Physics Reports 666 (2017) 1–109

Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

Cleaning large correlation matrices: Tools from Random Matrix Theory

Joël Bun [a,b,*], Jean-Philippe Bouchaud [a], Marc Potters [a]

[a] Capital Fund Management, 23–25, rue de l'Université, 75 007 Paris, France
[b] LPTMS, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay, France

ARTICLE INFO

Article history:
Accepted 24 October 2016
Available online 9 November 2016
editor: H. Orland

Keywords:
Random Matrix Theory
High dimensional statistics
Correlation matrix
Spectral decomposition
Rotational invariant estimator

ABSTRACT

This review covers recent results concerning the estimation of large covariance matrices using tools from Random Matrix Theory (RMT). We introduce several RMT methods and analytical techniques, such as the Replica formalism and Free Probability, with an emphasis on the Marčenko–Pastur equation that provides information on the resolvent of multiplicatively corrupted noisy matrices. Special care is devoted to the statistics of the eigenvectors of the empirical correlation matrix, which turn out to be crucial for many applications. We show in particular how these results can be used to build consistent "Rotationally Invariant" estimators (RIE) for large correlation matrices when there is no prior on the structure of the underlying process. The last part of this review is dedicated to some real-world applications within financial markets as a case in point. We establish empirically the efficacy of the RIE framework, which is found to be superior in this case to all previously proposed methods. The case of additively (rather than multiplicatively) corrupted noisy matrices is also dealt with in a special Appendix. Several open problems and interesting technical developments are discussed throughout the paper.

Contents

* Corresponding author at: LPTMS, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay, France.
E-mail addresses: joel.bun@gmail.com (J. Bun), jean-philippe.bouchaud@cfm.fr (J.-P. Bouchaud), marc.potters@cfm.fr (M. Potters).

http://dx.doi.org/10.1016/j.physrep.2016.10.005

1. Introduction

1.1. Motivations

In the present era of "Big Data", new statistical methods are needed to decipher large dimensional datasets 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. 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:

$$\mathbf{C}_{ij} = \mathbb{E}\big[y_i y_j\big], \quad i, j \in [\![1, N]\!]. \tag{1.1}$$

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 (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:

$$E_{ij} := \frac{1}{T} \sum_{t=1}^{T} Y_{it}\, Y_{jt} \equiv \frac{1}{T} \left(\mathbf{Y}\mathbf{Y}^*\right)_{ij}, \tag{1.2}$$

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

[1] This apparently innocuous assumption will be discussed in Section 3.

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 $\mathbf{C}$ [1]. 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 review 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:[2] 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

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.3}$$

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 is 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 [2] for a more detailed discussion):

$$\widehat{\boldsymbol{\beta}} = \left(X^*\mathbf{C}X\right)^{-1} X^*\mathbf{C}^{-1}\mathbf{y}. \tag{1.4}$$

We shall investigate this estimator in Section 7.

(ii) Generalized methods of moments (GMM): Suppose one wants to calibrate the parameters Θ of a model on some dataset. 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 [3]. Note that GLS is a special form of GMM.

(iii) Classification [4]: 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 π_1 and π_2, but same covariance matrix $\mathbf{C}$. The *Linear Discriminant Analysis* rule classifies $\mathbf{y}$ to class 2 if

$$\mathbf{x}^*\mathbf{C}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^*\mathbf{C}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \log(\pi_2/\pi_1). \tag{1.5}$$

(iv) Large portfolio optimization [5]: 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 ν. According to Markowitz's theory, the optimal investment strategy is a vector of weights $\mathbf{w} := (w_1, \ldots, w_p)^*$ 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 μ fixed. (Other constraints can also be implemented.) The optimal strategy reads

$$\mathbf{w} = \nu \frac{\mathbf{C}^{-1}\mathbf{g}}{\mathbf{g}^*\mathbf{C}^{-1}\mathbf{g}}. \tag{1.6}$$

As we shall see in Section 7, 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 [6] and in [7] 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.

1.2. 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

[2] See the monograph [8] for other examples.

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 review 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 [9] 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 $\mathbf{C}$ [9]:

$$\mathcal{P}_W(\mathbf{E}|\mathbf{C}) = \frac{T^{NT/2}}{2^{NT/2}\Gamma_N(T/2)} \frac{\det(\mathbf{E})^{\frac{T-N-1}{2}}}{\det(\mathbf{C})^{T/2}} e^{-\frac{T}{2}\mathrm{Tr}\mathbf{C}^{-1}\mathbf{E}}, \tag{1.7}$$

where $\Gamma_N(\cdot)$ is the multivariate Gamma function with parameter N.[3] 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 [10]:

$$\rho_N(\lambda) = \frac{1}{N}\sum_{k=0}^{N-1} \frac{k!}{T-N+k}\left[L_k^{T-N}(\lambda)\right]^2 \lambda^{T-N} e^{-\lambda}, \tag{1.8}$$

where we assumed that $T > N$ and L_k^l are the Laguerre polynomials.[4]

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 [11]. 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 [12] 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 [13] 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 Section 5). This was first shown for the *precision* matrix $\mathbf{C}^{-1}$ in [14,15] and then for the covariance matrix $\mathbf{C}$ in [16] and lead to the famous *linear shrinkage* estimator

$$\boldsymbol{\Xi} = \alpha_s \mathbf{E} + (1-\alpha_s)\mathbf{I}_N, \tag{1.9}$$

where Ξ denotes, here and henceforth, an estimator of $\mathbf{C}$ and $\alpha_s \in (0, 1)$ is the shrinkage intensity parameter. In [16], Haff proposed to estimate α_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 (1.9) rather than the sample covariance matrix $\mathbf{E}$ has been precisely quantified much later in 2004 [17] in the asymptotic regime $N \to \infty$, with an explicit and observable estimator for the shrinkage intensity α_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 Section 5.

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 [18] 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 [19], 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 Fig. 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 number of simultaneous noisy variables creates important systematic errors in the computation of the eigenvalues of the matrix.

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

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

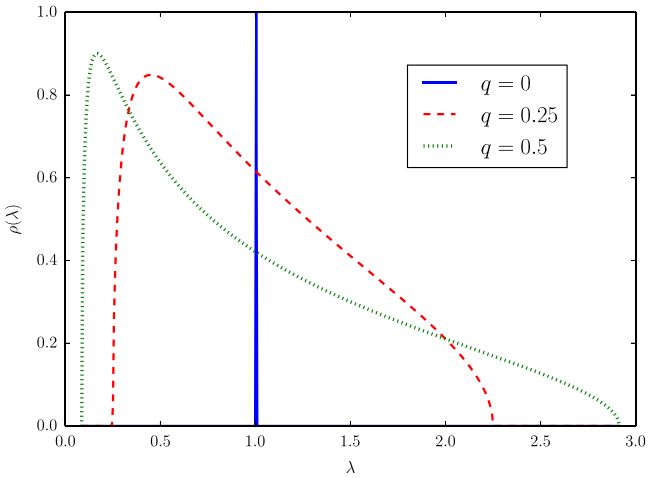


Fig. 1. Plot of the sample eigenvalues and the corresponding sample eigenvalues density under the null hypothesis with $N = 500$. The blue line ($q = 0$) corresponds to a perfect estimation of the population eigenvalues. The larger is the observation ratio q, the wider is the sample density. We see that 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 $\mathbf{C}$ [20–22]. This property is in fact at the core of RMT which makes this theory particularly appealing. At the same time, some empirical evidences of the relevance of these results for sample covariance matrices were provided in [23,24] using financial datasets, which are known to be non-Gaussian [25]. 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) [26]. 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 [27,26]. 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) [28,24]. 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 $\mathbf{C}$ [30]. 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 [29,31–33]. The first one consists in finding a parametric "true" spectral density that fits the data [29]. The method of [31], further improved in [32], 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 [33]. 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 [34,35], 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 [36] 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 [36]. 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

