraic manipulations

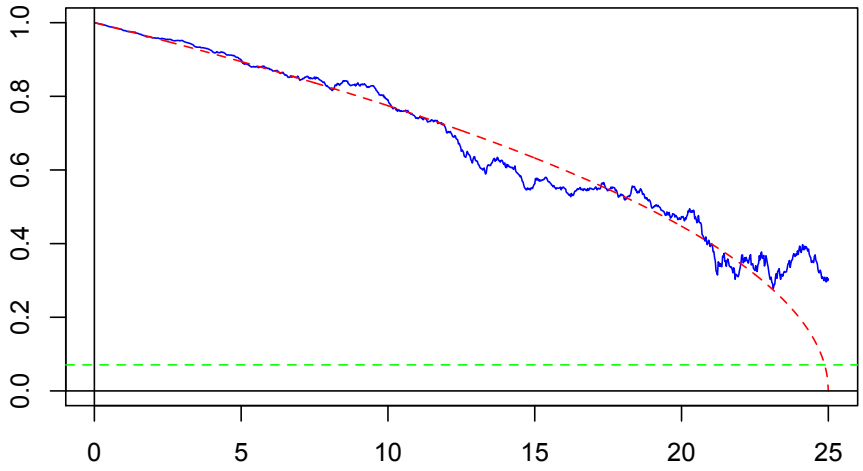



FIGURE 12.3.2. (Color online). Numerical simulation of the process $\langle \mathbf{u}_1^t | \mathbf{u}_1^0 \rangle$ as a function of time $t \in [0, a_1^2]$ (blue line) together with the theoretical limiting curve $\sqrt{1 - t/a_1^2}$ for $t \leqslant t_c = 25$ (red dashed line). The matrix $A$ has only one non zero eigenvalue $a_1 = 5$ and the dimension is $N = 200$. The horizontal green dashed line is $1/\sqrt{N}$. The agreement is good away from the right end point $t_c = a_1^2 = 25$ where a phase transition must occur near the critical point. We see that the convergence holds almost surely as predicted in subsection (12.3.2) from the convergence of the second moment.

similar to those of Section 5.2 (see also [41]), the general result:

$$\Phi_a(\lambda, \tilde{\lambda}) = \frac{(\sigma^2 + \tilde{\sigma}^2)(\xi_R - \tilde{\xi}_R)^2 + 2\sigma^2\tilde{\sigma}^2(\mathfrak{g}_R - \tilde{\mathfrak{g}}_R)(\xi_R - \tilde{\xi}_R) - (\sigma^2 - \tilde{\sigma}^2)(\xi_I^2 - \tilde{\xi}_I^2)}{\left[(\xi_R - \tilde{\xi}_R)^2 + (\xi_I + \tilde{\xi}_I)^2\right]\left[(\xi_R - \tilde{\xi}_R)^2 + (\xi_I - \tilde{\xi}_I)^2\right]}, \quad (12.4.3)$$

where $\Phi_a$ is defined in Eq. (5.0.4) with the subscript $a$ to denote "additive". We see that this latter equation is the analog of Eq. (5.2.15) for the deformed GOE. If we now specialize $\sigma = \tilde{\sigma}$, Eq. (12.4.3) simplifies to:

$$\Phi_a(\lambda, \tilde{\lambda}) = \frac{2\sigma^2(\lambda - \tilde{\lambda})(\xi_R(\lambda) - \xi(\tilde{\lambda}))}{\left[(\xi_R(\lambda) - \xi_R(\tilde{\lambda}))^2 + (\xi_I(\lambda) + \xi_I(\tilde{\lambda}))^2\right]\left[(\xi_R(\lambda) - \xi_R(\tilde{\lambda}))^2 + (\xi_I(\lambda) - \xi_I(\tilde{\lambda}))^2\right]}, \quad (12.4.4)$$

and as for sample covariance matrices, one can again set $\tilde{\lambda} = \lambda + \varepsilon$ to find the self-overlap formula:

$$\Phi_a(\lambda, \lambda) = \frac{\sigma^2 \partial_\lambda \xi_R(\lambda)}{2\xi_I^2\left([\partial_\lambda \xi_R(\lambda)]^2 + [\partial_\lambda \xi_I(\lambda)]^2\right)}. \quad (12.4.5)$$

As a consistency check, let us consider $\tilde{\sigma}^2 = 0$ (no noise). This implies that $\tilde{m}_I = 0$ and $\tilde{m}_R = \mu$. Then, one can easily see from Eq. (12.4.3) that this yields:

$$\Phi_a(\lambda, \mu) = \frac{\sigma^2}{(\xi_R - \mu)^2 + \xi_I^2}, \quad (12.4.6)$$

which is exactly the result derived in (12.2.6).

Finally, an important case is when the noises $\mathbf{B}$, $\tilde{\mathbf{B}}$ are correlated (with a coefficient $\rho$). Contrary to sample covariance matrices, it turns out that we can still find the mean squared overlaps (5.0.4). More importantly, the result is identical to (12.4.3) (up to the variance term) even if the above calculations referred to independent noises. Indeed, the trick is to realize that one can always write (in law) $\mathbf{B} = \sqrt{\rho}\,\mathbf{B}_0 + \sqrt{1-\rho}\,\mathbf{B}_1$ and $\tilde{\mathbf{B}} = \sqrt{\rho}\,\mathbf{W}_0 + \sqrt{1-\rho}\,\mathbf{W}_2$, where $\mathbf{B}_1$, $\mathbf{B}_2$ are now independent, as above. Since the formula (12.4.3) does not rely on the common matrix $\mathbf{C}$, we can replace it by $\mathbf{C}+\sqrt{\rho}\mathbf{W}_0$ and we therefore conclude that (12.4.3) trivially holds with $\sigma^2$ simply multiplied by $1-\rho$. The corresponding shape of $\Phi_a(\lambda,\lambda)$ for different values of $\rho$ is shown in Fig. 12.4.1. We also provide in the inset a comparison with synthetic data for a fixed $\rho = 0.54$, $\sigma^2 = 1$. The empirical average is taken over 200 realizations of the noises and the agreement is excellent.

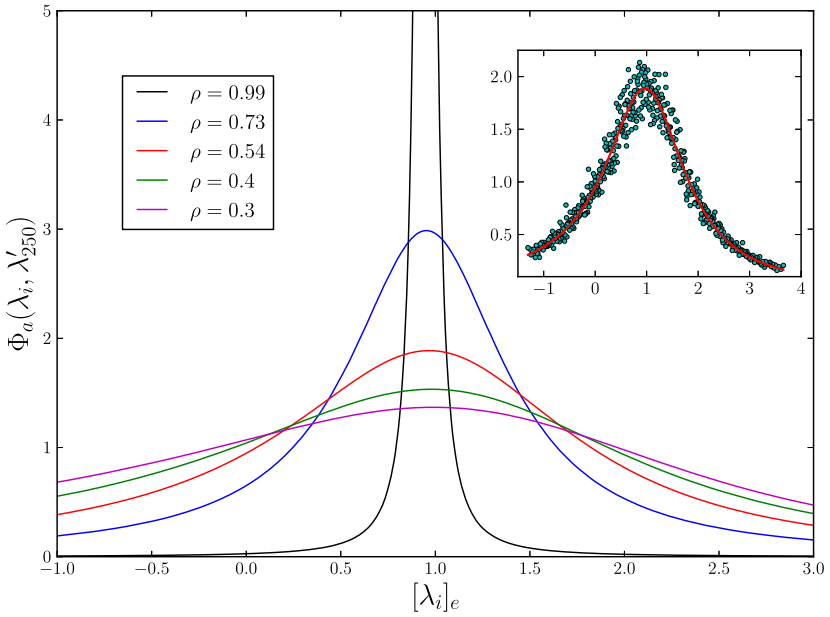



FIGURE 12.4.1. Main Figure: Evaluation of the self overlap $\Phi_a(\lambda,\lambda')$ for a fixed $\lambda' \approx 0.95$ as a function of $\lambda$ for $N = 500$, $\sigma^2 = 1$, and for different values of $\rho$. The population matrix $\mathbf{C}$ is given by a (white) Wishart matrix with parameter $T = 2N$. Inset: We compare the theoretical prediction $\Phi_a(\lambda,\lambda' \approx 0.95)$ for a fixed $\rho = 0.54$ with synthetic data. The empirical averages (blue points) are obtained from 200 independent realizations of $\mathbf{B}$.

In summary, we have provided general and exact formulas for the overlaps between the eigenvectors of large correlated deformed GOEs. As for the multiplicative case, these results do not require the knowledge of the underlying "pure" matrix $\mathbf{C}$ and we believe that this formula could have a broad range of applications in different contexts.

# Chapter 13

# Extension to an arbitrary rotational invariant noise

*This chapter is based on [40].*

We now turn on the general case where the noise term $\mathbf{B}$ is a (asymptotically) rotational invariant random matrix. In the previous chapter, we studied in details the statistics of the eigenvectors of the deformed GOE ensemble with possibly a finite and fixed number of outliers. We saw that the arguments for the outlier eigenvectors were quite different from the bulk ones, as is also the case for sample covariance matrices. In particular, the analysis of outliers relied on the mean squared overlaps dynamics (12.1.4) that is unfortunately unavailable when the noise term is not Gaussian. Hence, it seems quite difficult to find a unified framework that allows one to investigate the outlier eigenvectors for the more general model of free addition (11.1.1). On the other hand, the bulk eigenvectors can be studied in details since it only requires the characterization of the limiting behavior of the resolvent of the observed matrix $\mathbf{M}$. This latter quantity was actually given in Eq. (3.1.103) so that we have all the needed tools to analyze the mean squared overlaps for bulk eigenvectors of $\mathbf{M}$.

In the first part of this chapter, we first go back to the derivation of Eq. (3.1.103) and we will show that the derivation is deeply related to the free addition formula. Thus, we shall propose a formal but elementary derivation of Voiculescu's free addition (3.1.66) by following the arguments of [40]. From this result, we will be able to derive the asymptotic behavior of the resolvent of the model (11.1.1) using the Replica formalim of Section 3.1.4. Finally, we shall apply this resolvent relation in order to derive the mean squared overlaps (5.0.3) for the bulk eigenvectors and also the optimal RIE in the large $N$ limit.

## 13.1 An elementary derivation of the free addition formula

As in Section 3.1.3, the starting point is to notice that since the noise is rotationally invariant, we can always work in the basis where the matrix $\mathbf{C}$ is diagonal. Thus, we may specialize the Replica formalism (3.1.92) for the resolvent of (11.1.1) which yields[1]

$$\mathbf{G}_{\mathbf{M}}(z)_{i,j} = \int \left( \prod_{\alpha=1}^{n} \prod_{k=1}^{N} d\eta_k^\alpha \right) \eta_i^1 \eta_j^1 \prod_{\alpha=1}^{n} e^{-\frac{1}{2}\sum_{k=1}^{N} (\eta_k^\alpha)^2 (z-c_k)} \left\langle e^{-\frac{1}{2}\sum_{k,l=1}^{N} \eta_k^\alpha (\Omega \mathbf{B} \Omega^*)_{k,l} \eta_l^\alpha} \right\rangle_\Omega . \quad (13.1.1)$$

[1] One may also use the Replica formalism for the Stieltjes transform as well.

One recognizes that the average value in the RHS of the latter equation is again the finite rank version of HCIZ integrals studied in details in Section A.2. Hence, one deduces from (A.1.7) that

$$\mathcal{I}_1\left(\sum_{\alpha=1}^{n}\eta^\alpha(\eta^\alpha)^*,\mathbf{B}\right)=\exp\left[\frac{N}{2}\sum_{\alpha=1}^{n}\mathcal{W}_{\mathbf{B}}\left(\frac{1}{N}(\eta^\alpha)^\dagger\eta^\alpha\right)\right], \tag{13.1.2}$$

with $\mathcal{W}'_{\mathbf{B}}(.)=\mathcal{R}_{\mathbf{B}}(.)$ the primitive of the $\mathcal{R}$-transform of $\mathbf{B}$. As a result, the computation of the resolvent (13.1.1) becomes

$$\mathbf{G}_{\mathbf{M}}(z)_{i,j}=\int\left(\prod_{k=1}^{N}d\eta_k\right)\eta_i^1\eta_j^1\exp\left\{\frac{N}{2}\sum_{\alpha=1}^{n}\left[\mathcal{W}_{\mathbf{B}}\left(\frac{1}{N}(\eta^\alpha)^\dagger\eta^\alpha\right)-\frac{1}{2}\sum_{k=1}^{N}(\eta_k^\alpha)^2(z-\mu_k)\right]\right\}, \tag{13.1.3}$$

and by introducing a Lagrange multiplier $p^\alpha:=\frac{1}{N}(\eta^\alpha)^\dagger\eta^\alpha$, we obtain using Fourier transform (and renaming $\zeta^\alpha=-2i\zeta^\alpha/N$)

$$\begin{aligned}\mathbf{G}_{\mathbf{M}}(z)_{i,j}\ \propto\ &\int\int\left(\prod_{\alpha=1}^{n}dp^\alpha d\zeta^\alpha\right)\exp\left\{\frac{N}{2}\sum_{\alpha=1}^{n}\left[\mathcal{W}_{\mathbf{B}}(p^\alpha)-p^\alpha\zeta^\alpha\right]\right\}\\&\times\int\left(\prod_{\alpha=1}^{n}\prod_{k=1}^{N}d\eta_k^\alpha\right)\eta_i^1\eta_j^1\exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{k=1}^{N}(\eta_k^\alpha)^2(z-\zeta^\alpha-\mu_k)\right\}.\end{aligned}$$

One can readily find

$$\mathbf{G}_{\mathbf{M}}(z)_{i,j}\propto\int\int\left(\prod_{\alpha=1}^{n}dp^\alpha d\zeta^\alpha\right)\frac{\delta_{i,j}}{z+\zeta^1-\mu_i}\exp\left\{-\frac{Nn}{2}F_0(p^\alpha,\zeta^\alpha)\right\}, \tag{13.1.4}$$

where the 'free energy' $F_0$ is given by

$$F_0(p^\alpha,\zeta^\alpha)=\frac{1}{Nn}\sum_{\alpha=1}^{n}\left[\sum_{k=1}^{N}\log(z-\zeta^\alpha-\mu_k)-\mathcal{W}_{\mathbf{B}}(p^\alpha)+p^\alpha\zeta^\alpha\right]. \tag{13.1.5}$$

As in Section 3.1.3, the integral (13.1.4) can be evaluated by considering the saddle-point of the free energy $F_0$ as the other term is obviously sub-leading. Moreover, we use the *replica symmetric* ansatz that tells us if the free energy is invariant under the action of the symmetry group $\mathbf{O}(N)$, then we expect a saddle-point which is also invariant. This implies that we have at the saddle-point

$$p^\alpha=p\quad\text{ and }\quad\zeta^\alpha=\zeta,\qquad\forall\,\alpha\in\{1,\dots,n\}, \tag{13.1.6}$$

from which, we obtain the following set of equations:

$$\zeta^*=\mathcal{R}_{\mathbf{B}}(p^*)\qquad\text{and}\qquad p^*=\mathfrak{g}_{\mathbf{C}}(z-\zeta^*). \tag{13.1.7}$$

If we apply the Blue transform of $\mathbf{C}$ on the second equation of (13.1.7), we obtain

$$z=\mathcal{B}_{\mathbf{C}}(p^*)+\mathcal{R}_{\mathbf{B}}(p^*)\equiv\mathcal{R}_{\mathbf{C}}(p^*)+\mathcal{R}_{\mathbf{B}}(p^*)-\frac{1}{p^*}. \tag{13.1.8}$$

On the other hand, we see that the resolvent (13.1.4) is given in the large $N$ limit and the limit $n\to 0$ by

$$\mathbf{G}_{ij}(z)\sim\frac{\delta_{ij}}{z-\mathcal{R}_{\mathbf{B}}(p^*)-\mu_i}. \tag{13.1.9}$$

The trick is to see that we can get rid off one variable by taking the normalized trace in this later equation as it yields

$$\mathfrak{g}_{\mathbf{M}}(z) = \mathfrak{g}_{\mathbf{C}}(z - \mathcal{R}_{\mathbf{B}}(p^*)) = p^* \tag{13.1.10}$$

where the last equation follows from (13.1.7). Therefore, we conclude by plugging this last equation into (13.1.8) that

$$z - \frac{1}{\mathfrak{g}_{\mathbf{M}}(z)} = \mathcal{R}_{\mathbf{C}}(\mathfrak{g}_{\mathbf{M}}(z)) + \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z)),$$

from which one can check by renaming $z = \mathcal{B}_{\mathbf{M}}(\omega)$ that

$$\mathcal{R}_{\mathbf{M}}(\omega) = \mathcal{R}_{\mathbf{C}}(\omega) + \mathcal{R}_{\mathbf{B}}(\omega), \tag{13.1.11}$$

which is exactly the free addition formula (3.1.66).

## 13.2 Asymptotic resolvent of (11.1.1)

A trivial application of the result above is the evaluation of the resolvent entrywise for the general model (11.1.1). Indeed, we see by plugging Eq. (13.1.10) into Eq. (13.1.9) that

$$\mathbf{G}_{\mathbf{M}}(z)_{ij} \sim \frac{\delta_{ij}}{z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z)) - \mu_i}, \tag{13.2.1}$$

which is equivalent to

$$\mathbf{G}_{\mathbf{M}}(z) = \mathbf{G}_{\mathbf{C}}(Z(z)), \qquad Z(z) \;:=\; z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z)). \tag{13.2.2}$$

One notices that this formula is indeed the generalization of the formula (3.1.67) as a matrix. Moreover, we see that in the large $N$ limit, the random resolvent of $\mathbf{M}$ converges to a deterministic quantity that lies in the basis of $\mathbf{C}$. We therefore see that the additive case is even simpler than the multiplicative one as expected. It also means that all the computations we considered in Section 5 can be performed nearly verbatim for the additive model (11.1.1) and the exact results can be found in [40].

## 13.3 Overlap and Optimal RIE formulas in the additive case

**13.3.1. Mean squared overlaps.** We were able to show that the resolvent of $\mathbf{M}$ in the general additive model (11.1.1) converges to a deterministic limit that is given in Eq. (13.2.2). We see that this matrix relation can be simplified when written in the basis where $\mathbf{C}$ is diagonal, since in this case $\mathbf{G}_{\mathbf{C}}(Z)$ is also diagonal. Therefore, the evaluation of the mean squared overlap between a given sample and true eigenvectors, denoted as $\Phi(\lambda, \mu)$, is straightforward using the same techniques as in Section 5.1.1. We omit details that may be found in [40] and one finds that the overlap for the free additive noise is given by:

$$\Phi(\lambda, \mu) = \frac{\beta_1(\lambda)}{(\lambda - \mu - \alpha_a(\lambda))^2 + \pi^2 \beta_a(\lambda)^2 \rho_{\mathbf{M}}(\lambda)^2}, \tag{13.3.1}$$

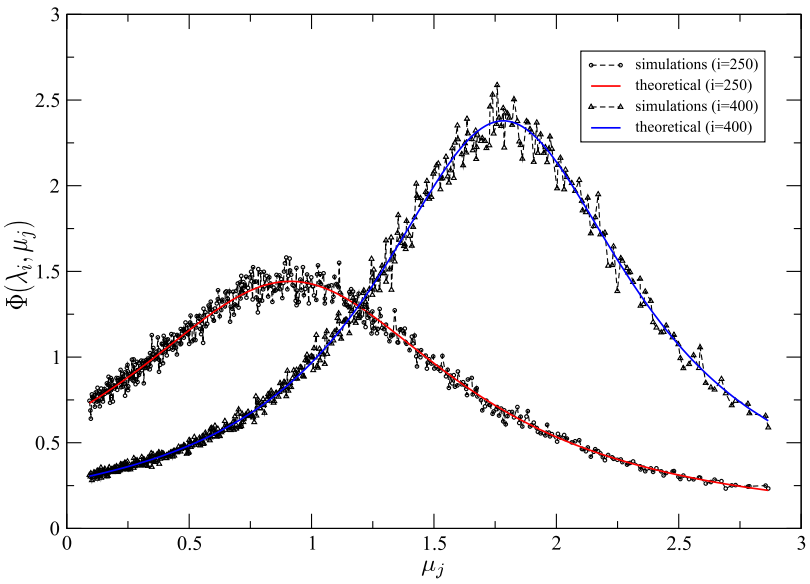



FIGURE 13.3.1. Computations of the rescaled overlap $\Phi(\lambda, \mu)$ as a function of $\mu$ in the free addition perturbation. We chose $i = 250$, $\mathbf{C}$ a Wishart matrix with parameter $q = 0.5$ and $\mathbf{B}$ a Wigner matrix with $\sigma^2 = 1$. The black dotted points are computed using numerical simulations and the plain red curve is the theoretical predictions Eq. (13.3.1). The agreement is excellent. For $i = 250$, we have $\mu_i \approx 0.83$ and we see that the peak of the curve is in that region. The same observation holds for $i = 400$ where $\mu_i \approx 1.66$. The numerical curves display the empirical mean values of the overlaps over 1000 samples of $\mathbf{M}$ given by Eq. (11.1.1) with $\mathbf{C}$ fixed.

where $\mu$ is the corresponding eigenvalue of the true matrix $\mathbf{C}$, and where we defined:

$$\begin{cases} \alpha_a(\lambda) &:= \mathrm{Re}[\mathcal{R}_{\mathbf{B}}\left(\mathfrak{h}_{\mathbf{M}}(\lambda) + \mathrm{i}\pi\rho_{\mathbf{M}}(\lambda)\right)], \\ \beta_a(\lambda) &:= \dfrac{\mathrm{Im}[\mathcal{R}_{\mathbf{B}}\left(\mathfrak{h}_{\mathbf{M}}(\lambda) + \mathrm{i}\pi\rho_{\mathbf{M}}(\lambda)\right)]}{\pi\rho_{\mathbf{M}}(\lambda)}. \end{cases} \tag{13.3.2}$$

As a simple consistency check, we specialize our result to the case where $\mathbf{\Omega B \Omega}^*$ is a GOE matrix such that the entries have a variance equal to $\sigma^2/N$. Then, one has $\mathcal{R}_{\mathbf{B}}(z) = \sigma^2 z$ meaning that $Z(z)$ of Eq. (13.2.2) simply becomes $Z(z) = z - \sigma^2 \mathfrak{g}_{\mathbf{M}}(z)$. This allows us to get a simpler expression for the overlap:

$$\Phi(\lambda, \mu) = \frac{\sigma^2}{(c - \lambda + \sigma^2 \mathfrak{h}_{\mathbf{M}}(\lambda))^2 + \sigma^4 \pi^2 \rho_{\mathbf{M}}(\lambda)^2}, \tag{13.3.3}$$

which is exactly the result obtained in Eq. (12.2.6). In Fig. 13.3.1, we illustrate this formula in the case where $\mathbf{C} = \boldsymbol{\mathcal{W}}$ with parameter $q$. We set $N = 500$, $T = 1000$, and take $\mathbf{\Omega B \Omega}^*$ as a GOE matrix with variance $1/N$. For a fixed $\mathbf{C}$, we generate 200 samples of $\mathbf{M}$ given by Eq. (11.1.1) for which we can measure numerically the overlap (5.0.3). We see that the theoretical prediction (13.3.3) agrees remarkably with the numerical simulations.

**13.3.2. Optimal RIE.** Since the overlaps are explicit in this general model, it is easy to compute the asymptotic limit of the oracle estimator (7.1.2) for the bulk eigenvalues in the model (11.1.1).

Indeed, it is easy to see from Eqs. (3.1.6) and (7.1.2) that:

$$\xi_i^{\text{ora.}} \sim \frac{1}{\pi\rho_{\mathbf{M}}(\lambda_i)} \lim_{z\to\lambda_i - i0^+} \text{Im}\left[\int \frac{\mu\,\rho_{\mathbf{C}}(\mu)}{Z(z)-\mu}\,\text{d}\mu\right] = \frac{1}{N\pi\rho_{\mathbf{M}}(\lambda_i)} \lim_{z\to\lambda_i - i0^+} \text{Im}\,\text{Tr}\left[\mathbf{G}_{\mathbf{M}}(z)\mathbf{C}\right], \quad (13.3.4)$$

where $Z(z)$ is given by Eq. (13.2.2). From Eq. (13.2.2) one also has $\text{Tr}[\mathbf{G}_{\mathbf{M}}(z)\mathbf{C}] = N(Z(z)\mathfrak{g}_{\mathbf{M}}(z)-1)$, and using Eqs. (13.2.2) and (13.3.2), we end up with:

$$\lim_{z\to\lambda - i0^+} \text{Im}\,\text{Tr}\left[\mathbf{G}_{\mathbf{M}}(z)\mathbf{C}\right] = N\pi\rho_M(\lambda)\left[\lambda - \alpha(\lambda) - \beta(\lambda)\mathfrak{h}_{\mathbf{M}}(\lambda)\right].$$

We therefore find the following optimal RIE nonlinear "shrinkage" function $F_1$:

$$\xi_i^{\text{ora.}} \sim F_a(\lambda_i); \qquad F_a(\lambda) = \lambda - \alpha_a(\lambda) - \beta_a(\lambda)\mathfrak{h}_{\mathbf{M}}(\lambda), \quad (13.3.5)$$

where $\alpha_a, \beta_a$ are defined in Eq. (13.3.2). This result states that if we consider a model where the signal $\mathbf{C}$ is perturbed with an additive noise (that is free with respect to $\mathbf{C}$), the optimal way to 'clean' the eigenvalues of $\mathbf{M}$ in order to get $\widehat{\Xi}(\mathbf{M})$ is to keep the eigenvectors of $\mathbf{M}$ and apply the nonlinear shrinkage formula (13.3.5). We see that the non-observable oracle estimator converges in the limit $N \to \infty$ towards a deterministic function of the observable eigenvalues.

As usual, let us consider the case where $\mathbf{B}$ is a GOE matrix in order to give more intuitions about (13.3.5). Using the definition of $\alpha_a$ and $\beta_a$ given in Eq. (13.3.2), the nonlinear shrinkage function is given by

$$F_a(\lambda) = \lambda - 2\sigma^2\mathfrak{h}_{\mathbf{M}}(\lambda). \quad (13.3.6)$$

Moreover, suppose that $\mathbf{C}$ is also a GOE matrix so that $\mathbf{M}$ is a also a GOE matrix with variance $\sigma_{\mathbf{M}}^2 = \sigma_{\mathbf{C}}^2 + \sigma^2$. As a consequence, the Hilbert transform of $\mathbf{M}$ can be computed straightforwardly from the Wigner semicircle law and we find

$$\mathfrak{h}_{\mathbf{M}}(\lambda) = \frac{\lambda}{2\sigma_{\mathbf{M}}^2}.$$

The optimal cleaning scheme to apply in this case is then given by:

$$F_a(\lambda) = \lambda\left(\frac{\sigma_{\mathbf{C}}^2}{\sigma_{\mathbf{C}}^2+\sigma^2}\right), \quad (13.3.7)$$

where one can see that the optimal cleaning is given by rescaling the empirical eigenvalues by the signal-to-noise ratio. This result is expected in the sense that we perturb a Gaussian signal by adding a Gaussian noise. We know in this case that the optimal estimator of the signal is given, element by element, by the Wiener filter [188], and this is exactly the result that we have obtained with (13.3.7). We can also notice that the ESD of the cleaned matrix is narrower than the true one. Indeed, let us define the signal-to-noise ratio $\text{SNR} = \sigma_{\mathbf{C}}^2/\sigma_{\mathbf{M}}^2 \in [0,1]$, and it is obvious from (13.3.7) that $\widehat{\Xi}(\mathbf{M})$ is a Wigner matrix with variance $\sigma_{\Xi}^2 \times \text{SNR}$ which leads to

$$\sigma_{\mathbf{M}}^2 \geq \sigma_{\mathbf{C}}^2 \geq \sigma_{\mathbf{C}}^2 \times \text{SNR}, \quad (13.3.8)$$

as it should be.

As a second example, we now consider a less trivial case and suppose that $\mathbf{C}$ is a white Wishart matrix with parameter $q_0$. For any $q_0 > 0$, it is well known that the Wishart matrix has nonnegative eigenvalues. However, we expect that the noisy effect coming from the GOE

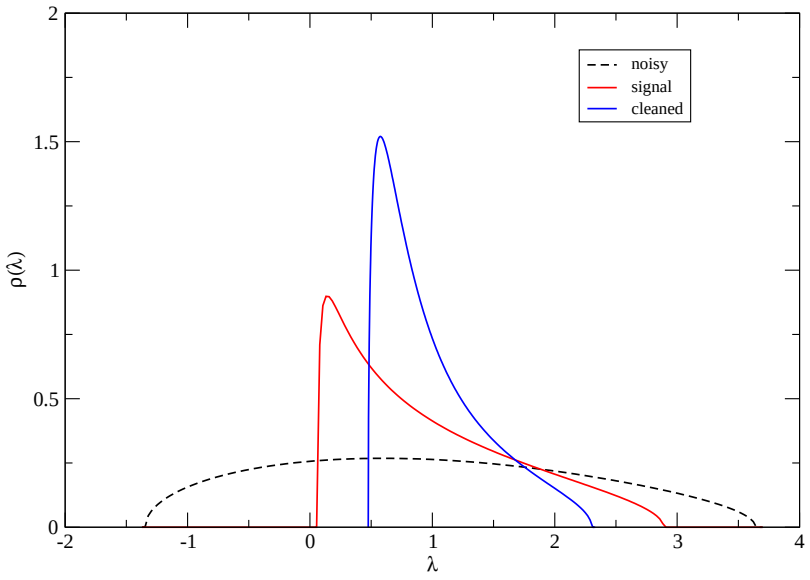



FIGURE 13.3.2. Eigenvalues of the noisy measurement $\mathbf{M}$ (black dotted line) compared to the true signal $\mathbf{C}$ drawn from a $500 \times 500$ Wishart matrix of parameter $q_0 = 0.5$ (red line). We have corrupted the signal by adding a GOE matrix with radius 1. The eigenvalues density of $\mathbf{M}$ allows negative values while the true one has only positive values. The blue line is the LSD of the optimally cleaned matrix. We clearly notice that the cleaned eigenvalues are all positive and its spectrum is narrower than the true one, while preserving the trace.

matrix pushes some true eigenvalues towards the negative side of the real axis. In Fig. 13.3.2, we clearly observe this effect and a good cleaning scheme should bring these negative eigenvalues back to positive values. In order to use Eq. (13.3.6), we invoke once again the free addition formula to find the following equation for the Stieltjes transform of $\mathbf{M}$:

$$-q_0\sigma^2\mathfrak{g}_{\mathbf{M}}(z)^3 + (\sigma^2 + q_0\, z)\mathfrak{g}_{\mathbf{M}}(z)^2 + (1 - q_0 - z)\mathfrak{g}_{\mathbf{M}}(z) + 1 = 0,$$

for any $z = \lambda - \mathrm{i}\eta$ with $\eta \to 0$. It then suffices to take the real part of the Stieltjes transform $\mathfrak{g}_{\mathbf{M}}(z)$ that solves this equation[2] to get the Hilbert transform. In order to check formula Eq. (13.3.5) using numerical simulations, we have generated a matrix of $\mathbf{M}$ given by Eq. (11.1.1) with $\mathbf{C}$ a fixed white Wishart matrix with parameter $q_0$ and $\mathbf{\Omega B \Omega}^*$ a GOE matrix with radius 1. As we know exactly $\mathbf{C}$, we can compute numerically the oracle estimator as given in (7.1.2) for each sample. In Fig. 13.3.3, we see that our theoretical prediction in the large $N$ limit compares very nicely with the mean values of the empirical oracle estimator computed from the sample. We can also notice in Fig. 13.3.2 that the spectrum of the cleaned matrix (represented by the ESD in green) is narrower than the standard Marčenko-Pastur density. This confirms the observation made in Chapter 7.

[2] We take the solution which has a strictly nonnegative imaginary part

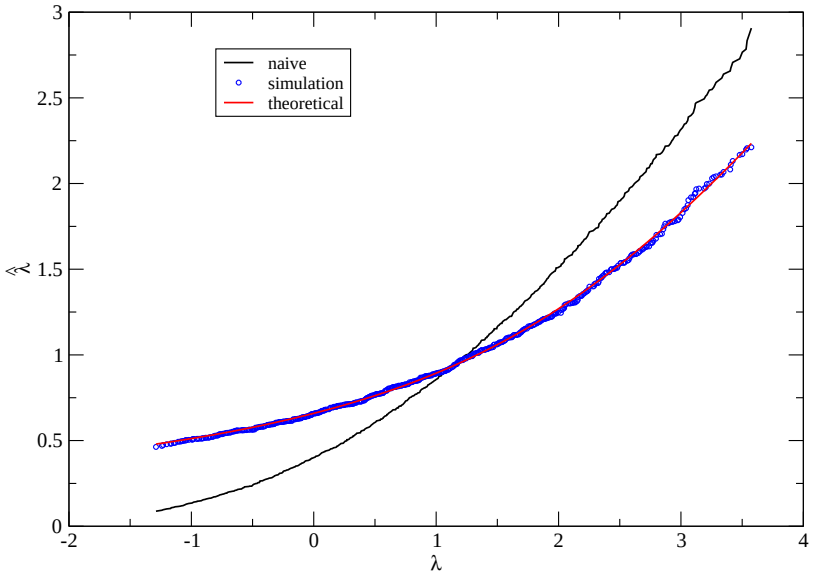



FIGURE 13.3.3. Eigenvalues according to the optimal cleaning formula (13.3.7) (red line) as a function of the observed noisy eigenvalues $\lambda$. The parameter are the same as in Fig. 13.3.2. We also provide a comparison against the naive eigenvalues substitution method (black line) and we see that the optimal cleaning scheme indeed narrows the spacing between eigenvalues.

# Bibliography


[1] Gernot Akemann, Jinho Baik, and Philippe Di Francesco, *The Oxford handbook of random matrix theory*, Oxford University Press, 2011.