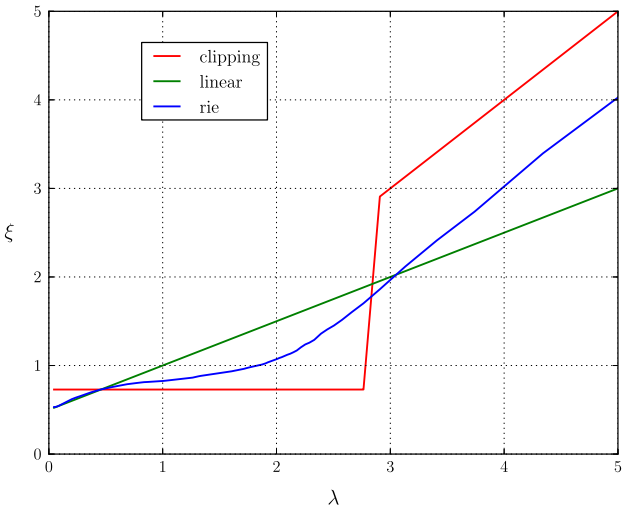


Fig. 2. Three shrinkage transformations: "cleaned" eigenvalues on the y-axis as a function of the sample eigenvalues (see Section 8 for more details). This figure is a quick summary the evolution of shrinkage estimators starting with the linear method (green), then the heuristic eigenvalues clipping method (red) to the optimal RIE (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

covariance matrices $\mathbf{C}$ [37–41]. 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*[5] (RIE) [37–39]. In this particular class of estimators, the main assumption is that any estimator Ξ 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 (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 $\mathbf{C}$ [39]. 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 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 Fig. 2, which can be seen as a thumbnail picture of the present review. Note that a very recent work [41] 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 allow one to construct "optimal" non-rotational invariant estimators. We shall address this issue at the end of this review.

1.3. Outline

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 [18] to the very recent "local" optimal RIE for general population covariance matrices [39]. We emphasize that this review 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 Section 2, 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 asymptotic regime. In fact, most of the computations in Section 2 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 [42]. 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 [43] or Marčenko–Pastur density [18]. This is the main purpose of Section 2.2. 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* [44]. Loosely speaking, two matrices $\mathbf{A}$ and $\mathbf{B}$ are mutually free if their eigenbasis are related to one another by a random rotation, or said differently if the eigenvectors of $\mathbf{A}$

[5] This is sometimes called rotation-equivariant estimators.

and **B** are almost surely orthogonal. Voiculescu discovered in 1991 [45] that some random matrices do satisfy asymptotically the freeness relation, which considerably influenced RMT. We present in Section 2.3 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 2.4, we present a more formal tool known as the Replica method in statistical physics of disordered systems [46,47]. While being less rigorous, this method turns out to be very powerful to compute the average behavior of large complex systems (see [48] 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.

In Sections 3 and 4, we study in details the different properties of large sample covariance matrices. Section 3 is dedicated to the statistics of the eigenvalues of **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 **C** using **E** such as the moment generating functions, or the edges of the support of the spectral density of **E**. We discuss the properties of the edges of the distribution for finite N and also the outliers. In Section 4, we focus the recent results concerning the eigenvectors of **E** for a general **C**. We distinguish two different cases. The first one is the angle between the true and estimated eigenvectors and we shall see that the initial results of [36] hold for a general **C**. The second case is the angle between two *independent* sample eigenvectors, a result that allows one to infer interesting properties about the structure of **C**.

After these three relatively technical sections, we then turn on the main theme of this review which is the estimation of large sample covariance matrices. In Section 5, we formalize the Bayesian method for covariance matrices. We present the class of conjugate prior from which we re-obtain the linear shrinkage (1.9) initially derived by Haff [16]. Next, we consider the class of Boltzmann-type, rotational invariant prior distributions. We then relate the Bayes optimal estimator with the least squares optimal oracle estimator of **C**. The so-called oracle estimator is the main quantity of interest in the following Section 6. 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 Section 4. Hence, there exists an optimal estimator of large population covariance **C** depending only on **E** inside the class of RIEs. The rest of Section 6 is dedicated to some theoretical and numerical applications of the optimal RIE.

Section 7 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 **E** in a large scale and out-of-sample framework. As alluded to above, we shall see that if **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 Section 6. Several alternative cleaning "recipes", proposed in previous work, are also reviewed in that section.

Finally, Section 8 contains empirical results using real financial datasets. We give further evidence that using a correctly regularized estimator of **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.

The appendices contain auxiliary results which are mentioned in the paper. 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. 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 Section 2 but is now standard in the RMT literature, which is why we relegate its presentation to an appendix. Finally, we devote a full appendix to the case where the noise in the matrix is *additive*, rather than *multiplicative* for correlation matrices. Although not directly relevant to the main issue discussed in the present review, the additive noise model is interesting in itself and finds many applications in different fields of science.

2. Random Matrix Theory: overview and analytical tools

2.1. RMT in a nutshell

2.1.1. 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 1950s as a statistical model for the energy levels of heavy nuclei, and leads to the well-known Wigner semi-circle distribution, as well as Dyson's Brownian motion (see e.g. [49,50] 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 [51,52]. Here, we will only consider the results of RMT that pertain to statistical inference, and leave aside

many topics—see e.g. [50,53–56] 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 [58] and can be dealt with the Singular Value Decomposition, for which partial results are available, see e.g. [59,60].

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[6] 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[7] $\mathbf{M}$ with a certain probability measure $\mathcal{P}_\beta(\mathbf{M})$, where β 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 [61] 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 $\mathbf{M} \to \boldsymbol{\Omega}\mathbf{M}\boldsymbol{\Omega}^\dagger$ for any matrix $\boldsymbol{\Omega}$ belonging to the Orthogonal group $\mathbf{O}(N)$, i.e. $\mathcal{P}_\beta(\mathbf{M}) = \mathcal{P}_\beta(\boldsymbol{\Omega}\mathbf{M}\boldsymbol{\Omega}^\dagger)$, $\forall \boldsymbol{\Omega} \in \mathbf{O}(N)$. 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:

$$\mathcal{P}_\beta(\mathbf{M})\mathcal{D}\mathbf{M} \propto e^{-\frac{\beta N}{2}\mathrm{Tr}V(\mathbf{M})}\mathcal{D}\mathbf{M} \tag{2.1}$$

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 (2.1) can alternatively be rewritten in terms of the eigenvalues and eigenvectors of $\mathbf{M}$ as:

$$\mathcal{P}_\beta(\mathbf{M})\mathcal{D}\mathbf{M} \propto e^{-\frac{\beta N}{2}\sum_{i=1}^N V(\nu_i)}\prod_{i<j}^N |\nu_i - \nu_j|^\beta \Big(\prod_{i=1}^N \mathrm{d}\nu_i\Big)\Big(\mathrm{d}\Omega\Big), \tag{2.2}$$

where the Vandermonde determinant ($\prod |\nu_i - \nu_j|^\beta$)comes from the change of variables (from the $\mathbf{M}_{ij}$ to the ν_i and Ω_{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 ν_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"):

$$\rho_{\mathbf{M}}^N(x) = \frac{1}{N}\sum_{i=1}^N \delta(x - \nu_i) \tag{2.3}$$

with δ 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 review, but see [56,57,62] 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

6 i.e. independent of the specific realization of the matrix itself.

7 Boldface letters will refer throughout this paper to matrices.

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 (2.3) for a specific $\mathbf{M}$ by the average according to the probability measure of $\mathbf{M}$ (e.g. over the measure (2.1)):

$$\rho_{\mathbf{M}}(x) = \lim_{N\to\infty} \rho^N_{\mathbf{M}}(x), \quad \text{with} \quad \rho^N_{\mathbf{M}}(x) = \left\langle \frac{1}{N}\sum_{i=1}^{N} \delta(x-\nu_i) \right\rangle_{\mathbf{M}}. \tag{2.4}$$

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 a 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 section, 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.

2.1.2. 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[8]

$$\mathbf{G}_{\mathbf{M}}(z) := (z\mathbf{I}_N - \mathbf{M})^{-1}, \tag{2.5}$$

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:

$$\mathbf{G}_{\mathbf{M}}(z) = \sum_{i=1}^{N} \frac{\mathbf{w}_i \mathbf{w}_i^*}{z - \nu_i}. \tag{2.6}$$

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 ν_i. We will show in Section 4 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 (2.4). For this aim, we define the normalized trace of Eq. (2.5) as

$$\mathfrak{g}^N_{\mathbf{M}}(z) := \frac{1}{N}\mathrm{Tr}\left[\mathbf{G}_{\mathbf{M}}(z)\right]. \tag{2.7}$$

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

$$\mathfrak{g}^N(z) \underset{N\to\infty}{\sim} \mathfrak{g}(z), \qquad \mathfrak{g}(z) := \int \frac{\rho(u)}{z-u}\mathrm{d}u \tag{2.8}$$

which is known as the *Stieltjes* (or *Cauchy*) transform of ρ. The Stieltjes transform has a lot of appealing properties. For instance, if the density function ρ 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 $z \in \mathbb{R}\setminus \mathrm{supp}[\rho_{\mathbf{M}}]$;
(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.:

$$\lim_{\eta\to 0^+} \mathfrak{g}(x \pm \mathrm{i}\eta) = \mathfrak{h}(x) \mp i\pi\rho(x), \qquad x \in \mathrm{supp}[\rho] \text{ and } \rho(x) \in \mathbb{R}^+, \tag{2.9}$$

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

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

$$\mathfrak{h}(x) := \fint_{\mathrm{supp}[\rho]} \frac{\rho(u)}{x-u}\mathrm{d}u \tag{2.10}$$

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 ρ can be retrieved by inverting the last property of the Riemann–Hilbert problem:

$$\rho(x) \equiv \frac{1}{\pi}\lim_{\eta\to 0^+}\mathrm{Im}(\mathfrak{g}(x-\mathrm{i}\eta)), \qquad x\in\mathrm{supp}[\rho]. \tag{2.11}$$

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

$$\mathfrak{g}(z) \underset{z\to\infty}{=} \frac{1}{z}+\frac{1}{N}\sum_{k=1}^{\infty}\frac{\mathrm{Tr}\mathbf{M}^k}{z^{k+1}} \equiv \frac{1}{z}+\sum_{k=1}^{\infty}\frac{\varphi(\mathbf{M}^k)}{z^{k+1}}, \tag{2.12}$$

where we defined the kth 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 to 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

$$\mathfrak{g}_{a\mathbf{M}}(z) = \frac{1}{a}\mathfrak{g}_{\mathbf{M}}\left(\frac{z}{a}\right), \tag{2.13}$$

for any $a\in\mathbb{R}\backslash\{0\}$. Moreover, suppose that $\mathbf{M}$ is invertible, then using (2.7) we also have

$$z\mathfrak{g}_{\mathbf{M}}(z)+\frac{1}{z}\mathfrak{g}_{\mathbf{M}^{-1}}\left(\frac{1}{z}\right) = 1, \tag{2.14}$$

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 of them will turn out to be important in the next section. We begin with the *free cumulant* generating function which is known as the $\mathcal{R}$-transform in the literature [63,55,64]. To define this quantity, it is convenient to introduce the functional inverse of the Stieltjes transform, also known as the *Blue* transform [65]

$$\mathcal{B}(\mathfrak{g}(z)) = z, \tag{2.15}$$

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

$$\mathcal{R}(\omega) = \mathcal{B}(\omega)-\frac{1}{\omega}. \tag{2.16}$$

Note that one may deduce from (2.13) the following property

$$\mathcal{R}_{a\mathbf{M}}(\omega) = a\mathcal{R}_{\mathbf{M}}(a\omega), \tag{2.17}$$

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. (2.16), we obtain the formula

$$\mathcal{R}(\mathfrak{g}(z))+\frac{1}{\mathfrak{g}(z)} = z. \tag{2.18}$$

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

$$\mathcal{R}(\omega) = \sum_{\ell=1}^{\infty} \kappa_\ell(\mathbf{M})\omega^{\ell-1} \tag{2.19}$$

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

$$\begin{aligned}
\kappa_1 &= \varphi_1 \\
\kappa_2 &= \varphi_2 - \varphi_1^2 \\
\kappa_3 &= \varphi_3 - 3\varphi_2\varphi_1 + 2\varphi_1^3 \\
\kappa_4 &= \varphi_4 - 4\varphi_3\varphi_1 - 2\varphi_2^2 + 10\varphi_2\varphi_1^2 - 5\varphi_1^4.
\end{aligned} \tag{2.20}$$

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 2.3.

Moment generating function and S-transform. The moment generating function of the LSD ρ is obtained by considering

$$\mathcal{T}(z) := z\mathfrak{g}(z) - 1 = \int \frac{\mathrm{d}u\rho(u)u}{z-u}, \tag{2.21}$$

frequently known as the $\mathcal{T}$ (or sometimes η [55]) transform [40]. Indeed, by taking $z \to \infty$, one readily finds

$$\mathcal{T}_{\mathbf{M}}(z) = \sum_{k=1}^{\infty} \frac{\varphi(\mathbf{M}^k)}{z^k}. \tag{2.22}$$

We can then introduce the so-called $\mathcal{S}$-transform as [63]:

$$\mathcal{S}(\omega) := \frac{\omega+1}{\omega\mathcal{T}^{-1}(\omega)} \tag{2.23}$$

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. (2.20), one finds that the $\mathcal{S}$-transform also admits a Taylor series which reads:

$$\begin{aligned}
\mathcal{S}_{\mathbf{M}}(\omega) &= \frac{1}{\varphi_1} + \frac{\omega}{\varphi_1^3}(\varphi_1^2 - \varphi_2) + \frac{\omega^2}{\varphi_1^5}(2\varphi_2^2 - \varphi_2\varphi_1^2 - \varphi_3\varphi_1) + \mathcal{O}(\omega^3) \\
&= \frac{1}{\kappa_1} - \frac{\kappa_2}{\kappa_1^3}\omega + \frac{2\kappa_2^2 - \kappa_1\kappa_3}{\kappa_1^5}\omega^2 + \mathcal{O}(\omega^3).
\end{aligned} \tag{2.24}$$

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 2.3.3)

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

$$\mathcal{R}(\omega) = \frac{1}{\mathcal{S}(\omega\mathcal{R}(\omega))}, \qquad \mathcal{S}(\omega) = \frac{1}{\mathcal{R}(\omega\mathcal{S}(\omega))} \tag{2.25}$$

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 [66].