[2] Romain Allez and Jean-Philippe Bouchaud, *Eigenvector dynamics: general theory and some applications*, Physical Review E **86** (2012), no. 4, 046202.

[3] ———, *Eigenvector dynamics under free addition*, Random Matrices: Theory and Applications **03** (2014), no. 03, 1450010.

[4] Romain Allez, Jean-Philippe Bouchaud, Satya N Majumdar, and Pierpaolo Vivo, *Invariant $\beta$-Wishart ensembles, crossover densities and asymptotic corrections to the Marčenko–Pastur law*, Journal of Physics A: Mathematical and Theoretical **46** (2012), no. 1, 015001.

[5] Romain Allez, Joël Bun, and Jean-Philippe Bouchaud, *The eigenvectors of Gaussian matrices with an external source*, arXiv preprint arXiv:1412.7108 (2014).

[6] Orly Alter, Patrick O Brown, and David Botstein, *Singular value decomposition for genome-wide expression data processing and modeling*, Proceedings of the National Academy of Sciences **97** (2000), no. 18, 10101–10106.

[7] Takeshi Amemiya, *Advanced econometrics*, Harvard university press, 1985.

[8] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni, *An introduction to random matrices*, no. 118, Cambridge University Press, 2010.

[9] Theodore W Anderson, *An introduction to multivariate statistics*, Wiley, New York (1984), 675.

[10] Theodore Wilbur Anderson, *Asymptotic theory for principal component analysis*, Annals of Mathematical Statistics (1963), 122–148.

[11] ZD Bai, Baiqi Miao, and Jian-Feng Yao, *Convergence rates of spectral distributions of large sample covariance matrices*, SIAM journal on matrix analysis and applications **25** (2003), no. 1, 105–127.

[12] ZD Bai, BQ Miao, GM Pan, et al., *On asymptotics of eigenvectors of large sample covariance matrix*, The Annals of Probability **35** (2007), no. 4, 1532–1572.

[13] Zhi Dong Bai, *Convergence rate of expected spectral distributions of large random matrices. part i. Wigner matrices*, The Annals of Probability (1993), 625–648.

[14] Zhidong Bai and Jack W Silverstein, *Spectral analysis of large dimensional random matrices*, Springer, 2009.

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

[16] Daniel Bartz, *Advances in high-dimensional covariance matrix estimation*, Technische Universität Berlin, Doctoral Thesis (2015).

[17] Daniel Bartz, Kerr Hatrick, Christian W Hesse, Klaus-Robert Müller, and Steven Lemm, *Directional variance adjustment: Bias reduction in covariance matrices based on factor analysis with an application to portfolio optimization*, PloS one **8** (2013), no. 7, e67503.

[18] Daniel Bartz and Klaus-Robert Müller, *Covariance shrinkage for autocorrelated data*, Advances in Neural Information Processing Systems, 2014, pp. 1592–1600.

[19] Carlo WJ Beenakker, *Random-matrix theory of quantum transport*, Reviews of modern physics **69** (1997), no. 3, 731.

[20] Gérard Ben Arous and Alice Guionnet, *The spectrum of heavy tailed random matrices*, Communications in Mathematical Physics **278** (2008), no. 3, 715–751.

[21] Florent Benaych-Georges, *Eigenvectors of Wigner matrices: universality of global fluctuations*, arXiv preprint arXiv:1104.1219 (2011).

[22] Florent Benaych-Georges and Antti Knowles, *Lectures on the local semicircle law for Wigner matrices*, arXiv preprint arXiv:1601.04055 (2016).

[23] Florent Benaych-Georges and Raj Rao Nadakuditi, *The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices*, Advances in Mathematics **227** (2011), no. 1, 494–521.

[24] ———, *The singular values and vectors of low rank perturbations of large rectangular random matrices*, Journal of Multivariate Analysis **111** (2012), 120–135.

[25] Giulio Biroli, J-P Bouchaud, and Marc Potters, *On the top eigenvalue of heavy-tailed random matrices*, EPL (Europhysics Letters) **78** (2007), no. 1, 10001.

[26] Giulio Biroli, Jean-Philippe Bouchaud, and Marc Potters, *The student ensemble of correlation matrices: eigenvalue spectrum and kullback-leibler entropy*, arXiv preprint arXiv:0710.0802 (2007).

[27] Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, and Jun Yin, *On the principal components of sample covariance matrices*, arXiv preprint arXiv:1404.0788 (2014).

[28] J-P Bouchaud, Laurent Laloux, M Augusta Miceli, and Marc Potters, *Large dimension forecasting models and random singular value spectra*, The European Physical Journal B **55** (2007), no. 2, 201–207.

[29] Jean-Philippe Bouchaud and M Potters, *Financial applications of random matrix theory: a short review*, The Oxford handbook of Random Matrix Theory, Oxford University Press, 2011.

[30] Jean-Philippe Bouchaud and Marc Potters, *Theory of financial risk and derivative pricing: from statistical physics to risk management*, Cambridge university press, 2003.

[31] Paul Bourgade, L Erdős, H-T Yau, and Jun Yin, *Fixed energy universality for generalized Wigner matrices*, Communications on Pure and Applied Mathematics (2015).

[32] Paul Bourgade and Horng-Tzer Yau, *The eigenvector moment flow and local quantum unique ergodicty*, arXiv preprint arXiv:1312.1301 (2013).

[33] Mark J Bowick and Édouard Brézin, *Universal scaling of the tail of the density of eigenvalues in random matrix models*, Physics Letters B **268** (1991), no. 1, 21–28.

[34] Alan Bray and Alan McKane, *Instanton calculation of the escape rate for activation over a potential barrier driven by colored noise*, Physical Review Letters **62** (1989), 493.

[35] E Brézin, S Hikami, and A Zee, *Universal correlations for deterministic plus random hamiltonians*, Physical Review E **51** (1995), no. 6, 5442.

[36] Edouard Brézin, Claude Itzykson, Giorgio Parisi, and Jean-Bernard Zuber, *Planar diagrams*, Communications in Mathematical Physics **59** (1978), no. 1, 35–51.

[37] Marie-France Bru, *Wishart processes*, Journal of Theoretical Probability **4** (1991), no. 4, 725–751.

[38] J. Bun, J. P. Bouchaud, S. N. Majumdar, and M. Potters, *Instanton approach to large n Harish-Chandra-Itzykson-Zuber integrals*, Phys. Rev. Lett. **113** (2014), 070201.

[39] Joël Bun, *Out-of-sample quadratic optimization using random matrix theory*, Université Paris-Diderot, Master Thesis (2013).

[40] Joël Bun, Romain Allez, Jean-Philippe Bouchaud, and Marc Potters, *Rotational invariant estimator for general noisy matrices*, arXiv preprint arXiv:1502.06736 (2015).

[41] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters, *On the overlaps between eigenvectors of correlated random matrices*, arXiv preprint arXiv:1603.04364 (2016).

[42] ———, *Cleaning correlation matrices*, Risk magazine (2016 (April)).

[43] Joël Bun and Antti Knowles, *An optimal rotational invariant estimator for general covariance matrices*, in preparation (2016).

[44] Z Burda, A Görlich, A Jarosz, and J Jurkiewicz, *Signal and noise in correlation matrix*, Physica A: Statistical Mechanics and its Applications **343** (2004), 295–310.

[45] Z Burda, RA Janik, and MA Nowak, *Multiplication law and S transform for non-hermitian random matrices*, Physical Review E **84** (2011), no. 6, 061125.

[46] Zdzislaw Burda, *Free products of large random matrices–a short review of recent developments*, Journal of Physics: Conference Series, vol. 473, IOP Publishing, 2013, p. 012002.

[47] Zdzisław Burda, Jerzy Jurkiewicz, Maciej A Nowak, Gabor Papp, and Ismail Zahed, *Free Lévy matrices and financial correlations*, Physica A: Statistical Mechanics and its Applications **343** (2004), 694–700.

[48] Zdzislaw Burda, Jerzy Jurkiewicz, and Bartlomiej Waclaw, *Spectral moments of correlated Wishart matrices*, arXiv preprint cond-mat/0405263 (2004).

[49] Fabio Caccioli, Imre Kondor, and Gábor Papp, *Portfolio optimization under expected shortfall: contour maps of estimation error*, arXiv preprint arXiv:1510.04943 (2015).

[50] Gary Chamberlain and Michael Rothschild, *Arbitrage, factor structure, and mean-variance analysis on large asset markets*, 1982.

[51] Rémy Chicheportiche, *Non-linear dependences in finance*, arXiv preprint arXiv:1309.5073 (2013).

[52] Rémy Chicheportiche and J-P Bouchaud, *A nested factor model for non-linear dependencies in stock returns*, Quantitative Finance (2015), no. ahead-of-print, 1–16.

[53] Stefano Ciliberti, Imre Kondor, and Marc Mézard, *On the feasibility of portfolio optimization under expected shortfall*, Quantitative Finance **7** (2007), no. 4, 389–396.

[54] Pierre Cizeau and Jean-Philippe Bouchaud, *Theory of Lévy matrices*, Physical Review E **50** (1994), no. 3, 1810.

[55] Benoît Collins, Alice Guionnet, and Edouard Maurel-Segala, *Asymptotics of unitary and orthogonal matrix integrals*, Advances in Mathematics **222** (2009), no. 1, 172–215.

[56] Benoît Collins, David McDonald, and Nadia Saad, *Compound Wishart matrices and noisy covariance matrices: Risk underestimation*, arXiv preprint arXiv:1306.5510 (2013).

[57] Romain Couillet, Merouane Debbah, et al., *Random matrix methods for wireless communications*, Cambridge University Press Cambridge, MA, 2011.

[58] Romain Couillet, Abla Kammoun, and Frédéric Pascal, *Second order statistics of robust estimators of scatter. application to GLRT detection for elliptical signals*, Journal of Multivariate Analysis **143** (2016), 249–274.

[59] Romain Couillet, Frédéric Pascal, and Jack W Silverstein, *The random matrix regime of Maronna's M-estimator with elliptically distributed samples*, Journal of Multivariate Analysis **139** (2015), 56–78.

[60] C. de Dominicis, *Techniques de renormalisation de la théorie des champs et dynamique des phénomènes critiques*, J. Phys. Colloques **37** (1976), C1–247–C1–253.

[61] David S Dean, *Langevin equation for the density of a system of interacting Langevin processes*, Journal of Physics A: Mathematical and General **29** (1996), no. 24, L613.

[62] David S Dean and Satya N Majumdar, *Large deviations of extreme eigenvalues of random matrices*, Physical review letters **97** (2006), no. 16, 160201.

[63] ———, *Extreme value statistics of eigenvalues of Gaussian random matrices*, Physical Review E **77** (2008), no. 4, 041108.

[64] Josh M. Deutsch, *Quantum statistical mechanics in a closed system*, Physical Review A **43** (1991), no. 4, 2046.

[65] Persi Diaconis and Peter J Forrester, *A. hurwitz and the origins of random matrix theory in mathematics*, arXiv preprint arXiv:1512.09229 (2015).

[66] Lee H Dicker et al., *Ridge regression and asymptotic minimax estimation over spheres of growing dimension*, Bernoulli **22** (2016), no. 1, 1–37.

[67] Ivailo I Dimov, Petter N Kolm, Lee Maclin, and Dan YC Shiber, *Hidden noise structure and random matrix models of stock correlations*, Quantitative Finance **12** (2012), no. 4, 567–572.

[68] Catherine Donati-Martin and Alain Rouault, *Random truncations of haar distributed matrices and bridges*, arXiv preprint arXiv:1302.6539 (2013).

[69] Paul S Dwyer, *Some applications of matrix derivatives in multivariate analysis*, Journal of the American Statistical Association **62** (1967), no. 318, 607–625.

[70] Freeman J Dyson, *A Brownian-motion model for the eigenvalues of a random matrix*, Journal of Mathematical Physics **3** (1962), 1191–1198.

[71] Bradley Efron and Carl Morris, *Multivariate empirical Bayes and estimation of covariance matrices*, The Annals of Statistics (1976), 22–32.

[72] Bradley Efron and Carl N Morris, *Stein's paradox in statistics*, WH Freeman, 1977.

[73] J Eisert, M Friesdorf, and Christian Gogolin, *Quantum many-body systems out of equilibrium*, Nature Physics **11** (2015), no. 2, 124–130.

[74] Noureddine El Karoui et al., *Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond*, The Annals of Applied Probability **19** (2009), no. 6, 2362–2405.

[75] László Erdős, *Universality of Wigner random matrices: a survey of recent results*, Russian Mathematical Surveys **66** (2011), no. 3, 507.

[76] László Erdős, Sandrine Péché, José A Ramírez, Benjamin Schlein, and Horng-Tzer Yau, *Bulk universality for Wigner matrices*, Communications on Pure and Applied Mathematics **63** (2010), no. 7, 895–925.

[77] Eugene F Fama and Kenneth R French, *Common risk factors in the returns on stocks and bonds*, Journal of financial economics **33** (1993), no. 1, 3–56.

[78] PJ Forrester, *Some exact correlations in the Dyson Brownian motion model for transitions to the CUE*, Physica A: Statistical Mechanics and its Applications **223** (1996), no. 3, 365–390.

[79] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics Springer, Berlin, 2001.

[80] George W Furnas, Scott Deerwester, Susan T Dumais, Thomas K Landauer, Richard A Harshman, Lynn A Streeter, and Karen E Lochbaum, *Information retrieval using a singular value decomposition model of latent semantic structure*, Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1988, pp. 465–480.

[81] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin, *Bayesian data analysis*, vol. 2, Taylor & Francis, 2014.

[82] A. Guionnet and M. Maïda, *A Fourier view on the r-transform and related asymptotics of spherical integrals*, Journal of Functional Analysis **222** (2005), no. 2, 435 – 490.

[83] Alice Guionnet and Ofer Zeitouni, *Large deviations asymptotics for spherical integrals*, Journal of functional analysis **188** (2002), no. 2, 461–515.

[84] Walid Hachem, Adrien Hardy, and Jamal Najim, *A survey on the eigenvalues local behavior of large complex correlated Wishart matrices*, ESAIM: Proceedings and Surveys **51** (2015), 150–174.

[85] LR Haff, *Minimax estimators for a multinormal precision matrix*, Journal of Multivariate Analysis **7** (1977), no. 3, 374–385.

[86] ———, *An identity for the Wishart distribution with applications*, Journal of Multivariate Analysis **9** (1979), no. 4, 531–544.

[87] ———, *Empirical Bayes estimation of the multivariate normal covariance matrix*, The Annals of Statistics (1980), 586–597.

[88] Lars Peter Hansen, *Large sample properties of generalized method of moments estimators*, Econometrica: Journal of the Econometric Society (1982), 1029–1054.

[89] Harish-Chandra, *Differential operators on a semisimple lie algebra*, American Journal of Mathematics (1957), 87–120.

[90] Gerard Hooft, *A planar diagram theory for strong interactions*, Nuclear Physics B **72** (1974), no. 3, 461–473.

[91] Harold Hotelling, *Relations between two sets of variates*, Biometrika **28** (1936), no. 3/4, 321–377.

[92] David Hoyle and Magnus Rattray, *Limiting form of the sample covariance eigenspectrum in pca and kernel pca*, Advances in Neural Information Processing Systems, 2003, p. None.

[93] Peter J Huber, *Robust statistics*, Springer, 2011.

[94] Gregoire Ithier and Florent Benaych-Georges, *Thermalisation of a quantum system from first principles*, arXiv preprint arXiv:1510.04352 (2015).

[95] C Itzykson and J-B Zuber, *The planar approximation. ii*, Journal of Mathematical Physics **21** (1980), 411–421.

[96] William James and Charles Stein, *Estimation with quadratic loss*, Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, vol. 1, 1961, pp. 361–379.

[97] Hans-Karl Janssen, *On a Lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties*, Zeitschrift für Physik B Condensed Matter **23** (1976), no. 4, 377–380.

[98] Kurt Johansson, *Shape fluctuations and random matrices*, Communications in mathematical physics **209** (2000), no. 2, 437–476.

[99] _____, *Universality of the local spacing distribution in certain ensembles of hermitian Wigner matrices*, Communication in Mathematical Physics **215** (2001), no. 3, 683–705.

[100] Iain M Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Annals of statistics (2001), 295–327.

[101] _____, *Multivariate analysis and jacobi ensembles: Largest eigenvalue, tracy–widom limits and rates of convergence*, Annals of statistics **36** (2008), no. 6, 2638.

[102] George Kapetanios, *A new method for determining the number of factors in factor models with large datasets*, Tech. report, Working Paper, Department of Economics, Queen Mary, University of London, 2004.

[103] V Kargin et al., *Subordination for the sum of two random matrices*, The Annals of Probability **43** (2015), no. 4, 2119–2150.

[104] Noureddine El Karoui, *Spectrum estimation for large dimensional covariance matrices using random matrix theory*, The Annals of Statistics (2008), 2757–2790.

[105] Noureddine El Karoui and Holger Kösters, *Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods*, arXiv preprint arXiv:1105.1404 (2011).

[106] K Kawasaki, *Simple derivations of generalized linear and nonlinear Langevin equations*, Journal of Physics A: Mathematical and General **6** (1973), 1289.

[107] Alexei M Khorunzhy and L Pastur, *On the eigenvalue distribution of the deformed Wigner ensemble of random matrices*, Advances in Soviet Mathematics **19** (1994), 97–127.

[108] Virginia C Klema and Alan J Laub, *The singular value decomposition: Its computation and some applications*, Automatic Control, IEEE Transactions on **25** (1980), no. 2, 164–176.

[109] Antti Knowles and Jun Yin, *Anisotropic local laws for random matrices*, arXiv preprint arXiv:1410.3516 (2014).

[110] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters, *Noise dressing of financial correlation matrices*, Physical review letters **83** (1999), no. 7, 1467.

[111] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud, *Random matrix theory and financial correlations*, International Journal of Theoretical and Applied Finance **3** (2000), no. 03, 391–397.

[112] Michel Le Bellac, Fabrice Mortessagne, and G. George Batrouni, *Equilibrium and non-equilibrium statistical thermodynamics*, Cambridge University Press, 2006.

[113] Olivier Ledoit and Sandrine Péché, *Eigenvectors of some large sample covariance matrix ensembles*, Probability Theory and Related Fields **151** (2011), no. 1-2, 233–264.

[114] Olivier Ledoit and Michael Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection*, Journal of Empirical Finance **10** (2003), no. 5, 603–621.

[115] ———, *A well-conditioned estimator for large-dimensional covariance matrices*, Journal of multivariate analysis **88** (2004), no. 2, 365–411.

[116] ———, *Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions*, Tech. report, Working Paper Series, Department of Economics, University of Zurich, 2013.

[117] ———, *Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks*, Available at SSRN 2383361 (2014).

[118] ———, *Numerical implementation of the quest function*, Tech. report, Department of Economics-University of Zurich, 2016.

[119] AMS Macêdo, *Universal parametric correlations at the soft edge of the spectrum of random matrix ensembles*, EPL (Europhysics Letters) **26** (1994), no. 9, 641.

[120] Satya N Majumdar and Grégory Schehr, *Top eigenvalue of a random matrix: large deviations and third order phase transition*, Journal of Statistical Mechanics: Theory and Experiment **2014** (2014), no. 1, P01012.

[121] Satya N Majumdar and Pierpaolo Vivo, *Number of relevant directions in principal component analysis and Wishart random matrices*, Physical review letters **108** (2012), no. 20, 200601.

[122] SN Majumdar, *Random matrices, the Ulam problem, directed polymers and growth models, and sequence matching, chapter 4*, Les Houches-Session LXXXV. Elsevier (2006), 179–216.

[123] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur, *Distribution of eigenvalues for some sets of random matrices*, Matematicheskii Sbornik **114** (1967), no. 4, 507–536.

[124] Enzo Marinari, Giorgio Parisi, and Felix Ritort, *Replica field theory for deterministic models. ii. a non-random spin glass with glassy behaviour*, Journal of Physics A: Mathematical and General **27** (1994), no. 23, 7647.

[125] Harry Markowitz, *Portfolio selection**, The journal of finance **7** (1952), no. 1, 77–91.

[126] Harry M Markowitz, *Portfolio selection: efficient diversification of investments*, vol. 16, Yale University Press, 1968.

[127] Ricardo A Maronna, Douglas R Martin, and Victor J Yohai, *Robust statistics: Theory and methods*, John Wiley and Sons, 2006.

[128] Matteo Marsili, *Dissecting financial markets: sectors and states*, Quantitative Finance **2** (2002), no. 4, 297–302.

[129] A Matytsin, *On the large-n limit of the Itzykson-Zuber integral*, Nuclear Physics B **411** (1994), 805–820.

[130] Peter McCullagh and John A Nelder, *Generalized linear models*, vol. 37, CRC press, 1989.

[131] Madan Lal Mehta, *Random matrices*, vol. 142, Academic press, 2004.

[132] Robert C Merton, *An intertemporal capital asset pricing model*, Econometrica: Journal of the Econometric Society (1973), 867–887.

[133] Xavier Mestre, *Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates*, Information Theory, IEEE Transactions on **54** (2008), no. 11, 5113–5129.

[134] Marc Mézard, Miguel Angel Virasoro, and Giorgio Parisi, *Spin glass theory and beyond*, World scientific, 1987.

[135] James A Mingo and Alexandru Nica, *Annular noncrossing permutations and partitions, and second-order asymptotics for random matrices*, International Mathematics Research Notices **2004** (2004), no. 28, 1413–1460.

[136] Rémi Monasson and Dario Villamaina, *Estimating the principal components of correlation matrices from all their empirical eigenvectors*, EPL (Europhysics Letters) **112** (2015), no. 5, 50001.

[137] Flaviano Morone, Francesco Caltagirone, Elizabeth Harrison, and Giorgio Parisi, *Replica theory and spin glasses*, arXiv preprint arXiv:1409.2722 (2014).

[138] Celine Nadal and Satya N Majumdar, *A simple derivation of the tracy–widom distribution of the maximal eigenvalue of a Gaussian unitary random matrix*, Journal of Statistical Mechanics: Theory and Experiment **2011** (2011), no. 04, P04001.

[139] Rahul Nandkishore and David A. Huse, *Many-Body Localization and Thermalization in Quantum Statistical Mechanics*, Annual Review of Condensed Matter Physics **6** (2015), no. 1, 15–38.

[140] Alexei Onatski, *Determining the number of factors from empirical distribution of eigenvalues*, The Review of Economics and Statistics **92** (2010), no. 4, 1004–1016.

[141] Szilárd Pafka and Imre Kondor, *Noisy covariance matrices and portfolio optimization ii*, Physica A: Statistical Mechanics and its Applications **319** (2003), 487–494.

[142] Szilárd Pafka, Marc Potters, and Imre Kondor, *Exponential weighting and random-matrix-theory-based filtering of financial covariance matrices for portfolio optimization*, arXiv preprint cond-mat/0402573 (2004).

[143] Giorgio Parisi, *A sequence of approximated solutions to the sk model for spin glasses*, Journal of Physics A: Mathematical and General **13** (1980), no. 4, L115.

[144] Debashis Paul, *Asymptotics of sample eigenstructure for a large dimensional spiked covariance model*, Statistica Sinica (2007), 1617–1642.

[145] Debashis Paul and Alexander Aue, *Random matrix theory in statistics: a review*, Journal of Statistical Planning and Inference **150** (2014), 1–29.

[146] Debashis Paul and Jack W Silverstein, *No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix*, Journal of Multivariate Analysis **100** (2009), no. 1, 37–57.

[147] Sandrine Péché, *Universality of local eigenvalue statistics for random sample covariance matrices*, Ph.D. thesis, EPFL, 2003.

[148] ———, *Universality results for the largest eigenvalues of some sample covariance matrix ensembles*, Probability Theory and Related Fields **143** (2009), no. 3-4, 481–516.

[149] Anthony Perret and Grégory Schehr, *Finite n corrections to the limiting distribution of the smallest eigenvalue of Wishart complex matrices*, Random Matrices: Theory and Applications (2015), 1650001.

[150] Philippe Biane, *Free probability for probabilists*, Quantum Probability Communications **11** (2003), no. 11, 55–71.

[151] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luis A Nunes Amaral, Thomas Guhr, and H Eugene Stanley, *Random matrix approach to cross correlations in financial data*, Physical Review E **65** (2002), no. 6, 066126.

[152] Marc Potters, *A random matrix Bayesian framework for out-of-sample quadratic optimization*, Oral presentation at IAS Princeton Workshop on random matrices (2013).

[153] Jose Ramirez, Brian Rider, and Bálint Virág, *Beta ensembles, stochastic Airy spectrum, and a diffusion*, Journal of the American Mathematical Society **24** (2011), no. 4, 919–944.

[154] Daniel Revuz and Marc Yor, *Continuous martingales and Brownian motion*, vol. 293, Springer Science & Business Media, 2013.

[155] LCG Rogers and Z Shi, *Interacting Brownian particles and the Wigner law*, Probability theory and related fields **95** (1993), no. 4, 555–570.

[156] Thilo A Schmitt, Desislava Chetalova, Rudi Schäfer, and Thomas Guhr, *Non-stationarity in financial time series: Generic features and tail behavior*, EPL (Europhysics Letters) **103** (2013), no. 5, 58003.

[157] AM Sengupta and Partha P Mitra, *Distributions of singular values for some random matrices*, Physical Review E **60** (1999), no. 3, 3389.

[158] Dimitri Shlyakhtenko, *Random Gaussian band matrices and freeness with amalgamation*, International Mathematics Research Notices **1996** (1996), no. 20, 1013–1025.

[159] Jack W Silverstein, *Eigenvalues and eigenvectors of large dimensional sample covariance matrices*, Contemporary Mathematics **50** (1986), 153–159.

[160] ———, *On the eigenvectors of large dimensional sample covariance matrices*, Journal of multivariate analysis **30** (1989), no. 1, 1–16.

[161] ———, *Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices*, Journal of Multivariate Analysis **55** (1995), no. 2, 331–339.

[162] Jack W Silverstein and ZD Bai, *On the empirical distribution of eigenvalues of a class of large dimensional random matrices*, Journal of Multivariate analysis **54** (1995), no. 2, 175–192.

[163] Jack W Silverstein and Sang-Il Choi, *Analysis of the limiting spectral distribution of large dimensional random matrices*, Journal of Multivariate Analysis **54** (1995), no. 2, 295–309.

[164] Roland Speicher, *Free probability theory*, The Oxford handbook of Random Matrix Theory, Oxford University Press, 2011.

[165] Charles Stein, *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, Proceedings of the Third Berkeley symposium on mathematical statistics and probability, vol. 1, 1956, pp. 197–206.

[166] Akimichi Takemura, *An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population*, Tech. report, DTIC Document, 1983.

[167] Michel Talagrand, *The parisi formula*, Annals of Mathematics (2006), 221–263.

[168] Toshiyuki Tanaka, *Asymptotics of Harish-Chandra-Itzykson-Zuber integrals and free probability theory*, Journal of Physics: Conference Series, vol. 95, IOP Publishing, 2008, p. 012002.

[169] Antti Tanskanen, Jani Lukkarinen, and Kari Vatanen, *Random factor approach for large sets of equity time-series*, arXiv preprint arXiv:1604.05896 (2016).

[170] Terence Tao, http://terrytao.wordpress.com/2013/02/08/the-harish-chandra-itzykson-zuber-integral-formula/.

[171] ———, *Topics in random matrix theory*, vol. 132, American Mathematical Soc., 2012.

[172] Terence Tao and Van Vu, *Random matrices: universality of local eigenvalue statistics*, Acta mathematica **206** (2011), no. 1, 127–204.

[173] ———, *Random matrices: Universal properties of eigenvectors*, Random Matrices: Theory and Applications **1** (2012), no. 01, 1150001.

[174] Elena Tarquini, Giulio Biroli, and Marco Tarzia, *Level statistics and localization transitions of Lévy matrices*, Physical Review Letters **116** (2016), no. 1, 010601.

[175] Craig A Tracy and Harold Widom, *Level-spacing distributions and the Airy kernel*, Communications in Mathematical Physics **159** (1994), no. 1, 151–174.

[176] Antonio M Tulino and Sergio Verdú, *Random matrix theory and wireless communications*, Communications and Information theory **1** (2004), no. 1, 1–182.

[177] Michele Tumminello, Fabrizio Lillo, and Rosario N Mantegna, *Correlation, hierarchies, and networks in financial markets*, Journal of Economic Behavior & Organization **75** (2010), no. 1, 40–58.

[178] David E Tyler, *A distribution-free m-estimator of multivariate scatter*, The Annals of Statistics **15** (1987), no. 1, 234–251.

[179] Aad W Van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge university press, 2000.

[180] Pierpaolo Vivo, Satya N Majumdar, and Oriol Bohigas, *Large deviations of the maximum eigenvalue in Wishart random matrices*, Journal of Physics A: Mathematical and Theoretical **40** (2007), no. 16, 4317.

[181] Dan Voiculescu, *Symmetries of some reduced free product C*-algebras*, Springer, 1985.

[182] _____, *Limit laws for random matrices and free products*, Inventiones mathematicae **104** (1991), no. 1, 201–220.

[183] DV Voiculescu, KJ Dykema, and A. Nica, *Free random variables*, no. 1, American Mathematical Soc., 1992.

[184] Kenneth W Wachter, *The limiting empirical measure of multiple discriminant ratios*, The Annals of Statistics (1980), 937–957.

[185] Shanshan Wang, Rudi Schäfer, and Thomas Guhr, *Average cross-responses in correlated financial market*, arXiv preprint arXiv:1603.01586 (2016).

[186] HA Weidenmüller and GE Mitchell, *Random matrices and chaos in nuclear physics: Nuclear structure*, Reviews of Modern Physics **81** (2009), no. 2, 539.

[187] Hermann Weyl, *Inequalities between the two kinds of eigenvalues of a linear transformation*, Proceedings of the national academy of sciences **35** (1949), no. 7, 408–411.

[188] Norbert Wiener, *Extrapolation, interpolation, and smoothing of stationary time series*, vol. 2, MIT press Cambridge, MA, 1949.

[189] Eugene P Wigner, *On the statistical distribution of the widths and spacings of nuclear resonance levels*, Mathematical Proceedings of the Cambridge Philosophical Society, vol. 47, Cambridge Univ Press, 1951, pp. 790–798.

[190] Michael Wilkinson and Paul N Walker, *A Brownian motion model for the parameter dependence of matrix elements*, Journal of Physics A: Mathematical and General **28** (1995), no. 21, 6143.

[191] Tim Wirtz and Thomas Guhr, *Distribution of the smallest eigenvalue in the correlated Wishart model*, Physical review letters **111** (2013), no. 9, 094101.

[192] John Wishart, *The generalised product moment distribution in samples from a normal multivariate population*, Biometrika (1928), 32–52.

[193] Yanrong Yang, Guangming Pan, et al., *Independence test for high dimensional data based on regularized canonical correlation coefficients*, The Annals of Statistics **43** (2015), no. 2, 467–500.

[194] Jianfeng Yao, Abla Kammoun, and Jamal Najim, *Eigenvalue estimation of parameterized covariance matrices of large dimensional data*, Signal Processing, IEEE Transactions on **60** (2012), no. 11, 5893–5905.

[195] Yong Q Yin, *Limiting spectral distribution for a class of random matrices*, Journal of multivariate analysis **20** (1986), no. 1, 50–68.

[196] Anthony Zee, *Law of addition in random matrix theory*, Nuclear Physics B **474** (1996), no. 3, 726–744.

[197] Teng Zhang, Xiuyuan Cheng, and Amit Singer, *Marchenko-Pastur law for Tyler's and Maronna's M-estimators*, arXiv preprint arXiv:1401.3424 (2014).

[198] P Zinn-Justin, *Adding and multiplying random matrices: a generalization of Voiculescu's formulas*, Physical Review E **59** (1999), no. 5, 4884.

[199] Jean-Bernard Zuber, *The large-N limit of matrix integrals over the orthogonal group*, Journal of Physics A: Mathematical and Theoretical **41** (2008), no. 38, 382001.

[200] ———, *Introduction to random matrices*, 2012.

# Appendix A

# Harish-Chandra–Itzykson-Zuber integrals

## A.1 Definitions and results

The (generalized) Harish-Chandra-Itzykson-Zuber (HCIZ) integral [89, 95] $\mathcal{I}_\beta(\mathbf{A}, \mathbf{B})$ is defined as:

$$\mathcal{I}_\beta(\mathbf{A}, \mathbf{B}) = \int_{G(N)} \mathcal{D}\Omega \; e^{\frac{\beta N}{2} \mathrm{Tr} \mathbf{A}\Omega\mathbf{B}\Omega^*}, \tag{A.1.1}$$

where the integral is over the (flat) Haar measure of the compact group $\Omega \in G(N) = \mathbf{O}(N), \mathbf{U}(N)$ or $Sp(N)$ in $N$ dimensions and $\mathbf{A}, \mathbf{B}$ are arbitrary $N \times N$ symmetric (hermitian or symplectic) matrices. The parameter $\beta$ is the usual Dyson "inverse temperature", with $\beta = 1, 2$, or 4, respectively for the three groups.