Let us show the second equality of (2.25) for the sake of completeness. The derivation of the first identity is similar and we omit details. Using (2.16) and (2.23), one obtains

$$\mathcal{R}(\omega\mathcal{S}(\omega)) = \mathcal{B}\left(\frac{\omega+1}{\mathcal{T}^{-1}(\omega)}\right) - \frac{\mathcal{T}^{-1}(\omega)}{\omega+1}. \tag{2.26}$$

Next, by setting $z = \mathcal{T}^{-1}(\omega)$, we can rewrite (2.21) as

$$\frac{\omega+1}{\mathcal{T}^{-1}(\omega)} = \mathfrak{g}\big(\mathcal{T}^{-1}(\omega)\big). \tag{2.27}$$

Hence, we conclude that

$$\mathcal{R}(\omega\mathcal{S}(\omega)) = \mathcal{T}^{-1}(\omega) - \frac{1}{\mathfrak{g}\big(\mathcal{T}^{-1}(\omega)\big)} = \frac{\omega}{\mathfrak{g}\big(\mathcal{T}^{-1}(\omega)\big)}. \tag{2.28}$$

The conclusion then follows from (2.27).

2.2. 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 C and D.1.2. We refer to [53] for the moment methods (ii), to [42,67] for Feynman diagrams (iii) or to [68] and references therein for symmetry 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 [50,53,54,57] 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 [69] for a self-contained introduction or to e.g. [42,51,70] 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. (2.1).

2.2.1. Stieltjes transform and potential function

First, we write from (2.1) 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 [42] and we repeat here the main idea of the derivation (see also [71, Section 2.1]). Let us first express the partition function in terms of the eigenvalues and eigenvectors of $\mathbf{M}$, using (2.2):

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

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:

$$\mathcal{Z} \propto \int \prod_{i=1}^{N} d\nu_i e^{-N^2 S(\{\nu_i\})} \quad \text{with} \quad S(\{\nu_i\}) = \frac{1}{N}\sum_{i=1}^{N} V(\nu_i) - \frac{\beta}{2N^2}\sum_{i\neq j}\log|\nu_i-\nu_j|. \tag{2.29}$$

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 Fig. 3. We recall that we consider only densities which are defined on a 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" ℓ 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 $\ell = 0$.

In the large N limit, the integral over eigenvalues can be computed by the saddle-point method which yields the following "force equilibrium" condition[9]:

$$V'(\nu_i) = \frac{\beta}{N}\sum_{j=1:j\neq i}^{N}\frac{1}{\nu_i-\nu_j}, \quad \forall i = 1,\ldots,N. \tag{2.30}$$

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 [42]:

$$\mathfrak{g}(z) = V'(z) - Q(z)\sqrt{(z-\nu_+)}\sqrt{(z-\nu_-)}, \tag{2.31}$$

[9] 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. [70] for a detailed discussion. Entropy effects start playing a role for extended β ensembles where $\beta = c/N$ where c is finite, see [72].

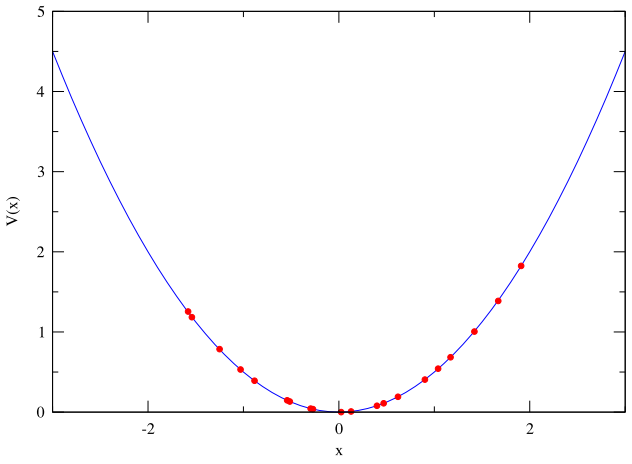


Fig. 3. Typical configuration of a repulsive Coulomb gas with $N = 20$ particles (red dots) in the potential $V(x) = x^2/2$ as a function of x.

where $\nu_- < \nu_+$ denote the edges of supp$[\rho]$ and $Q(z)$ is also a Laurent polynomial with degree $d-1$ and order ℓ. Therefore, we see that we have $d+1$ unknowns to determine, namely the coefficients of $Q(z)$, ν_- and ν_+ which are determined using the series expansion (2.12). We shall give a detailed illustration of this procedure in Section 2.2.3.[10]

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 (2.31). In the following we set $\beta = 1$. First, we introduce the normalized trace of the resolvent $\mathfrak{g}(z)$ in (2.30) by multiplying on both sides by $N^{-1}(z-\nu_i)^{-1}$ and summing over all i, which yields

$$\frac{1}{N}\sum_{i=1}^{N}\frac{V'(\nu_i)}{z-\nu_i} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1;j\neq i}^{N}\frac{1}{(z-\nu_i)(\nu_i-\nu_j)}. \tag{2.32}$$

Notice that this last equation is indeed an analytical function for $z \in \mathbb{C}\backslash\text{Supp}[\rho_{\mathbf{M}}]$. 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 (2.32) 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

$$\mathfrak{g}^2(z) - 2V'(z)\mathfrak{g}(z) + \frac{2}{N}\sum_{i=1}^{N}\frac{V'(z)-V'(\nu_i)}{z-\nu_i} = 0. \tag{2.33}$$

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

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

[10] In the case of positive definite covariance matrices, we can use the series (3.23) that corresponds to the limit $z \to 0$.

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. (2.33) 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 ρ) 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 ν_- and ν_+ denote the edges of supp$[\rho]$ and $Q(z)$ is a polynomial with degree $d-1$ and this gives (2.31).

2.2.2. Wigner's semicircle law

As a warm-up exercise, we begin with Wigner's semi-circle law [43], 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. [53], 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[11]:

$$\mathbb{E}\big[\mathbf{M}_{ij}\big] = 0, \quad \text{and} \quad \mathbb{E}\big[\mathbf{M}_{ij}^2\big] = \sigma^2/N. \tag{2.34}$$

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. (2.31), 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:

$$\mathfrak{g}_W(z) = \frac{z - \sqrt{z+2\sigma}\sqrt{z-2\sigma}}{2\sigma^2}, \tag{2.35}$$

where $\sqrt{\cdot}$ denotes throughout the following the principal square root, that is the non-negative square root of a non-negative real number. Eq. (2.35) 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 then retrieved using the inversion formula (2.11) that yields the celebrated *Wigner's semicircle law*:

$$\rho_W(x) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - x^2}, \qquad |x| < 2\sigma. \tag{2.36}$$

We plot in Fig. 4 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 [76]).

Due to the relative simplicity of the expression of Eq. (2.35), one can easily invert this expression to find the Blue transform to find that the $\mathcal{R}$-transform of the semicircle law reads

$$\mathcal{R}_W(z) = \sigma^2 z. \tag{2.37}$$

Since the average trace φ_1 is exactly 0, the $\mathcal{S}$-transform of a Wigner matrix is an ill-defined object.

2.2.3. The Marčenko–Pastur law

As stated in the introduction, the study of random matrices began with John Wishart [9]. 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

[11] The case where the variance of the matrix elements diverge corresponds to *Lévy matrices*, introduced in [73]. For a rigorous approach, we refer the readers to [74]. For recent developments, see [75].

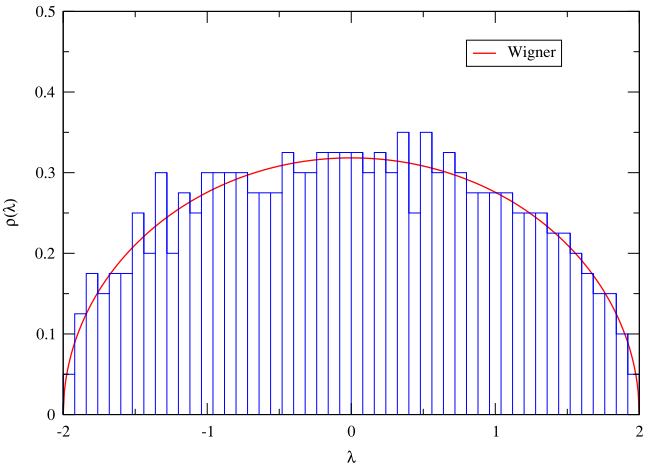


Fig. 4. Wigner semi-circle density (2.36) compared with empirical results with $N = 500$ (histogram) from one sample, illustrating the convergence of the ESD at finite N to the asymptotic LSD.

$\mathbf{C}$, 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 Section 3). For *any* N and $T > N$, Wishart derived the exact PDF of the entries $\mathbf{M}$ which reads:

$$\mathcal{P}_{\mathrm{w}}(\mathbf{M}|\mathbf{C}) = \frac{1}{2^{NT/2}\Gamma_N(T/2)} \frac{\det(\mathbf{M})^{\frac{T-N-1}{2}}}{\det(\mathbf{C})^{T/2}} e^{-\frac{T}{2}\mathrm{Tr}\mathbf{C}^{-1}\mathbf{M}}. \tag{2.38}$$

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 (2.38)

$$\mathcal{P}_{\mathrm{w}}(\mathbf{M}|\mathbf{I}_N) \propto \det(\mathbf{M})^{\frac{T-N-1}{2}} e^{-\frac{T}{2}\mathrm{Tr}\mathbf{M}} := e^{-\frac{T}{2}\mathrm{Tr}\mathbf{M} + \frac{T-N-1}{2}\mathrm{Tr}\log\mathbf{M}}, \tag{2.39}$$

which clearly belongs to the class of Boltzmann ensembles (2.1). Throughout the following, we shall denote by $\mathcal{W}$ the $N \times N$ matrix whose distribution is given by (2.39). Ignoring sub-leading terms, the corresponding potential function is given by:

$$V(z) = \frac{1}{2q}\left[z - (1-q)\log z\right], \qquad \text{with} \qquad q := N/T. \tag{2.40}$$

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 (2.32) is of the form c/z with c a constant to be determined using (2.12). We postpone the computation of the Stieltjes transform $\mathfrak{g}(z)$ to the end of this section. The final result reads:

$$\mathfrak{g}(z) = \frac{(z+q-1) - \sqrt{z-\nu_-}\sqrt{z-\nu_+}}{2qz}, \qquad \nu_\pm := (1 \pm \sqrt{q})^2, \tag{2.41}$$

and this is the solution found by Marčenko and Pastur in [18] in the special case $\mathbf{C} = \mathbf{I}_N$. We can now use the inversion formula (2.11) to find the celebrated Marčenko–Pastur (MP) law (for $q \in (0, 1)$)

$$\rho_{\mathrm{MP}}(\nu) = \frac{\sqrt{4\nu q - (\nu + q - 1)^2}}{2q\pi\nu}, \qquad \forall\, \nu \in \left[\nu_-, \nu_+\right]. \tag{2.42}$$

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. (2.42). 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 [77]).

Again, the expression of $\mathfrak{g}(z)$ is simple enough to obtain a closed formula for the Blue transform, and deduce from Eq. (2.41) the $\mathcal{R}$-transform of the MP law:

$$\mathcal{R}_{\mathrm{MP}}(\omega) = \frac{1}{1 - q\omega}. \tag{2.43}$$

One can compute the $\mathcal{S}$-transform of the MP law using the relation (2.25):

$$\mathcal{S}_{\mathrm{MP}}(\omega) = \frac{1}{1+q\omega}. \tag{2.44}$$

We now derive the Stieltjes transform (2.41) through a complete application of the BIPZ formalism introduced in Eq. (2.32). As alluded to above, the Stieltjes transform (2.32) for the isotropic Wishart matrix has the form

$$\mathfrak{g}(z) = \frac{1}{2q}\left[1 - \frac{1-q}{z}\right] - \frac{c}{z}\sqrt{z-\nu_+}\sqrt{z-\nu_-}, \tag{2.45}$$

and the constants that we have to determine are c, ν_+ and ν_-. To that end, we use (2.12) that tells us that when $|z| \to \infty$

$$\mathfrak{g}(z) = \frac{1}{z} + \frac{\varphi(\mathbf{M})}{z^2} + \mathcal{O}(z^{-3}). \tag{2.46}$$

On the other hand, one finds by taking the limit $z \to \infty$ into (2.45) that

$$\mathfrak{g}(z) = \frac{1}{2q}\left[1 - \frac{1-q}{z}\right] - c\left[1 - \frac{\nu_+ + \nu_-}{2z} - \frac{(\nu_+ - \nu_-)^2}{8z^2}\right] + \mathcal{O}(z^{-3}). \tag{2.47}$$

Then, by comparing this last equation to (2.46), 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

$$c = \frac{1}{2q}. \tag{2.48}$$

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

$$1 = -\frac{(1-q)}{2q} + \frac{\nu_+ + \nu_-}{4q}, \tag{2.49}$$

that is to say

$$\nu_+ = 2(1+q) - \nu_-. \tag{2.50}$$

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

$$\varphi(\mathbf{M}) = \frac{(\nu_+ - \nu_-)^2}{16q}, \tag{2.51}$$

which is equivalent to

$$\nu_- = \nu_+ - 4\sqrt{q\varphi(\mathbf{M})} = (1+q) - 2\sqrt{q} = (1-\sqrt{q})^2, \tag{2.52}$$

where we used (2.50) and $\varphi(\mathbf{M}) = 1$ in the third step. Consequently, we deduce from (2.50) that $\nu_+ = (1+\sqrt{q})^2$ and the result (2.41) follows from the Eqs. (2.48), (2.50) and (2.52).

2.2.4. *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 (2.42). Indeed, one just needs to make the change of variable $u = ((1-q)v)^{-1}$ into Eq. (2.42) to obtain[12]:

$$\rho_{\mathrm{IMP}}(u) = \frac{\kappa}{\pi u^2}\sqrt{(u_+ - u)(u - u_-)}, \qquad u_\pm := \frac{1}{\kappa}\left[\kappa + 1 \pm \sqrt{2\kappa+1}\right], \tag{2.53}$$

where the subscript IMP stands for "Inverse Marčenko–Pastur" and κ is related to q through

$$q = \frac{1}{2\kappa+1} \in (0,1). \tag{2.54}$$

In particular, one notices that $u_\pm = (1-q)/v_\mp$ where $v_\mp$ is defined in Eq. (2.41). We plot in Fig. 5 the density of the Marčenko–Pastur (2.42) and of its inverse (2.53) both with parameter $q = 0.5$.

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

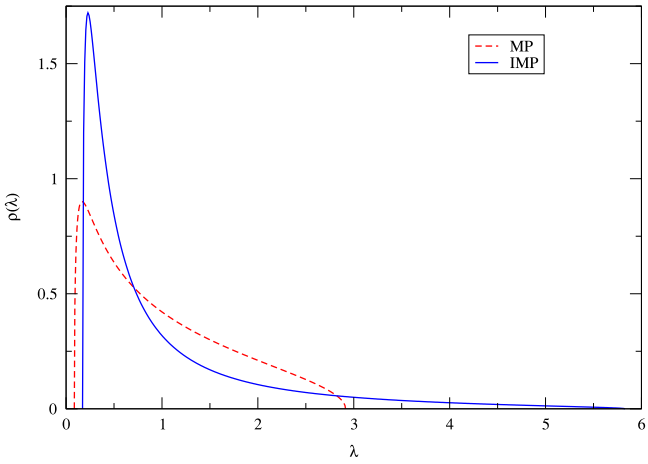


Fig. 5. The red dotted curve corresponds to the Marčenko–Pastur density (2.42) with $q = 0.5$. We repeat the experiment with the Inverse Wishart matrix still with $q = 0.5$ (plain blue curve).