This integral has found several applications in many different fields, including Random Matrix Theory, disordered systems or quantum gravity (for a particularly insightful introduction, see [170]). In RMT, this integral naturally appears in many problems, e.g. the derivation of the free addition and multiplication or the evaluation of eigenvalues density of states of a partition function whose potential is subject to a multiplicative external field. In statistics, this integral is also of particular interest. Indeed, let us reconsider the Bayesian framework of Chapter 6. We saw in Eq. (6.4.2) that the posterior distribution of the population covariance matrix $\mathbf{C}$ given the sample covariance $\mathbf{E}$ under a Gaussian assumption may be written as:

$$\mathcal{P}(\mathbf{C}|\mathbf{E}) = \frac{1}{Z} \exp\left[-\frac{N}{2} \mathrm{Tr}\, \boldsymbol{\mathcal{V}}(\mathbf{C}, \mathbf{E})\right], \qquad \boldsymbol{\mathcal{V}}(\mathbf{C}, \mathbf{E}) \; := \; \frac{1}{q}\left[\log \mathbf{C} + \mathbf{E}\mathbf{C}^{-1}\right] + V_0(\mathbf{C}), \tag{A.1.2}$$

where $V_0(\mathbf{C})$ is the potential function of the prior distribution and $Z$ the partition function. Using the BIPZ formalism of Section 3.1.2, the asymptotic behavior of the eigenvalues associated to this posterior distribution can be studied through the asymptotic of the partition function

$$Z \propto \int \exp\left[-\frac{N}{2} \mathrm{Tr}\, \boldsymbol{\mathcal{V}}(\mathbf{C}, \mathbf{E})\right] \mathrm{d}\mathbf{C}. \tag{A.1.3}$$

By integrating over the Haar measure of the Orthogonal group $\mathbf{O}(N)$, it is not hard to see that we end up with $\mathcal{I}_1(\mathbf{E}, \mathbf{C}^{-1})$. Therefore, the asymptotic behavior of the HCIZ integral is also

an important quantity in high dimensional statistics and this motivated the article [38] with Jean-Philippe Bouchaud, Satya N. Majumdar and Marc Potters.

In the unitary case $G(N) = U(N)$ and $\beta = 2$, it turns out that the HCIZ integral can be expressed exactly, for all $N$, as the ratio of determinants that depend on $\mathbf{A}, \mathbf{B}$, and additional $N$-dependent prefactors:

$$\mathcal{I}_{\beta=2}(\mathbf{A},\mathbf{B}) = \frac{c_N}{N^{(N^2-N)/2}} \frac{\det\left((e^{Na_ib_j})_{1\leq i,j\leq N}\right)}{\Delta(\mathbf{A})\Delta(\mathbf{B})} \tag{A.1.4}$$

with $\{a_i\}$, $\{b_i\}$ the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$, $\Delta(\mathbf{A}) = \prod_{i<j} |a_i - a_j|$ the Vandermonde determinant of $\mathbf{A}$ [and, similarly, for $\Delta(\mathbf{B})$], and $c_N = \prod_i^N i!$. Finding the expression of $\beta = 1$ or $\beta = 4$ is still an open problem.

Also, as is well known, determinants contain $N!$ terms of alternating signs, which makes their order of magnitude very hard to estimate *a priori*. This difficulty appears clearly when one is interested in the large $N$ asymptotics of HCIZ integrals, for which one would naively expect to have a simplified, explicit expression as a functional $F_2(\rho_{\mathbf{A}}, \rho_{\mathbf{B}}) = \lim_{N\to\infty} N^{-2} \ln \mathcal{I}_{\beta=2}(\mathbf{A},\mathbf{B})$ of the eigenvalue densities $\rho_{\mathbf{A},\mathbf{B}}$ of $\mathbf{A}, \mathbf{B}$ [129]. Using Dyson's Brownian motion, one can finds [38, 82]: $F_{\beta=2}(\mathbf{A},\mathbf{B}) = \lim_{N\to\infty} N^{-2} \ln \mathcal{I}_2(\mathbf{A},\mathbf{B})$:

$$F_2(\mathbf{A},\mathbf{B}) = -\frac{3}{4} - S_2(\mathbf{A},\mathbf{B}) + \frac{1}{2}\int \mathrm{d}x\, x^2(\rho_{\mathbf{A}}(x)+\rho_{\mathbf{B}}(x)) - \frac{1}{2}\int \mathrm{d}x\mathrm{d}y\, [\rho_{\mathbf{A}}(x)\rho_{\mathbf{A}}(y)+\rho_{\mathbf{B}}(x)\rho_{\mathbf{B}}(y)] \ln|x-y|,$$

where

$$S_2(\mathbf{A},\mathbf{B}) = \frac{1}{2}\int dt \int d\lambda\, \rho(\lambda,t) \left\{ v^2(\lambda,t) + \frac{\pi^2}{3}\rho^2(\lambda,t) \right\} \tag{A.1.5}$$

with $\rho(\lambda,t)$ and $v(\lambda,t)$ solution of the following Euler equation

$$\begin{cases} \partial_t \rho(\lambda,t) + \partial_\lambda[\rho(\lambda,t)v(\lambda,t)] = 0, \\ \partial_t v(\lambda,t) + v(\lambda,t)\partial_\lambda v(\lambda,t) = \dfrac{\pi^2}{2}\partial_\lambda \rho^2(\lambda,t), \\ \text{with } \rho(\lambda,0) = \rho_A(\lambda), \text{ and } \rho(\lambda,1) = \rho_B(\lambda). \end{cases} \tag{A.1.6}$$

In fact, this result can be extended to arbitrary value of $\beta$ with the final (simple) result $F_\beta(\mathbf{A},\mathbf{B}) = \beta F_2(\mathbf{A},\mathbf{B})/2$. This coincides with the result obtained by Zuber in the orthogonal case $\beta = 1$ [199] (see also [55, 83, 168] for arbitrary $\beta$).

Nonetheless, explicit results concerning the asymptotics of this integral are scarce. When $A$ and $B$ are both Wigner matrices, the Euler-Matytsin system of equation can be solved explicitly [38]. Another soluble case is when one of the two matrix has a Flat distribution [83]. Last but not least, a beautiful explicit result is available when one of the matrices has lower rank $n \ll N$. Precisely, let us assume that $\mathbf{A}$ has $n$ eigenvalues $a_1, a_2, \ldots, a_n$ when $N - n$ zero eigenvalues. Then we have [82, 124, 168]:

$$\mathcal{I}_\beta(\mathbf{A},\mathbf{B}) = \exp\left[ \frac{N\beta}{2} \sum_{i=1}^{n} \mathcal{W}_{\mathbf{B}}(a_i) \right], \tag{A.1.7}$$

where $\mathcal{W}_{\mathbf{B}}$ is the primitive of the $\mathcal{R}$-transform of $\mathbf{B}$. This result is of particular importance when we do Replica analysis since we introduce a finite number $n$ of "replicas" (see Section 3.1.4). We provide hereafter a complete derivation with elementary calculus in the rank-one case in the following section and explain how to generalize it to the rank-$n$ case.

## A.2 Derivation of (A.1.7) in the Rank-1 case

This section is devoted to the derivation of the result (A.1.7) in the sample case where $\mathbf{A} = \mathrm{diag}(a_1, 0, \dots, 0)$ and $\mathbf{B} = \mathrm{diag}(b_1, \dots, b_N)$. Firstly, we rewrite (A.1.1) (we set $\beta = 1$ for simplicity):

$$\mathcal{I}_1(\mathbf{A}, \mathbf{B}) = \frac{1}{\mathcal{Z}} \int \left( \prod_{k=1}^{N} \mathrm{d}\Omega_{1k} \right) \exp\left[ \frac{N}{2} a_1 \sum_{k=1}^{N} \Omega_{1k}^2 b_k \right] \delta\left( \sum_{k=1}^{N} \Omega_{1k}^2 - 1 \right), \tag{A.2.1}$$

where the Dirac delta function enforces the orthogonality and $\mathcal{Z}$ is normalization constant defined as:

$$\mathcal{Z} := \int \left( \prod_{k=1}^{N} \mathrm{d}\Omega_{1k} \right) \delta\left( \sum_{k=1}^{N} \Omega_{1k}^2 - 1 \right), \tag{A.2.2}$$

which allows us to omit constant variables in the following. We then use the following integral representation of the delta function:

$$\delta\left( \sum_{k=1}^{N} \Omega_{1k}^2 - 1 \right) = \frac{1}{2\pi} \int \exp\left[ \mathrm{i}\zeta \Big( \sum_{k=1}^{N} \Omega_{1k}^2 - 1 \Big) \right] \mathrm{d}\zeta, \tag{A.2.3}$$

so that we have (after renaming $\zeta = -2\mathrm{i}\zeta/N$)

$$\begin{aligned} \mathcal{I}_1(\mathbf{A}, \mathbf{B}) \quad &\propto \quad \frac{N}{4\pi} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}\zeta \int \left( \prod_{k=1}^{N} \mathrm{d}\Omega_{1k} \right) \exp\left[ \frac{N}{2} \left( a_1 \sum_{k=1}^{N} \Omega_{1k}^2 b_k + \zeta \Big( \sum_{k=1}^{N} \Omega_{1k}^2 - 1 \Big) \right) \right] \\ &= \quad \frac{N}{4\pi} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}\zeta \exp\left[ \frac{N\zeta}{2} \right] \int \left( \prod_{k=1}^{N} \mathrm{d}\Omega_{1k} \right) \exp\left[ -\frac{N}{2} \sum_{k=1}^{N} \Omega_{1k}^2 \Big( \zeta - a_1 b_k \Big) \right] \\ &= \quad \frac{N}{4\pi} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \exp\left[ -\frac{N}{2} \left( \frac{1}{N} \sum_{k=1}^{N} \log(\zeta - a_1 b_k) - \zeta \right) \right] \mathrm{d}\zeta. \end{aligned} \tag{A.2.4}$$

Since we consider $N \to \infty$, the integral over $\zeta$ is performed by a saddle-point method, leading to the following equation:

$$\frac{1}{N} \sum_{k=1}^{N} \frac{1}{\zeta - a_1 b_k} = 1, \tag{A.2.5}$$

which is equivalent to

$$\mathfrak{g}_{\mathbf{B}}(\zeta / a_1) = a_1. \tag{A.2.6}$$

We therefore find that

$$\zeta = a_1 \mathcal{B}_{\mathbf{B}}(a_1) = a_1 \mathcal{R}_{\mathbf{B}}(a_1) + 1. \tag{A.2.7}$$

By plugging this solution into (A.2.4), we obtain

$$\frac{2}{N} \log \mathcal{I}_1(\mathbf{A}, \mathbf{B}) \sim a_1 \mathcal{R}_{\mathbf{B}}(a_1) - \frac{1}{N} \sum_{k=1}^{N} \log\Big( 1 + a_1 (\mathcal{R}_{\mathbf{B}}(a_1 - b_k)) \Big). \tag{A.2.8}$$

One can then check, by taking the derivative of both sides, that

$$a_1 \mathcal{R}_{\mathbf{B}}(a_1) - \frac{1}{N} \sum_{k=1}^{N} \log\Big( 1 + a_1 (\mathcal{R}_{\mathbf{B}}(a_1 - b_k)) \Big) = \mathcal{W}_{\mathbf{B}}(a_1), \tag{A.2.9}$$

where $\mathcal{W}_{\mathbf{B}}$ is the primitive integral of the $\mathcal{R}$-transform of $\mathbf{B}$ satisfying $\mathcal{W}'_{\mathbf{B}}(\omega) = \mathcal{R}_{\mathbf{B}}(\omega)$. We therefore conclude that

$$\frac{2}{N}\log \mathcal{I}_1(\mathbf{A},\mathbf{B}) \sim \mathcal{W}_{\mathbf{B}}(a_1), \tag{A.2.10}$$

which is the claim.

Let us now explain briefly how to extend this derivation to the rank-$n$ case. Formally, the integral reads

$$\mathcal{I}_1(\mathbf{A},\mathbf{B}) = \frac{1}{\mathcal{Z}}\int \left(\prod_{i=1}^{n}\prod_{k=1}^{N} \mathrm{d}\Omega_{ik}\right) \exp\left[\frac{N}{2}\sum_{i=1}^{n} a_i \sum_{k=1}^{N} \Omega_{ik}^2 b_k\right] \prod_{i,j=1}^{n} \delta\Big(\sum_{k=1}^{N} \Omega_{ik}\Omega_{jk} - \delta_{ij}\Big), \tag{A.2.11}$$

where the normalization $\mathcal{Z}$ is easily deduced from (A.2.2), and $\mathbf{A} = \mathrm{diag}(a_1, a_2, \ldots, a_n, \ldots, 0)$. When $n = \mathcal{O}(N)$, i.e. when $\mathbf{A}$ has close to full rank, the orthogonality constraint $\sum_{k=1}^{N} \Omega_{ik}\Omega_{jk} = 0$ for $i \neq j$ becomes dominant and makes the calculation difficult. However, when $n \ll N$, this constraint is nearly automatically satisfied since two random unit vectors in $N$ dimensions have naturally a scalar product of order $1/\sqrt{N}$. In this limit, only the normalisation constraint is operative, i.e. $\sum_{k=1}^{N} \Omega_{ik}^2 = 1$, $\forall i = 1, \ldots, n$. But one then easily sees that the above integral factorizes into $n$ independent integrals of the type we considered above, hence leading to result (A.1.7) above. For a more rigourous proof that this result holds as long as $n \ll \sqrt{N}$, see [82].

## A.3 Instantion calculations for the full rank HCIZ integral

*This section is based on the article [38] written with Jean-Philippe Bouchaud, Satya N. Majumdar and Marc Potters.*

**A.3.1. Non-intersecting Brownian motions and HCIZ integral.** The first part of the derivation of (A.1.5) is relate the HCIZ integral (3.1.94) with the transition probability of non-intersecting Brownian motions with constrained initial and final positions. Consider N non-intersecting Brownian particles $\lambda_i(t), i \in \{1, \ldots, N\}$ where we assume that

$$\lambda_1(t) < \lambda_2(t) < \cdots < \lambda_N(t), \qquad \forall\, t \in [0,1] \tag{A.3.1}$$

and such that the $\{\lambda_i\}$ start with a given body density $\rho_{\mathbf{A}}$ and force to end at a body density $\rho_{\mathbf{B}}$ at time $t = 1$. This can be interpreted as the motion of the eigenvalues of a Gaussian matrix $\mathbf{M}(t)$ for $t \in [0,1]$ whose initial and final positions are constrained. The probability of each particles can be seen as a conditional probability that can be computed using the transition probability of the Brownian particles

$$\mathcal{P}(\mathbf{M}(t)|\mathbf{A}) = A_{N,\beta} \exp\left(-\frac{\beta N}{4}\mathrm{Tr}(\mathbf{M}(t) - \mathbf{A})^2\right), \tag{A.3.2}$$

where $A_{N,\beta}$ is a normalization constant defined by

$$A_{N,\beta} = \frac{1}{2^{N/2}}\left(\frac{\beta N}{2\pi}\right)^{N/2+\beta N(N-1)/4}. \tag{A.3.3}$$

The aim is to relate the conditional probability (A.3.2) with the HCIZ integral so the first part is to rewrite Eq. (A.3.2) as a function of the eigenvalues (or particles) $\{\lambda_i\}$ of $\mathbf{M}(t)$. As the

matrix $\mathbf{A}$ is fixed, we can assume without loss of generality that $\mathbf{A}$ is a diagonal matrix of eigenvalues $\{a_i\}_i$. Then, we perform a spectral decomposition of the matrix $\mathbf{M}(t) = \mathbf{U}\mathbf{\Lambda}(t)\mathbf{U}^*$ where $\mathbf{U}$ denotes the eigenvectors of $\mathbf{M}(t)$ that belongs to the Haar measure over unitary matrices ($\beta = 2$) or real orthogonal matrices ($\beta = 1$). One can check that this transformation implies a a Jacobian in Eq. (A.3.2) given by $\Delta^\beta(\mathbf{M}(t))\mathrm{d}(\mathbf{U}\Lambda(t)\mathbf{U}^*)$ where[1]

$$\Delta^\beta(\mathbf{M}(t)) = \prod_{i<j} |\lambda_i - \lambda_j|^\beta \,, \tag{A.3.4}$$

is the well known vandermonde determinant. Hence, for $t \in (0, 1]$, using that the RHS of Eq. (A.3.2) is invariant under rotation, one can integrate out over the Haar measure associated to $\mathbf{U}$ (see [8] for details) and one has:

$$\mathcal{P}(\mathbf{M}(t)|\mathbf{A}) = A_{N,\beta}C_{N,\beta} \int d\lambda_1 ... d\lambda_N \Delta^\beta(\{\lambda_i\}) e^{-\frac{\beta N}{4}\mathrm{Tr}[\mathbf{\Lambda}^2(t)+\mathbf{A}^2]} \int \mathcal{D}\mathbf{U} e^{\frac{\beta N}{2}\mathrm{Tr}\,\mathbf{U}\mathbf{\Lambda}(t)\mathbf{U}^*\mathbf{A}} \,, \tag{A.3.5}$$

where

$$C_{N,\beta} = \frac{\pi^{\beta N(N-1)/4} \prod_{j=1}^N \left(\frac{\beta}{2}\right)!}{\prod_{j=1}^N \left(\frac{j\beta}{2}\right)!} \,, \tag{A.3.6}$$

such that

$$K_{N,\beta} = A_{N,\beta}C_{N,\beta} = (2\pi)^{-N/2} \left(\frac{\beta N}{2}\right)^{N/2+\beta N(N-1)/4} \prod_{j=1}^N \left(\frac{\beta}{2}\right)! \prod_{j=1}^N \left[\left(\frac{j\beta}{2}\right)!\right]^{-1} \,, \tag{A.3.7}$$

and that indeed describle the joint distribution for the GOE (and GUE) [8,99]. So, we managed to find a relation between the HCIZ integral (for $\beta = \{1, 2\}$) and the conditional probability that the particles $\{\lambda_i\}$ (i.e. the eigenvalues) end with a body density $\rho_B$ at t = 1 knowing that they start from a density $\rho_A$ at t = 0:

$$\mathcal{P}_\beta(\{b_i\}|\{a_j\}) = K_{N,\beta}\Delta^\beta(\{\lambda_i\}) e^{-\frac{\beta N}{4}\mathrm{Tr}[\mathbf{B}^2+\mathbf{A}^2]} \mathcal{I}_\beta(\mathbf{B}, \mathbf{A}) \,, \tag{A.3.8}$$

with $\mathcal{I}_\beta(\mathbf{B}, \mathbf{A})$ defined in Eq. (3.1.94) and where we used that $\lambda_i(1) = b_i$ by assumption. Therefore, we find the following result:

$$\mathcal{I}_\beta(\mathbf{B}, \mathbf{A}) = \frac{\mathcal{P}_\beta(\{b_i\}|\{a_j\}) e^{\frac{\beta N}{4}\mathrm{Tr}[\mathbf{B}^2+\mathbf{A}^2]}}{K_{N,\beta}\Delta^\beta(\{b_i\})} \,. \tag{A.3.9}$$

This result indicates that we can evaluate the large N limit of the Orthogonal version of HCIZ integral by considering that the $\{\lambda_i(t)\}$ obey Dyson's Brownian motion with $\beta = 1$, which is the purpose of the next subsection. Firstly, we emphasize as a consistency check that in the Unitary case ($\beta = 2$), we know that the HCIZ integral has an explicit expression [89,95]:

$$I_\beta(\mathbf{B}, \mathbf{A}) = \frac{\prod_{j=1}^N j!}{N^{N(N-1)/2}} \frac{\det\left(e^{Nb_ia_j}\right)}{\Delta^2(\mathbf{B})\Delta^2(\mathbf{A})} \tag{A.3.10}$$

and hence, by plugging this latter equation into Eq. (A.3.8), one finds

$$\mathcal{P}_{\beta=2}(\{\lambda_i(t)\}) = \left(\frac{N}{2\pi}\right)^{N/2} \frac{\Delta(\mathbf{B})}{\Delta(\mathbf{A})} \det\left(e^{-\frac{N}{2}(b_i-a_j)^2}\right) \tag{A.3.11}$$

like in [99, Lemma 2.1].

[1] We shall omit the parameter $t$ in our notation when there is no confusion.

**A.3.2. Dyson Brownian motion argument.** We start by assuming that our $N$ particles (actually eigenvalues) obey Dyson's Brownian motion equation, which is a consequence of assuming that the matrix $\mathbf{M}(t)$ starts at $\mathbf{A}$ at time $t = 0$ and "diffuses" (in a $N(N-1)/2$ space) up to $\mathbf{B}$ at time $t = 1$. Restricted to the $N$ dimensional space of eigenvalues $\{\lambda_i\}$, one has for any $i = 1, \dots, N$:

$$d\lambda_i(t) = \sqrt{\frac{2}{\beta N}}\, \mathrm{d}W_i(t) + \frac{1}{N} \sum_{k \neq i} \frac{\mathrm{d}t}{\lambda_i(t) - \lambda_k(t)}\,, \tag{A.3.12}$$

with the $\{W_i(t)\}_i$ is a family of independent Brownian motion. As noticed in Eq. (A.3.9), we may set without loss of generality $\beta = 2$ (unitary matrices) as the result for $\beta = 1$ is identical up to a factor $1/2$.

Our question is: what is the (exponentially small) probability that the $\{\lambda_i\}$ start from a configuration with one body density $\rho_{\mathbf{A}}(\lambda)$ and ends at time $t$ with a one body density $\rho_{\mathbf{B}}(\lambda)$ ? This conditional probability is exactly the one that appears in Eq. (A.3.9). We will use instanton calculations to address this issue, using two different (but complementary) languages: that of particle trajectories and that of densities.

We start by the particles point of view. We introduce the total potential energy $U(\{\lambda_i\}) = -\frac{1}{N}\sum_{i<j} \ln|\lambda_i - \lambda_j|$, and the corresponding "force" $f_i = -\partial_{\lambda_i} U$. The probability of a given trajectory for the $N$ Brownian motions is given by:[2]

$$\mathcal{P}(\{\lambda_i(t)\}|\{a_i(t)\}) = \mathcal{N} \exp - \left[ \frac{N}{2} \int_0^1 \mathrm{d}t \sum_i \left( \dot{\lambda}_i + \partial_{\lambda_i} U \right)^2 \right]. \tag{A.3.13}$$

The action $S$ (i.e. the term in the exponential) can be decomposed in two separate terms. More precisely, we have $S = S_1 + S_2$ so that

$$\mathcal{P}(\{\lambda_i(t)\}|\{a_i(t)\}) \propto \exp\Big[-\big(S_1 + S_2\big)\Big], \tag{A.3.14}$$

where $S_1$ contains a total derivative, leading to:

$$S_1 \;:=\; -\frac{N}{2} \int \mathrm{d}\lambda \mathrm{d}\lambda' \rho_{\mathbf{A}}(\lambda)\rho_{\mathbf{A}}(\lambda') \ln|\lambda - \lambda'| + \frac{N}{2} \int \mathrm{d}\lambda \mathrm{d}\lambda' \rho_{\mathbf{B}}(\lambda)\rho_{\mathbf{B}}(\lambda') \ln|\lambda - \lambda'|\,, \tag{A.3.15}$$

and the second one is given by

$$S_2 \;:=\; \frac{N}{2} \int_0^1 \mathrm{d}t \sum_i \left[ \dot{\lambda}_i^2 + (\partial_{\lambda_i} U)^2 \right] \tag{A.3.16}$$

The "instanton" trajectory that dominates the probability for large $N$ is such that the functional derivative with respect to all $\lambda_i(t)$ is zero (see e.g. [34]):

$$-2\frac{d^2\lambda_i}{dt^2} + 2\sum_j \partial^2_{\lambda_i,\lambda_j} U \partial_{\lambda_j} U = 0 \tag{A.3.17}$$

which is equivalent after some algebraic manipulations to:

$$\frac{d^2\lambda_i}{dt^2} = -\frac{2}{N^2} \sum_{\ell \neq i} \frac{1}{(\lambda_i - \lambda_\ell)^3}. \tag{A.3.18}$$

[2] We neglect the Jacobian which is small in the large $N$ (small temperature) limit, as usual.

This can be interpreted as the motion of unit mass particles, accelerated by an *attractive* force that derives from an effective two-body potential $\phi(r) = -(Nr)^{-2}$. The hydrodynamical description of such a fluid is given by the Euler equations for the density $\rho(\lambda, t)$ and the velocity field $v(\lambda, t)$:

$$\begin{cases} \partial_t \rho(\lambda,t) + \partial_\lambda[\rho(\lambda,t)v(\lambda,t)] = 0, \\ \partial_t v(\lambda,t) + v(\lambda,t)\partial_\lambda v(\lambda,t) = -\dfrac{1}{\rho(\lambda,t)}\partial_\lambda P(\lambda,t)\,, \\ \text{with } \rho(\lambda,0) = \rho_A(\lambda), \text{ and } \rho(\lambda,1) = \rho_B(\lambda). \end{cases} \tag{A.3.19}$$

where $P(x,t)$ is the pressure field which reads, from the virial formula in one dimension [112, p. 138]:

$$P = \rho k T - \frac{1}{2}\rho \sum_{\ell \neq i} |\lambda_i - \lambda_\ell| \phi'(\lambda_i - \lambda_\ell) \approx -\frac{\rho}{N^2} \sum_{\ell \neq i} \frac{1}{(\lambda_i - \lambda_\ell)^2}, \tag{A.3.20}$$

because the fluid is at temperature $N^{-1}$. Using the same argument as Matytsin [129], i.e, writing

$$\lambda_i - \lambda_\ell \approx (i - \ell)/(N\rho) \tag{A.3.21}$$

and

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \tag{A.3.22}$$

one finally finds:

$$P(\lambda, t) = -\frac{\pi^2}{3}\rho^3(\lambda, t)\,. \tag{A.3.23}$$

By plugging this equation into the second equation of (A.3.19), we see that we retrieve Eq. (A.1.6), as expected. Finally, going back to the action term $S_2$, and going to the continuous limit, one finds from the above results that:

$$S_2 = \sum_i \left[\dot{\lambda}_i^2 + (\partial_{\lambda_i} U)^2\right] \approx N \int \mathrm{d}\lambda \rho(\lambda, t) \left[v^2(\lambda, t) + \frac{\pi^2}{3}\rho^2(\lambda, t)\right], \tag{A.3.24}$$

which yields exactly the action (A.1.5) by plugging this into (A.3.14) with (A.3.15), apart from the $-3/4$ term which comes from the prefactor in Eq. (A.3.9).

**A.3.3. Dean-Kawasaki equation.** We now consider the densities point of view. Suppose that we have $N$ particles and each of them obeys the following Langevin equation with an arbitrary two body potential interaction $\phi(x - y)$:

$$\mathrm{d}\lambda_i = \frac{1}{\sqrt{N}}\mathrm{d}W_i(t) - \frac{1}{N}\mathrm{d}t \sum_{j \neq i} \partial_{\lambda_i}\phi(\lambda_i - \lambda_j)\,. \tag{A.3.25}$$

Since the work of Kawasaki [106] (see the work of Dean [61] for a much more understandable derivation) we know that there is an exact Langevin equation for the density field associated to these particles and it reads [61]:

$$\partial_t \rho(\lambda, t) + \partial_\lambda J(\lambda, t) = 0 \tag{A.3.26}$$

with:

$$J(\lambda, t) = \frac{1}{N}\xi(\lambda, t)\sqrt{\rho(\lambda, t)} - \rho(\lambda, t) \int \mathrm{d}\lambda' \partial_\lambda \phi(\lambda - \lambda')\rho(\lambda', t) - \frac{1}{2N}\partial_\lambda \rho(\lambda, t), \tag{A.3.27}$$

where $\xi(\lambda,t)$ is a normalized Gaussian white noise (in time and in space) and $\rho := \frac{1}{N}\sum_i \rho_i$.

One can again write the weight of histories of $\{\rho(\lambda,t)\}$ using Martin-Siggia-Rose path integrals [60, 97]. This reads:

$$\mathcal{P}(\{\rho(\lambda,t)\}) = \left\langle \mathcal{N} \int \mathcal{D}\psi \exp\left[\int_0^1 \mathrm{d}t \int \mathrm{d}\lambda N^2 i\psi(\lambda,t)\left(\partial_t\rho + \partial_\lambda J\right)\right]\right\rangle_\xi \tag{A.3.28}$$

Performing the average over $\xi$ gives the following action (and renaming $-i\psi \to \psi$):

$$\mathcal{P}(\{\rho(\lambda,t)\}) = \exp\left[-S_{\mathrm{DK}}\right], \quad S_{\mathrm{DK}} := N^2 \int_0^1 \mathrm{d}t \int \mathrm{d}\lambda \left[\psi\partial_t\rho + F(\lambda,t)\rho\partial_\lambda\psi - \frac{\psi}{2N}\partial_\lambda^2\rho + \frac{1}{2}\rho(\partial_\lambda\psi)^2\right], \tag{A.3.29}$$

with $F(\lambda,t) = \int \mathrm{d}\lambda' \partial_\lambda\phi(\lambda-\lambda')\rho(\lambda',t)$. Taking functional derivatives with respect to $\rho$ and $\psi$ then leads to the following set of equations:

$$\begin{cases} \partial_t\rho = \partial_\lambda(\rho F) + \partial_\lambda(\rho\partial_\lambda\psi) + \dfrac{1}{2N}\partial_\lambda^2\rho\,, \\ \partial_t\psi - \dfrac{1}{2}(\partial_\lambda\psi)^2 = F\partial_\lambda\psi - \partial_\lambda \displaystyle\int \mathrm{d}\lambda' \phi(\lambda-\lambda')\rho(\lambda',t)\partial_{\lambda'}\psi(\lambda',t) - \dfrac{1}{2N}\partial_\lambda^2\psi\,, \\ \text{with } \rho(\lambda,0) = \rho_A(\lambda), \text{ and } \rho(\lambda,1) = \rho_B(\lambda). \end{cases} \tag{A.3.30}$$

Note that there are additional "diffusion" terms that are of order $1/N$. This is interesting since this should regularize the shocks in the (Burgers) equation for $\psi$. We will show that Eq. (A.3.30) yields the system of PDE (A.1.6). Compared to (A.1.6), the first equation of Eq. (A.3.30) suggests to identify the velocity field as $v(\lambda,t) = -F(\lambda,t) - \partial_\lambda\psi(\lambda,t)$, leading to:

$$\partial_t\rho + \partial_\lambda(\rho v) = \frac{1}{2N}\partial_\lambda^2\rho = \mathcal{O}(N^{-1})\,. \tag{A.3.31}$$

Applying this transformation for the second PDE of Eq. (A.3.30) leads to:

$$\partial_t v + v\partial_\lambda v = -\partial_t F + F\partial_\lambda F + \partial_\lambda^2 \int \mathrm{d}\lambda' \phi(\lambda-\lambda')\rho(\lambda',t)(v(\lambda',t) + F(\lambda',t)) + \mathcal{O}(N^{-1})\,,$$

that can be rewritten as:

$$\partial_t v + v\partial_\lambda v = F\partial_\lambda F - \partial_\lambda \int \mathrm{d}\lambda' \partial_\lambda\phi(\lambda-\lambda')\rho(\lambda',t) \int \mathrm{d}z \partial_{\lambda'}\phi(\lambda'-z)\rho(z,t) + \mathcal{O}(N^{-1}). \tag{A.3.32}$$

If we consider the logarithmic repulsive interaction as in Eq. (A.3.12), we may write:

$$\int \mathrm{d}\lambda' \partial_\lambda\phi(\lambda-\lambda')\rho(\lambda',t) \int \mathrm{d}z \partial_{\lambda'}\phi(\lambda'-z)\rho(z,t) = \frac{1}{N^2} \sum_{j\neq i; k\neq j} \frac{1}{(\lambda_i-\lambda_j)(\lambda_j-\lambda_k)} \tag{A.3.33}$$

which can be transformed into

$$\frac{1}{N^2} \sum_{j\neq i; k\neq j} \frac{1}{(\lambda_i-\lambda_j)(\lambda_j-\lambda_k)} = \frac{1}{2N^2} \sum_{j\neq k\neq i} \frac{1}{\lambda_j-\lambda_k}\left[\frac{1}{\lambda_i-\lambda_j} - \frac{1}{\lambda_i-\lambda_k}\right] - \frac{1}{N^2}\sum_{j\neq i} \frac{1}{(\lambda_i-\lambda_j)^2}\,. \tag{A.3.34}$$

The RHS of the latter equation can be simplified using simple algebraic manipulations:

$$\begin{aligned}\frac{1}{N^2}\sum_{j\neq i;k\neq j}\frac{1}{(\lambda_i-\lambda_j)(\lambda_j-\lambda_k)} &= \frac{1}{2N^2}\sum_{j\neq k\neq i}\frac{1}{\lambda_j-\lambda_k}\left[\frac{1}{\lambda_i-\lambda_j}-\frac{1}{\lambda_i-\lambda_k}\right]-\frac{1}{N^2}\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}\,,\\ &= \frac{1}{2N^2}\sum_{j\neq k\neq i}\frac{1}{(\lambda_i-\lambda_j)(\lambda_i-\lambda_k)}-\frac{1}{N^2}\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}\,,\\ &= \frac{1}{2}\left(\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)}\right)^2-\frac{3}{2}\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}. \end{aligned} \tag{A.3.35}$$

Using Eq. (A.3.21), she second term of the RHS is given by

$$-\frac{3}{2N^2}\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}\approx-\frac{3}{2}\times\frac{\pi^2}{3}\rho(\lambda,t)^2\,. \tag{A.3.36}$$

Using the logarithmic repulsion, one has

$$\frac{1}{2}\partial_\lambda F^2=\frac{1}{2}\partial_\lambda\left(\frac{1}{N}\sum_{j\neq i}\frac{1}{\lambda_i-\lambda_j}\right)^2. \tag{A.3.37}$$

All in all, it suffices to plug this last equation and (A.3.35) into (A.3.32) to find for the second PDE of (A.3.30):

$$\partial_t v+v\partial_\lambda v=\frac{\pi^2}{2}\partial_\lambda\rho^2+O(N^{-1})\,, \tag{A.3.38}$$

as expected.