In addition to the eigenvalue density (2.53), one can also derive explicit expressions for the other transforms presented in Section 2.1.2. For the Stieltjes transform, it suffices to apply the same change of variable $u = ((1-q)z)^{-1}$ and to use the properties (2.13) and (2.14) to obtain:

$$\mathfrak{g}_{\text{iw}}(u) = \frac{u(\kappa+1) - \kappa - \kappa\sqrt{u-u_-}\sqrt{u-u_+}}{u^2}, \tag{2.55}$$

where the bounds $u_\pm$ are given in Eq. (2.53). One can easily check with the inversion formula (2.9) that we indeed retrieve the density of states (2.53) as expected.

Using the Stieltjes transform (2.55), one can then compute the $\mathcal{R}$-transform of the Inverse Marčenko–Pastur density to find

$$\mathcal{R}_{\text{IMP}}(\omega) = \frac{\kappa - \sqrt{\kappa(\kappa-2\omega)}}{\omega}, \qquad \kappa > 0, \tag{2.56}$$

and then, from (2.25), the $\mathcal{S}$-transform reads

$$\mathcal{S}_{\text{IMP}}(\omega) = 1 - \frac{\omega}{2\kappa}. \tag{2.57}$$

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 [10]:

$$\mathcal{P}_{\text{iw}}(\mathbf{M}^{-1}|\mathbf{C}) = \frac{1}{2^{NT/2}\Gamma_N(T/2)} \frac{\det(\mathbf{C})^{T/2}}{\det(\mathbf{M})^{((T+N+1)/2)}} e^{-\frac{1}{2}\operatorname{Tr}\mathbf{C}\mathbf{M}^{-1}}. \tag{2.58}$$

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. (2.2). A detailed derivation of this change of variable may be found e.g. in [78, Eq. (15.15)].

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

$$\langle \mathbf{M} \rangle_{\mathcal{P}_{\text{iw}}} = \frac{\mathbf{C}}{T-N-1}. \tag{2.59}$$

The derivation of this result can be obtained using the different identities of [79].

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 (2.59) that

$$\langle \mathbf{M} \rangle_{\mathcal{P}_{\text{iw}}} = \frac{T}{T-N-1}\mathbf{I}_N \underset{\text{LDL}}{\sim} \frac{1}{1-q}\mathbf{I}_N. \tag{2.60}$$

In order to have a normalized spectral density, i.e. $N^{-1}\operatorname{Tr}\mathbf{M} = 1$, 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. (2.53).

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.

2.3. 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. [64] or [66] for a more exhaustive presentation.

2.3.1. Freeness

Free probability theory was initiated in 1985 by Dan Voiculescu in order to understand special classes of von Neumann algebras [44], 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 [63] and Speicher [80] 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 [63]: 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 [80].

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

$$\varphi(\mathbf{M}) := \frac{1}{N}\mathrm{Tr}\mathbf{M}, \tag{2.61}$$

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:

$$\varphi\left(\mathbf{A}^{n_1}\mathbf{B}^{m_1}\mathbf{A}^{n_2}\mathbf{B}^{m_2}\dots\mathbf{A}^{n_k}\mathbf{B}^{m_k}\right) = \varphi(\mathbf{A}^{n_1})\varphi(\mathbf{B}^{m_1})\varphi(\mathbf{A}^{n_2})\varphi(\mathbf{B}^{m_2})\dots\varphi(\mathbf{A}^{n_k})\varphi(\mathbf{B}^{m_k}), \tag{2.62}$$

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

$$\varphi\left((\mathbf{A} - \varphi(\mathbf{A}))(\mathbf{B} - \varphi(\mathbf{B}))\right) = 0, \tag{2.63}$$

from which we deduce $\varphi(\mathbf{AB}) = \varphi(\mathbf{A})\varphi(\mathbf{B})$. Hence, if one thinks of the trace operator (2.61) 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

$$\varphi\left((\mathbf{A} - \varphi(\mathbf{A}))(\mathbf{B} - \varphi(\mathbf{B}))(\mathbf{A} - \varphi(\mathbf{A}))\right) = 0, \tag{2.64}$$

we can deduce that

$$\varphi(\mathbf{ABA}) = \varphi(\mathbf{A}^2\mathbf{B}) = \varphi(\mathbf{A}^2)\varphi(\mathbf{B}). \tag{2.65}$$

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 Ω 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 [81]. As we shall see in Section 7, the computation of mixed moments will be used to derive some useful relations for estimating over-fitting for statistical estimation problems.

2.3.2. Sums of free matrices

In addition to the computation of mixed moments such as Eq. (2.64), 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.,

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

for any $N \times N$ matrix $\boldsymbol{\Omega}$ 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*

$$\mathcal{R}_{\mathbf{M}}(\omega) = \mathcal{R}_{\mathbf{A}}(\omega) + \mathcal{R}_{\mathbf{B}}(\omega). \tag{2.66}$$

Hence, we can interpret the $\mathcal{R}$-transform (2.16) as the analogue in RMT of the logarithm of the Fourier transform for standard additive convolution. It is possible to rewrite Eq. (2.66) as a function of the Stieltjes transform of $\mathbf{M}$ that contains all the information about the spectral density of $\mathbf{M}$. Eq. (2.66) 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

$$\mathfrak{g}_{\mathbf{M}}(z) = \mathfrak{g}_{\mathbf{A}}(z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z))). \tag{2.67}$$

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 σ^2/N. We know from Eq. (2.37) 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[13]:

$$\mathfrak{g}_{\mathbf{M}}(z) = \mathfrak{g}_{\mathbf{A}}(z - \sigma_{\mathbf{B}}^2 \mathfrak{g}_{\mathbf{M}}(z)). \tag{2.68}$$

This is the Stieltjes transform of the deformed GOE matrix[14] 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 $\mathbf{B}$ [83]. 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 $z \in \mathbb{R} \setminus \operatorname{supp}[\rho_{\mathbf{A}}]$ such that

$$\mathfrak{g}_{\mathbf{A}}(z) = \frac{1}{\beta}. \tag{2.69}$$

For instance, if $\mathbf{A}$ is a Wigner matrix with variance $\sigma^2 > 0$, one can easily check from (2.69) and (2.37) that the largest eigenvalue ν_1 of $\mathbf{M}$ is given by

$$\nu_1 = \begin{cases} \beta + \sigma^2/\beta & \text{if } \beta > \sigma \\ 2\sigma & \text{otherwise.} \end{cases} \tag{2.70}$$

When $\beta > \sigma$, we say that ν_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 (2.69). First we need to compute the $\mathcal{R}$-transform of the rank one matrix $\mathbf{B}$ in order to use (2.66). From (2.8), we easily find that

$$\mathfrak{g}_{\mathbf{B}}(u) = \frac{1}{N}\frac{1}{u-\beta} + \left(1 - \frac{1}{N}\right)\frac{1}{u} = \frac{1}{u}\left[1 + \frac{1}{N}\frac{\beta}{1-u^{-1}\beta}\right]. \tag{2.71}$$

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

$$\mathcal{B}_{\mathbf{B}}(\omega) = \frac{1}{\omega} + \frac{\beta}{N(1-\omega\beta)} + \mathcal{O}(N^{-2}). \tag{2.72}$$

[13] 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 D for more about this.