It remains to prove that we indeed retrieve (A.1.5) using the action $S_{DK}$ defined in (A.3.29). By using the saddle-point equation Eq. (A.3.30), we obtain that

$$\begin{aligned}S_{\rm DK} &= N^2\int_0^1\mathrm{d}t\int\mathrm{d}\lambda\left[\psi(\partial_\lambda(\rho F)+\partial_\lambda(\rho\partial_\lambda\psi)+\frac{1}{2N}\partial_\lambda^2\rho)+F\rho\partial_\lambda\psi-\frac{\psi}{2N}\partial_\lambda^2\rho+\frac{1}{2}\rho(\partial_x\psi)^2\right],\\ &= N^2\int_0^1\mathrm{d}t\int\mathrm{d}\lambda\left[-\rho F\partial_\lambda\psi-\rho(\partial_\lambda\psi)^2+F\rho\partial_\lambda\psi+\frac{1}{2}\rho(\partial_\lambda\psi)^2\right],\\ &= N^2\int_0^1\mathrm{d}t\int\mathrm{d}\lambda\left[-\frac{1}{2}\rho(\partial_\lambda\psi)^2\right]. \end{aligned} \tag{A.3.39}$$

Next, observing that $\partial_\lambda\psi=-(F+v)$, we have

$$S_{\rm DK}=-N^2\int_0^1\mathrm{d}t\int\mathrm{d}\lambda\left[\frac{1}{2}\rho(F^2+2Fv+v^2)\right], \tag{A.3.40}$$

where we see the $\rho v^2$ term as in Eq. (A.1.5). It suffices to show that the remaining two terms yields the desired result. To that end, we first consider the $F^2$ term and we have for $N\to\infty$:

$$\begin{aligned}F^2 &= \int\mathrm{d}\lambda'\partial_\lambda\phi(\lambda-\lambda')\rho(\lambda',t)\int\mathrm{d}z\partial_\lambda\phi(\lambda-z)\rho(z,t)\\ &\approx \frac{1}{N^2}\sum_{j\neq i}\frac{1}{\lambda_i-\lambda_j}\sum_{k\neq i}\frac{1}{\lambda_i-\lambda_k}\\ &= \frac{1}{N^2}\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}+\frac{1}{N^2}\sum_{j\neq k\neq i}\frac{1}{(\lambda_i-\lambda_j)(\lambda_i-\lambda_k)}. \end{aligned} \tag{A.3.41}$$

Now, using symmetry arguments, we may find that

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}F^2(\lambda_i) &= \frac{1}{N^3}\sum_{i=1}^{N}\left[\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}+\sum_{j\neq k\neq i}\frac{1}{(\lambda_i-\lambda_j)(\lambda_i-\lambda_k)}\right],\\ &= \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}\,. \end{aligned} \tag{A.3.42}$$

Hence, by invoking once again (A.3.21), we get in the large N limit:

$$\int d\lambda\rho(\lambda)F^2(\lambda)=\frac{\pi^2}{3}\int\rho^3(\lambda)\mathrm{d}\lambda\,. \tag{A.3.43}$$

It remains to consider the last term in (A.3.40). The last quantity to handle is

$$\begin{aligned}S_3 &:= \int_0^1 dt\int d\lambda\rho(\lambda,t)F(\lambda,t)v(\lambda,t)\,,\\ &\equiv \int_0^1 dt\int d\lambda\rho(\lambda,t)v(\lambda,t)\int d\lambda'\partial_\lambda\phi(\lambda-\lambda')\rho(\lambda',t)\,, \end{aligned} \tag{A.3.44}$$

where we used that $F(\lambda,t):=\int\mathrm{d}\lambda'\partial_\lambda\phi(\lambda-\lambda')\rho(\lambda',t)$. By integration by parts, we get

$$\begin{aligned}S_3 &= -\int_0^1 dt\int d\lambda\partial_\lambda(\rho(\lambda,t)v(\lambda,t))\int d\lambda'\phi(\lambda-\lambda')\rho(\lambda',t)\,,\\ &= \int_0^1 dt\int d\lambda\partial_t(\rho(\lambda,t))\int d\lambda'\phi(\lambda-\lambda')\rho(\lambda',t)\,,\\ &= \frac{1}{2}\int_0^1 dt\partial_t\int d\lambda\int d\lambda'\phi(\lambda-\lambda')\rho(\lambda,t)\rho(\lambda',t)\,, \end{aligned} \tag{A.3.45}$$

where we invoked the equation of motions (e.g. the first equation of (A.3.30)) in the second line and symmetry argument in $t$ in the last one. Using the boundary condition, given in the last line of Eq. (A.3.30) and the logarithmic two body potential interaction for $\phi(\lambda-\lambda')$, we conclude that

$$S_3=\frac{1}{2}\left[\int d\lambda d\lambda'\log|\lambda-\lambda'|\rho_A(\lambda)\rho_A(\lambda')-\int d\lambda d\lambda'\log|\lambda-\lambda'|\rho_B(\lambda)\rho_B(\lambda')\right]. \tag{A.3.46}$$

The conclusion then easily follows by plugging Eqs. (A.3.43) and (A.3.46) into (A.3.40).

# Appendix B

# Reminders on linear algebra

## B.1 Schur complement

The derivation of recursion relation mostly relies on linear algebra. More specifically, let us define the $(N+M)\times(N+M)$ matrix $\mathbf{M}$ by

$$\mathbf{M} := \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}, \tag{B.1.1}$$

where the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and $\mathbf{D}$ are respectively of dimension $N\times N, N\times M, M\times N$ and $M\times M$. Suppose that $\mathbf{D}$ is invertible, then the *Schur complement* of the block $\mathbf{D}$ of the matrix $\mathbf{M}$ is given by the $N\times N$ matrix

$$\mathbf{M}/\mathbf{D} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}. \tag{B.1.2}$$

Using it, one obtains after using block Gaussian elimination (or LU decomposition) that the determinant of $\mathbf{M}$ can be expressed as

$$\det(\mathbf{M}) = \det(\mathbf{D})\det(\mathbf{M}/\mathbf{D}). \tag{B.1.3}$$

Moreover, one can write the inverse matrix $\mathbf{M}^{-1}$ in terms of $\mathbf{D}^{-1}$ and the inverse of the Schur complement (B.1.2)

$$\mathbf{M}^{-1} = \begin{pmatrix} (\mathbf{M}/\mathbf{D})^{-1} & -(\mathbf{M}/\mathbf{D})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{M}/\mathbf{D})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{M}/\mathbf{D})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}. \tag{B.1.4}$$

Similarly, if $\mathbf{A}$ is invertible, the Schur complement of the block $\mathbf{A}$ of the matrix $\mathbf{M}$ is given by the $M\times M$ matrix

$$\mathbf{M}/\mathbf{A} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}. \tag{B.1.5}$$

One easily obtains $\det(\mathbf{M})$ in terms of $\mathbf{A}$ and $\mathbf{M}/\mathbf{A}$ from (B.1.3) by replacing $\mathbf{D}$ by $\mathbf{A}$

$$\det(\mathbf{M}) = \det(\mathbf{A})\det(\mathbf{M}/\mathbf{A}). \tag{B.1.6}$$

The inverse matrix $\mathbf{M}^{-1}$ can also be written in terms of $\mathbf{A}^{-1}$ and the inverse of the Schur complement (B.1.5)

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{M}/\mathbf{A})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{M}/\mathbf{A})^{-1} \\ -(\mathbf{M}/\mathbf{A})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{M}/\mathbf{A})^{-1} \end{pmatrix}. \tag{B.1.7}$$

## B.2 Matrix identities

There are several useful identities that can be inferred from Schur complement formula. Firstly, using (B.1.4) and (B.1.7), we may immediately deduce the so-called *Woodbury* matrix identity

$$(\mathbf{A}+\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}+\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}. \tag{B.2.1}$$

Moreover, if $\mathbf{D} = I_M$, we get the *matrix determinant lemma* from (B.1.3) and (B.1.6)

$$\det(\mathbf{A}-\mathbf{B}\mathbf{C}) = \det(\mathbf{A})\det(\mathbf{I}_M - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}), \tag{B.2.2}$$

and if $\mathbf{A} = \mathbf{I}_N$ in addition, one gets *Sylvester's determinant identity*

$$\det(\mathbf{I}_N - \mathbf{B}\mathbf{C}) = \det(\mathbf{I}_M - \mathbf{C}\mathbf{B}). \tag{B.2.3}$$

Now, assuming that both $\mathbf{B}$ and $\mathbf{C}$ are column vectors, one readily find from (B.2.1) the *Sherman-Morrison* formula.

## B.3 Resolvent identities

Another useful application of Schur complement formula concerns the resolvent. We keep the notations of Section 3.1.1 and thus

$$\mathbf{G}(z) = \mathbf{H}^{-1}(z), \qquad \mathbf{H}(z) \;:=\; z\mathbf{I}_N - \mathbf{M}\,, \tag{B.3.1}$$

with $\mathbf{G}$ a $N \times N$ symmetric matrix. We now rewrite $\mathbf{H}(z)$ as a block matrix:

$$\mathbf{H}(z) \;=\; \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{pmatrix}, \tag{B.3.2}$$

where the matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ are respectively of dimension $K \times K$, $K \times M$ and $M \times M$ with $N = K + M$. Next, we define from (B.1.2) the schur complement $\mathbf{D} := \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^*$. In the following, we consider $K = 2$ for simplicity. We have for any $i, j \in \{1, 2\}$, we have from (B.1.4):

$$G_{ij} = (\mathbf{D}^{-1})_{ij}. \tag{B.3.3}$$

As a warmup exercise, let us first consider the simplest case $i = j$ ($K = 1$) and we set without loss of generality that $i = 1$. Then $\mathbf{A}$ becomes a scalar and so is $\mathbf{D}$. Using Eq. (B.3.1), one obtains $\mathbf{A} = z - M_{11}$, $\mathbf{B} = [M_{12}, \dots, M_{1N}]$ and $\mathbf{C} = \mathbf{H}^{(1)}(z)$ where $\mathbf{H}^{(i)}$ denotes the "minor" of $\mathbf{H}$, i.e. $\mathbf{H}^{(i)} := \big(H_{st} : s, t \in [\![1, N]\!]\backslash\{i\}\big)$. Hence, it is easy to see from the very definition of $\mathbf{D}$ that

$$\mathbf{D} \equiv D_{11} = z - M_{11} - \sum_{\alpha,\beta}^{(1)} M_{1\alpha} G^{(1)}_{\alpha,\beta} M_{\beta 1}, \tag{B.3.4}$$

where and we used the abbreviation

$$\sum_{\alpha,\beta}^{(i)} \equiv \sum_{\alpha,\beta \in [\![1,N]\!]\backslash\{i\}}. \tag{B.3.5}$$

Therefore, we deduce from (B.3.3) that

$$G_{11}(z) = \frac{1}{z - M_{11} - \sum_{\alpha\beta}^{(1)} M_{1\alpha} G^{(1)}_{\alpha,\beta} M_{\beta 1}}. \tag{B.3.6}$$

This last result holds for any other diagonal term of the resolvent $\mathbf{G}$.

Next, we consider the general case $K = 2$ so that $\mathbf{D}$ is a $2 \times 2$ matrix. Again, using the block representation (B.3.2) and Eq. (B.3.1), one deduces that:

$$D_{kl} = z\delta_{kl} - M_{kl} - \sum_{\alpha,\beta}^{(kl)} M_{k\alpha} G^{(kl)}_{\alpha,\beta} M_{\beta l}, \quad k, l \in [\![i, j]\!]. \tag{B.3.7}$$

It is not hard to see that $D_{kk}$ yields Eq. (B.3.4) as it should. Using that (B.3.7) is a $2 \times 2$ matrix, one can readily invert the matrix $\mathbf{D}$ to obtain the relation

$$G_{ij} - G^{(m)}_{ij} = \frac{G_{im} G_{mj}}{G_{mm}}, \tag{B.3.8}$$

for any $i, j \in [\![1, K]\!]$ and $m \in [\![1, N]\!]$ with $i, j \neq m$. This last equation allows one to write a recursion relation on the entries of the resolvent (see the following appendix).

## B.4 Applications: Self-consistent relation for resolvent and Central Limit Theorem

We focus in this section on another frequently used analytical tool in RMT based on recursion relation for the resolvent of a given matrix $\mathbf{M}$. This technique has many advantages compared to the method compared to the Replica analysis: (i) the entries of the matrix need not to be identically distributed, (ii) no ansatz is required to perform the calculations. In the limit of $N \to \infty$, an interesting application of the Central Limit Theorem (CLT) concerns the spectral properties of random matrices. Precisely, we shall see that relations like that of Eq. (5.1.1) are actually a consequence of the CLT.

**B.4.1. Wigner matrices.** As a warmup exercise, we consider the simplest ensemble of random matrices where all elements of the matrix $\mathbf{M}$ are iid random variables, with the only constraint that the matrix be symmetrical. This is the well-known Wigner ensemble where we assume that

$$\mathbb{E}[M_{ij}] = 0, \quad \mathbb{E}[M_{ij}^2] = \frac{\sigma^2}{N}, \tag{B.4.1}$$

for any $i, j \in [\![1, N]\!]$. Note that the scaling with $N^{-1}$ for the variance comes from the fact that we want the eigenvalues of $\mathbf{M}$ to stay bounded when $N \to \infty$. This allows to conclude that $M_{ij} \sim 1/\sqrt{N}$ for any $i, j \in [\![N]\!]$.

In order to derive a self-consistent equation for the resolvent of $\mathbf{M}$, we use (B.4.1) and Wick's theorem into (B.3.7) and one can check that

$$\begin{aligned} \mathbb{E}\left[\sum_{\alpha,\beta}^{(kl)} M_{k\alpha} G^{(kl)}_{\alpha\beta} M_{\beta l}\right] &= \delta_{kl} \frac{\sigma^2}{N} \sum_{\alpha}^{(k)} G^{(k)}_{\alpha\alpha} \\ \mathbb{V}\left[\sum_{\alpha,\beta}^{(kl)} M_{k\alpha} G^{(kl)}_{\alpha\beta} M_{\beta l}\right] &\sim \frac{\sigma^4}{N}. \end{aligned} \tag{B.4.2}$$

Consequently, using the Central Limit Theorem, we conclude that for Wigner matrices, (B.3.7) converges for large $N$ towards

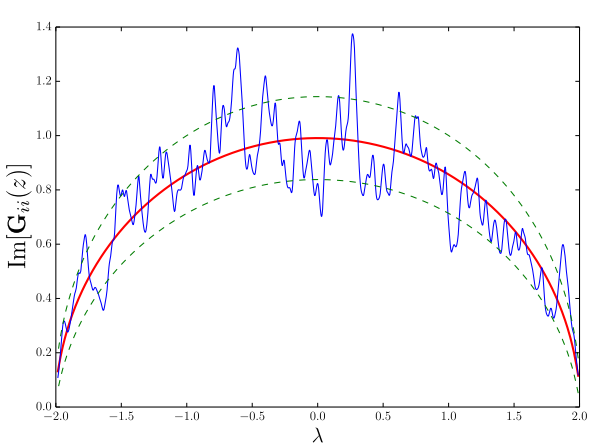
$$D_{kl} = \delta_{kl}\left(z - \frac{\sigma^2}{N}\sum_{\alpha}^{(k)} G^{(k)}_{\alpha\alpha}\right) + O(N^{-1/2}) \quad k, l \in \{i, j\}, \tag{B.4.3}$$

from which one deduces that $G_{ij} \sim N^{-1/2}$ using (B.3.3). Moreover, we may consistently check that $G^{(k)}_{\ell\ell} \sim G_{\ell\ell} + O(N^{-1})$ for any $\ell \in [\![1, N]\!]$ thanks to (B.3.8) and we therefore obtain for any $i \in [\![1, N]\!]$:

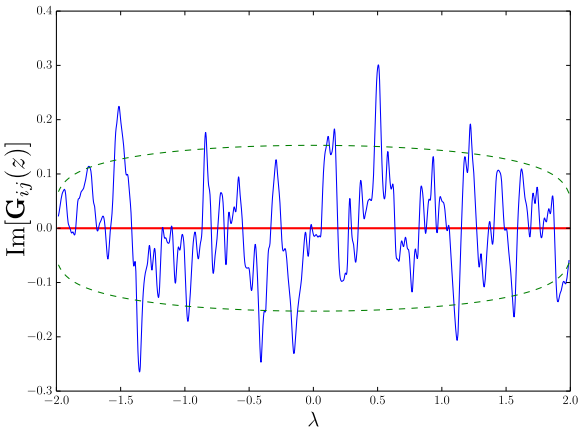
$$G_{ii} \sim \frac{1}{z - \sigma^2 \mathfrak{g}(z)} + O(N^{-1/2}). \tag{B.4.4}$$

By taking the normalized trace in this last equation, we obtain at leading order the equation of the semi-circle law's Stieltjes transform

$$\mathfrak{g}(z) = \frac{1}{z - \sigma^2(z)\mathfrak{g}}, \tag{B.4.5}$$




(A) Diagonal entry of $\mathrm{Im}[\mathbf{G}_{\mathbf{E}}(z)]$ with $i = 1$.




(B) Off diagonal entry of $\mathrm{Im}[\mathbf{G}_{\mathbf{E}}(z)]$ with $i = 1$ and $j = 2$.