[14] This result can be generalized to the class of deformed Wigner matrices, i.e. where the noise is given by (2.34) but not necessarily Gaussian, see e.g. [82].

We may therefore conclude from (2.16) that

$$\mathcal{R}_{\mathbf{B}}(\omega) = \frac{\beta}{N(1-\beta\omega)} + \mathcal{O}(N^{-2}). \tag{2.73}$$

Hence, we obtain by applying (2.66) and (2.16) that

$$\mathcal{B}_{\mathbf{M}}(\omega) = \mathcal{B}_{\mathbf{A}}(\omega) + \frac{\beta}{N(1-\beta\omega)} + \mathcal{O}(N^{-2}). \tag{2.74}$$

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

$$z = \mathcal{B}_{\mathbf{A}}(\mathfrak{g}_{\mathbf{M}}(z)) + \frac{\beta}{N(1-\beta\mathfrak{g}_{\mathbf{M}}(z))} + \mathcal{O}(N^{-2}). \tag{2.75}$$

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

$$\mathfrak{g}_{\mathbf{M}}(z) = \mathfrak{g}_0(z) + \frac{\mathfrak{g}_1(z)}{N} + \mathcal{O}(N^{-2}). \tag{2.76}$$

By plugging this ansatz into (2.75), we see that $\mathfrak{g}_0(z)$ and $\mathfrak{g}_1(z)$ satisfy

$$\begin{aligned} z &= \mathcal{B}_{\mathbf{A}}(\mathfrak{g}_0(z)) \\ \mathfrak{g}_1(z) &= -\frac{\beta}{\mathcal{B}'_{\mathbf{A}}(\mathfrak{g}_0(z))(1-\mathfrak{g}_0(z)\beta)}. \end{aligned} \tag{2.77}$$

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

$$\mathfrak{g}_1(z) = -\frac{\beta\mathfrak{g}'_{\mathbf{A}}(z)}{1-\mathfrak{g}_{\mathbf{A}}(z)\beta}. \tag{2.78}$$

Finally, we obtained that

$$\mathfrak{g}_{\mathbf{M}}(z) \approx \mathfrak{g}_{\mathbf{A}}(z) - \frac{1}{N}\frac{\beta\mathfrak{g}'_{\mathbf{A}}(z)}{1-\mathfrak{g}_{\mathbf{A}}(z)\beta}, \tag{2.79}$$

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 $z \in \mathbb{R}\setminus \mathrm{supp}[\rho_{\mathbf{A}}]$ such that $\mathfrak{g}_{\mathbf{A}}(z) = \beta^{-1}$ and this leads to the criterion (2.69).

2.3.3. *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

$$\mathbf{M} := \sqrt{\mathbf{A}}\mathbf{B}\sqrt{\mathbf{A}}. \tag{2.80}$$

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

$$\mathcal{S}_{\mathbf{M}}(\omega) = \mathcal{S}_{\mathbf{A}}(\omega)\mathcal{S}_{\mathbf{B}}(\omega). \tag{2.81}$$

This is the so-called *free multiplication*, which has been first obtained by Voiculescu [63] and then by [84] in a physics formalism.

Again, if one is interested in the limiting spectral density of $\mathbf{M}$, one would like to write (2.81) in terms of its Stieltjes transform. Using the very definition of the $\mathcal{S}$-transform, we rewrite (2.81) as

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

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

$$\mathcal{T}^{-1}_{\mathbf{A}}(\mathcal{T}_{\mathbf{M}}(z)) = z\mathcal{S}_{\mathbf{B}}(\mathcal{T}_{\mathbf{M}}(z)). \tag{2.82}$$

It is now immediate to get the analogue of (2.67) for the multiplicative case

$$\mathcal{T}_{\mathbf{M}}(z) = \mathcal{T}_{\mathbf{A}}\left(z\mathcal{S}_{\mathbf{B}}(\mathcal{T}_{\mathbf{M}}(z))\right), \tag{2.83}$$

that gives in terms of the Stieltjes transform

$$z\mathfrak{g}_{\mathbf{M}}(z) = Z(z)\mathfrak{g}_{\mathbf{A}}(Z(z)), \qquad Z(z) := z\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1). \tag{2.84}$$

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 [66] or [85] for a recent review on the multiplication of random matrices.

2.4. Replica analysis

2.4.1. Resolvent and the Replica trick

As we noticed above (Eq. (2.6)), 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. [47,46], or [48] 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 C. Other available techniques include Feynman diagrams [67,86].

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

$$\mathfrak{g}(z) = \sum_{i=1}^{N} \frac{1}{z - \nu_i} = \frac{\partial}{\partial z} \log \prod_{i=1}^{N} (z - \nu_i) = \frac{\partial}{\partial z} \log \det(zI - \mathbf{M}). \tag{2.85}$$

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

$$\mathcal{Z}(z) \equiv (\det(zI - \mathbf{M}))^{-1/2} = \int \exp\left[-\frac{1}{2}\sum_{i,j=1}^{N} \eta_i (zI - \mathbf{M})_{ij} \eta_j\right] \prod_{j=1}^{N} \left(\frac{\mathrm{d}\eta_j}{\sqrt{2\pi}}\right). \tag{2.86}$$

Plugging this last equation into (2.85) and assuming that the Stieltjes transform is self-averaging, we see that we need to compute the average of the logarithm of $\mathcal{Z}(z)$:

$$\mathfrak{g}(z) = -2\frac{\partial}{\partial z}\mathbb{E}\log \mathcal{Z}(z), \tag{2.87}$$

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

$$\log \mathcal{Z} = \lim_{n\to 0} \frac{\mathcal{Z}^n - 1}{n}, \tag{2.88}$$

so that one formally has

$$\mathfrak{g}(z) = \lim_{n\to 0} \frac{\partial}{\partial z} \frac{\mathbb{E}\mathcal{Z}^n - 1}{n}. \tag{2.89}$$

We have thus transformed the problem (2.87) 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 (2.88) 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 [87], 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 review.

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 (2.88), 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

$$(z\mathbf{I}_N - \mathbf{M})^{-1}_{ij} = \frac{\int \left(\prod_{k=1}^{N} d\eta_k\right) \eta_i \eta_j \exp\left\{-\frac{1}{2}\sum_{k,l=1}^{N} \eta_k (z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\}}{\int \left(\prod_{k=1}^{N} d\eta_k\right) \exp\left\{-\frac{1}{2}\sum_{k,l=1}^{N} \eta_k (z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\}}. \tag{2.90}$$

As explained in Appendix C, we expect that (2.90) is self-averaging in the LDL thanks to the Central Limit Theorem, so that:

$$(z\mathbf{I}_N - \mathbf{M})^{-1}_{ij} = \left\langle \frac{1}{\mathcal{Z}} \int \left(\prod_{k=1}^{N} d\eta_k\right) \eta_i \eta_j \exp\left\{-\frac{1}{2}\sum_{k,l=1}^{N} \eta_k (z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\} \right\rangle_{\mathcal{P}_\mathbf{M}}, \tag{2.91}$$

where $\mathcal{Z}$ is as above the partition function, i.e. the denominator in Eq. (2.90). The replica identity for resolvent is given by

$$\begin{aligned} G_{ij}(z) &= \lim_{n\to 0} \left\langle \mathcal{Z}^{n-1} \int \left(\prod_{k=1}^{N} d\eta_k\right) \eta_i \eta_j \exp\left\{-\frac{1}{2}\sum_{k,l=1}^{N} \eta_k (z\delta_{kl} - \mathbf{M}_{kl})\eta_l\right\} \right\rangle_{\mathcal{P}_\mathbf{M}} \\ &= \lim_{n\to 0} \int \left(\prod_{k=1}^{N}\prod_{\alpha=1}^{n} d\eta_k^\alpha\right) \eta_i^1 \eta_j^1 \left\langle \prod_{\alpha=1}^{n} \exp\left\{-\frac{1}{2}\sum_{k,l=1}^{N} \eta_k^\alpha (z\delta_{kl} - \mathbf{M}_{kl})\eta_l^\alpha\right\} \right\rangle_{\mathcal{P}_\mathbf{M}}. \end{aligned} \tag{2.92}$$

Again, we managed to rewrite the initial problem (2.91) as the computation of n replicas. We emphasize that (2.92) 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 (2.92) 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 [38]. As will become apparent below, we consider a model of random matrices inspired by Free Probability theory, i.e. $\mathbf{M} = \mathbf{A} + \boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^*$ and $\mathbf{M} = \mathbf{A}^{1/2}\boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^*\mathbf{A}^{1/2}$ (see Section 2.3 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 D).