FIGURE B.4.1. Illustration of the imaginary part of Eq. (B.4.6) with $N = 1000$. The empirical estimate of $\mathbf{G}_{\mathbf{E}}(z)$ (blue line) is computed for any $z = \lambda_i - \mathrm{i}N^{-1/2}$ with $i \in [\![1, N]\!]$ and comes from one sample. The theoretical one (red line) is given by the RHS of Eq. (B.4.6). The green dotted corresponds to the confidence interval whose formula is given by Eq. (B.4.7).

so that we conclude

$$G_{ij}(z) \sim \delta_{ij}\mathfrak{g}(z) + O(N^{-1/2}). \tag{B.4.6}$$

This result has been extended in a much more general framework – see e.g. the recent reviews [22, 75]. In particular, it is possible to show that the error term we obtain in Eq. (B.4.6) is quite similar to (5.1.3) and reads for $\eta = \widehat{\eta}N$ with $\widehat{\eta} \gg 1$:

$$\Psi_{\mathrm{GOE}}(z) \;:=\; \sqrt{\frac{\mathrm{Im}\,\mathfrak{g}_{\mathrm{SC}}(z)}{\widehat{\eta}}} + \frac{1}{\widehat{\eta}}, \tag{B.4.7}$$

provided that $N$ is large enough. We illustrate this ergodic behavior for the GOE in Figure 5.1.1, and we see the agreement is excellent and each diagonal entry indeed converges to the semicircle law.

**B.4.2. Sample covariance matrices.** We now want to derive (5.1.1) using the same type of arguments than in the previous section. Suppose that $\mathbf{E}$ is defined as in (4.1.3) and we denote by $\mathbf{G}(z)$ its resolvent. Let us assume for simplicity that $\mathbf{C} = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_N)$. Since $\mathbf{E}$ is a product of two rectangular matrices, it is convenient to introduce the $(N+T) \times (N+T)$ block matrix $\mathbf{R} \;:=\; (R_{ij}) \in \mathbb{R}^{(N+T)\times(N+T)}$ defined as:

$$\mathbf{R}(z) := \mathbf{H}^{-1}(z), \quad \mathbf{H}(z) := \begin{pmatrix} \mathbf{C}^{-1} & \mathbf{X} \\ \mathbf{X}^* & z\,\mathbf{I}_T \end{pmatrix}. \tag{B.4.8}$$

To simplify the notations, we introduce the set of indexes $\mathcal{I}_N := [\![1, N]\!]$ and $\mathcal{I}_T := [\![1, T]\!]$. Then using (B.1.4) and (B.1.7), we see that

$$R_{ij}(z) = z(\mathbf{C}^{1/2}\mathbf{G}_{\mathbf{E}}(z)\mathbf{C}^{1/2})_{ij}, \quad i, j \in \mathcal{I}_N, \tag{B.4.9}$$

where $\mathbf{E}$ is the sample covariance matrix defined in Eqs. (4.1.3) and (4.1.4), but also

$$R_{\alpha\beta}(z) = (\mathbf{G}_{\mathbf{S}}(z))_{\alpha\beta}, \qquad \alpha, \beta \in \mathcal{I}_T, \tag{B.4.10}$$

where the $T \times T$ matrix $\mathbf{S}$ is defined in Eq. (4.2.24).

We are interested in the computations of $R_{ij}$ for $i, j \in \mathcal{I}_N$ and this can be done using (B.3.3) and (B.3.7). Note that one can finds $R_{\alpha\beta}$ by proceeding in the same way. We obtain from (B.3.3) and (B.3.7) that

$$R_{ij}(z) = (\mathbf{D}^{-1})_{ij}, \qquad D_{kl} := \frac{\delta_{kl}}{\mu_k} - \sum_{\alpha,\beta \in \mathcal{I}_T} X_{k\alpha} R^{(kl)}_{\alpha\beta} X_{l\alpha}. \tag{B.4.11}$$

for any $k, l \in \{i, j\}$. Using that $\mathbb{E}[X_{it}] = 0$ and $\mathbb{E}[X_{it}^2] = T^{-1}$ from (4.1.5), we remark thanks to Wick's theorem that the sum in the term $D_{kl}$ obeys

$$\begin{aligned} \mathbb{E}\left[ \sum_{\alpha,\beta \in \mathcal{I}_T} X_{k\alpha} R^{(kl)}_{\alpha\beta} X_{l\alpha} \right] &= \frac{\delta_{kl}}{T} \sum_{\alpha}^{(k)} R^{(k)}_{\alpha\alpha} \\ \mathbb{V}\left[ \sum_{\alpha,\beta \in \mathcal{I}_T} X_{k\alpha} R^{(kl)}_{\alpha\beta} X_{l\alpha} \right] &\sim \frac{1}{T}, \end{aligned} \tag{B.4.12}$$

where we used the notation (B.3.5) for the sum. Invoking once again the CLT, we find that the entry $D_{kl}$ converges for large $N$ towards

$$D_{kl} \sim \delta_{kl} \left( \frac{1}{\mu_k} - \frac{1}{T} \sum_{\alpha \in \mathcal{I}_T} R^{(k)}_{\alpha\alpha} \right) + O(T^{-1/2}), \tag{B.4.13}$$

so that we may conclude from (B.4.11) that $R_{ij} \sim O(T^{-1/2})$ for $i \neq j$. Note that one may repeat the same arguments for $R_{\alpha\beta}$ with $\alpha, \beta \in \mathcal{I}_T$ to obtain

$$D_{\alpha\beta} \sim \delta_{\alpha\beta} \left( z - \frac{1}{T} \sum_{k \in \mathcal{I}_N} R^{(\alpha)}_{kk} \right) + \mathcal{O}(T^{-1/2}), . \tag{B.4.14}$$

Let us now investigate $R^{(k)}_{\alpha\alpha}$ which can be rewritten thanks to (B.3.8) as:

$$R^{(k)}_{\alpha\alpha} = R_{\alpha\alpha} - \frac{R_{k\alpha} R_{\alpha k}}{R_{kk}}. \tag{B.4.15}$$

We deduce from (B.4.13) that $R_{kk} \sim O(1)$. We will now show that $R_{k\alpha}$ (and $R_{\alpha k}$) are vanishing as $T^{-1/2}$. To that end, we apply (B.1.7) to (B.4.8) to find

$$R_{k\alpha} = -\big(\mathbf{C}\mathbf{X}\mathbf{G}_{\mathbf{S}}\big)_{k\alpha} = -\mu_k \sum_{\beta \in \mathcal{I}_T} X_{k\beta} (\mathbf{G}_{\mathbf{S}})_{\beta\alpha}. \tag{B.4.16}$$

Using Eqs. (B.4.10), (B.4.14) and that $X_{k\beta} \sim T^{-1/2}$, one can self-consistently check that $R_{k\alpha} \sim T^{-1/2}$. This is also true for $R_{\alpha k}$. Hence, if we plug this into Eq. (B.4.15), we see that for $N \to \infty$:

$$\frac{1}{T} \sum_{\alpha}^{(k)} R^{(k)}_{\alpha\alpha} = \frac{1}{T} \sum_{\alpha}^{(k)} R_{\alpha\alpha} + O(T^{-1}) = \mathfrak{g}_{\mathbf{S}}(z) + O(T^{-1}), \tag{B.4.17}$$

and we therefore have from Eqs. (B.4.13) and (B.4.11):

$$R_{ij}(z) = \delta_{ij}\left(\frac{\mu_k}{1-\mu_k \mathfrak{g}_S(z)}\right) + O(T^{-1/2}). \tag{B.4.18}$$

Finally, recalling that $\mathfrak{g}_{\mathbf{S}}(z) = q\mathfrak{g}_{\mathbf{E}}(z) + (1-q)/z$ from Eq. (4.2.25) and $R_{ii} = z\mu_i G_{ii}$ from Eq. (B.4.9), we conclude that

$$(\mathbf{G}_{\mathbf{E}}(z))_{ij} = \delta_{ij}\left(\frac{1}{z-\mu_k(1-q+qz\mathfrak{g}_{\mathbf{E}}(z))}\right) + O(T^{-1/2}), \qquad i,j \in [\![1,N]\!], \tag{B.4.19}$$

which is the prediction obtained in (5.1.2) with the Replica method. Similarly, we obtain for the $T \times T$ block that:

$$(\mathbf{G}_{\mathbf{S}}(z))_{\alpha\beta} = \frac{\delta_{\alpha\beta}}{z - \frac{1}{T}\sum_{k\in\mathcal{I}_N}(\mathbf{G}_{\mathbf{E}}(z))_{ij}} + O(T^{-1/2})\,. \tag{B.4.20}$$

Moreover, by using (B.4.18) and (4.2.26), we see that for $N \to \infty$

$$z - \frac{1}{T}\sum_{k\in\mathcal{I}_N}(\mathbf{G}_{\mathbf{E}}(z))_{kk} = \frac{1}{\mathfrak{g}_{\mathbf{S}}(z)}, \tag{B.4.21}$$

so that we may conclude

$$(\mathbf{G}_{\mathbf{S}}(z))_{\alpha\beta} = \delta_{\alpha\beta}\mathfrak{g}_{\mathbf{S}}(z) + O(T^{-1/2})\,. \tag{B.4.22}$$

This last result highlights that it is often easier to work with the $T \times T$ sample covariance matrix $\mathbf{S}$ rather than with the $N \times N$ matrix $\mathbf{E}$ since the resolvent can be approximated simply by its normalized trace. All these results can be found in a much more general and rigorous context in [109].

# Appendix C

# Conventions, notations and abbreviations

### Conventions

We use bold capital letters for matrices and bold lowercase letters for vectors, which we regard as $N \times 1$ matrices. The superscript $*$ denotes the transpose operator. We use the abbreviations $[\![a, b]\!] := [a, b] \cap \mathbb{N}$ and $[\![a]\!] \equiv [\![1, a]\!]$ for $a, b \in \mathbb{N}$.

### Mathematical symbols

We list here some of the most important notations of this thesis.

| Symbol | Description |
|---|---|
| $\mathcal{B}_{\mathbf{M}}$ | Blue transform of $\mathbf{M}$ (3.1.15) |
| $\mathbf{C}$ | Population/True covariance matrix (4.1.1) |
| $\underline{\mathbf{C}}$ | Spikeless version of $\mathbf{C}$ (4.3.6) |
| $\mathbb{C}_{\pm}$ | Complex upper/lower half plane |
| $\mathbf{E}$ | Sample/Empirical covariance matrix (4.1.3) |
| $\mathbb{E}$ | Expectation value over the noise |
| $\mathbf{G}_{\mathbf{M}}$ | Resolvent of $\mathbf{M}$, (3.1.5) |
| $\mathfrak{g}_{\mathbf{M}}^{N}$ | Empirical Stieltjes transform of $\rho_{\mathbf{M}}$ (3.1.7) |
| $\mathfrak{g}_{\mathbf{M}}$ | Stieltjes transform of $\rho_{\mathbf{M}}$ (3.1.8) |
| $\mathrm{i}$ | $\sqrt{-1}$ |
| $i$ | integer index |
| $N$ | Number of variables |
| $\mathbf{O}(N)$ | Orthogonal group on $\mathbb{R}^{N \times N}$ |
| $\mathcal{O}$ | Big $O$ notation |
| $\mathcal{P}(\cdot)$ | Probability density function |
| $\mathcal{P}(\cdot\|\cdot)$ | Conditional probability measure |
| $q$ | Observation ratio $(N/T)$ |
| $r$ | Number of outliers |
| $\mathcal{R}_{\mathbf{M}}$ | R-transform of $\mathbf{M}$ (3.1.16) |
| $\mathcal{R}^2_{\text{in}}$ | In-sample/predicted risk (8.1.7) |
| $\mathcal{R}^2_{\text{out}}$ | Out-of-sample/realized risk (8.1.9) |
| $\mathcal{R}^2_{\text{true}}$ | True risk (8.1.5) |
| $\mathbf{S}$ | "Dual" sample covariance matrix (4.2.24) |

| | |
|---|---|
| $\mathcal{S}_{\mathbf{M}}$ | S-transform of $\mathbf{M}$ (3.1.23) |
| $T$ | Sample size |
| $\mathcal{T}_{\mathbf{M}}$ | T-transform of $\mathbf{M}$ (3.1.21) |
| $\mathbf{u}_i$ | Sample eigenvector associated to $\lambda_i$ |
| $\mathbf{v}_i$ | Population eigenvector associated to $\mu_i$ |
| $\mathcal{W}_{\mathbf{M}}$ | Primitive of the $\mathcal{R}$-Transform of $\mathbf{M}$ (3.1.96) |
| $\mathbf{Y}$ | $N \times T$ normalized data matrix |
| $\alpha_s$ | Linear shrinkage intensity (6.3.7) |
| $\lambda_i$ | $i$th sample eigenvalue |
| $\mu_i$ | $i$th population (true) eigenvalue |
| $\Xi^{\text{lin.}}$ | Linear Shrinkage estimator (6.3.7) |
| $\hat{\Xi}(\mathbf{E})$ | Optimal RIE of $\mathbf{C}$ depending on $\mathbf{E}$ |
| $\Xi^{\text{ora.}}$ | Oracle estimator (7.1.2) |
| $\Xi(\mathbf{E})$ | RIE of $\mathbf{C}$ depending on $\mathbf{E}$ |
| $\rho^N_{\mathbf{M}}$ | Empirical spectral density of $\mathbf{M}$ (3.1.3) |
| $\rho_{\mathbf{M}}$ | Limiting spectral density of $\mathbf{M}$ (3.1.4) |
| $\Phi$ | Rescaled mean squared overlap (5.0.3) and (5.0.4) |
| $\varphi(\mathbf{M})$ | Normalized trace of $\mathbf{M}$ (3.1.61) |
| $\mathbf{\Omega}$ | Rotation matrix |
| $\langle\cdot\rangle_{\mathbf{M}}$ | Expectation value with respect to $\mathcal{P}(\mathbf{M})$ |
| $\langle\,,\,\rangle$ | inner product |
| $\langle\,\cdot\,\vert\,\cdot\,\rangle$ | Dirac bra-ket |

**Abbreviations**

| **Symbol** | **Description** |
|---|---|
| CCA | Canonical Correlation Analysis |
| ESD | Empirical Spectral Density |
| GOE | Gaussian Orthogonal Ensemble |
| IW | Inverse Wishart |
| IWs | Inverse Wishart + sorting |
| LDA | Linear discriminant analysis |
| LDL | Large dimension limit |
| LHS | Left Hand Side |
| LSD | Limiting Spectral Density |
| MMSE | Minimum Mean Squared Error |
| MSE | Mean Squared Error |
| MP | Marčenko-Pastur |
| PCA | Principal Component Analysis |
| PDE | Partial Differential Equation |
| PDF | Probability Density Function |
| RHS | Right Hand Side |
| QuEST | Quantized Eigenvalues Sampling Transform |
| RHS | Right Hand Side |
| RI | Rotational Invariance |
| RIE | Rotational Invariant Estimator |

| | |
|---|---|
| RP | Relative Performance |
| RMT | Random Matrix Theory |
| SCM | Sample Covariance Matrix |
| SVD | Singular Value Decomposition |

ÉCOLE DOCTORALE
Physique en île-de-France (EDPIF)

**Titre :** Application de la théorie des matrices aléatoires pour les statistiques en grande dimension

**Mots clés :** Matrices aléatoires, statistiques en grande dimension, estimation, décomposition spectrale

**Résumé :** De nos jours, il est de plus en plus fréquent de travailler sur des bases de données de très grandes tailles dans plein de domaines différents. Cela ouvre la voie à de nouvelles possibilités d'exploitation ou d'exploration de l'information, et de nombreuses technologies numériques ont été crées récemment dans cette optique. D'un point de vue théorique, ce problème nous contraint à revoir notre manière d'analyser et de comprendre les données enregistrées. En effet, dans cet univers communément appelé « Big Data », un bon nombre de méthodes traditionnelles d'inférence statistique multivariée deviennent inadaptées. Le but de cette thèse est donc de mieux comprendre ce phénomène, appelé fléau (ou malédiction) de la dimension, et ensuite de proposer différents outils statistiques exploitant explicitement la dimension du problème et permettant d'extraire des informations fiables des données. Pour cela, nous nous intéresserons beaucoup aux vecteurs propres de matrices symétriques. Nous verrons qu'il est possible d'extraire de l'information présentant un certain degré d'universalité. En particulier, cela nous permettra de construire des estimateurs optimaux, observables, et cohérents avec le régime de grande dimension.

**Title :** Application of random matrix theory to high dimensional statistics

**Keywords :** random matrices, high dimensional statistics, estimation, spectral decomposition

**Abstract :** Nowadays, it is easy to get a lot of quantitative or qualitative data in a lot of different fields. This access to new data brought new challenges about data processing and there are now many different numerical tools to exploit very large database. In a theoretical standpoint, this framework appeals for new or refined results to deal with this amount of data. Indeed, it appears that most results of classical multivariate statistics become inaccurate in this era of "Big Data". The aim of this thesis is twofold: the first one is to understand theoretically this so-called curse of dimensionality that describes phenomena which arise in high-dimensional space. Then, we shall see how we can use these tools to extract signals that are consistent with the dimension of the problem. We shall study the statistics of the eigenvalues and especially the eigenvectors of large symmetrical matrices. We will highlight that we can extract some universal properties of these eigenvectors and that will help us to construct estimators that are optimal, observable and consistent with the high dimensional framework.

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France