2.4.2. Matrix multiplication using replicas

We reconsider the model (2.80) 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 (2.92) to $\mathbf{M} = \mathbf{A}^{1/2}\boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^*\mathbf{A}^{1/2}$ so that we get

$$G_{ij}(z) = \lim_{n\to 0} \int \left(\prod_{k=1}^{N}\prod_{\alpha=1}^{n} d\eta_k^\alpha\right) \eta_i^1 \eta_j^1 e^{-\frac{z}{2}\sum_{\alpha=1}^{n}\sum_{k=1}^{N}(\eta_k^\alpha)^2} \mathcal{I}_1\left(\sum_{\alpha=1}^{n} \left(\eta^\alpha \mathbf{A}^{1/2}\right)\left(\eta^\alpha \mathbf{A}^{1/2}\right)^*, \mathbf{B}\right), \tag{2.93}$$

where

$$\mathcal{I}_\beta(\mathbf{A}', \mathbf{B}) := \int \exp\left[-\frac{\beta N}{2} \operatorname{Tr} \mathbf{A}' \Omega \mathbf{B} \Omega^*\right] \mathcal{D}\Omega, \tag{2.94}$$

is the so-called *Harish-Chandra–Itzykson–Zuber* integral [88,89]. 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 (2.94) 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.5) in Appendix A[15]:

$$\mathcal{I}_1\left(\sum_{\alpha=1}^{n} \left(\eta^\alpha \mathbf{A}^{1/2}\right)\left(\eta^\alpha \mathbf{A}^{1/2}\right)^*, \mathbf{B}\right) \underset{N\to\infty}{\sim} \exp\left[\frac{N}{2}\sum_{\alpha=1}^{n} \mathcal{W}_\mathbf{B}\left(\frac{1}{N}\sum_{i=1}^{N} (\eta_i^\alpha)^2 a_i\right)\right], \tag{2.95}$$

with

$$\mathcal{W}'_\mathbf{B}(.) = \mathcal{R}_\mathbf{B}(.), \tag{2.96}$$

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

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 (2.93) 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

$$\delta\Big(p^\alpha - \frac{1}{N}\sum_{i=1}^N (\eta_i^\alpha)^2 a_i\Big) = \int \frac{1}{2\pi}\exp\Big[\mathrm{i}\zeta^\alpha\Big(p^\alpha - \frac{1}{N}\sum_{i=1}^N (\eta_i^\alpha)^2 a_i\Big)\Big]\mathrm{d}\zeta^\alpha, \tag{2.97}$$

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

$$G_{ij}(z) \propto \int\int \left(\prod_{\alpha=1}^n \mathrm{d}p^\alpha \mathrm{d}\zeta^\alpha\right) \frac{\delta_{ij}}{z - \zeta^1 a_i}\exp\Big[-\frac{Nn}{2}F_0(p^\alpha, \zeta^\alpha)\Big] \tag{2.98}$$

where F_0 is the free energy given by

$$F_0(p^\alpha, \zeta^\alpha) = \frac{1}{n}\sum_{\alpha=1}^n \left[\frac{1}{N}\sum_{k=1}^N \log(z - \zeta^\alpha a_k) + \zeta^\alpha p^\alpha - \mathcal{W}_{\mathbf{B}}(p^\alpha)\right]. \tag{2.99}$$

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 $\zeta^\alpha = \zeta^*, \forall \alpha = 1, \dots, n$. 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. [47,87] or [90] 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 $\mathbf{M}$[16]:

$$z\mathbf{G}_{\mathbf{M}}(z)_{i,j} \underset{N\to\infty}{\sim} Z(z)\mathbf{G}_{\mathbf{A}}(Z(z))_{i,j}, \qquad Z(z) := z\,\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1), \tag{2.100}$$

which is often referred to as a *subordination* relation between the resolvent of $\mathbf{M}$ and $\mathbf{A}$. Taking the trace of both sides of the above equation, one notices that (2.100) is a generalization of the formula (2.84) as a matrix. We should emphasize that Eq. (2.100) is self-averaging 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 $\langle \mathbf{G}_{\mathbf{M}}^2(z)\rangle$. When considering the whole matrix $\mathbf{G}_{\mathbf{M}}(z)$ one should rather write:

$$z\,\langle \mathbf{G}_{\mathbf{M}}(z)\rangle \underset{N\to\infty}{\sim} Z(z)\mathbf{G}_{\mathbf{A}}(Z(z)), \qquad Z(z) := z\,\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1). \tag{2.101}$$

Note that the average resolvent $\langle \mathbf{G}_{\mathbf{M}}(z)\rangle$ is diagonal in the eigenbasis of $\mathbf{A}$, as expected by symmetry.

We can redo the exact same calculations for the free addition model $\mathbf{M} = \mathbf{A} + \Omega\mathbf{B}\Omega^*$, still with $\mathbf{A} = \mathrm{diag}(a_1, a_2, \dots, a_N)$ (see Appendix D). Starting from the replica identity (2.92) and then applying (A.5), we obtain the following expression [38]:

$$G_{ij}(z) \propto \int\int \left(\prod_{\alpha=1}^n d p^\alpha d\zeta^\alpha\right) \frac{\delta_{ij}}{z - \zeta^1 - a_i}\exp\left\{-\frac{Nn}{2}F_0^a(p^\alpha, \zeta^\alpha)\right\}, \tag{2.102}$$

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

$$F_0^a(p, \zeta) := \frac{1}{Nn}\sum_{\alpha=1}^n \left[\sum_{k=1}^N \log(z - \zeta^\alpha - a_k) - \mathcal{W}_{\mathbf{B}}(p^\alpha) + p^\alpha\zeta^\alpha\right]. \tag{2.103}$$

Invoking once again the replica symmetric ansatz, the subordination for the resolvent under the free addition model follows from a saddle-point analysis [38]

$$\mathbf{G}_{\mathbf{M}}(z)_{i,j} \underset{N\to\infty}{\sim} \mathbf{G}_{\mathbf{A}}(Z_a(z))_{i,j}, \qquad Z_a(z) := z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z)), \tag{2.104}$$

which is exactly the result obtained in [92] in a mathematical formalism. Again taking the trace of both sides of this equation allows one to recover the relation (2.67) between Stieltjes transforms.

16 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 [91] for a detail presentation of this concept for Wigner matrices).

2.4.3. Free multiplication: replica saddle-point analysis

We now present the derivation of (2.100) from (2.98). We shall see that it actually provides an elementary derivation of the free multiplication formula (2.81). 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

$$\zeta^* = \mathcal{R}_{\mathbf{B}}(p^*). \tag{2.105}$$

The other derivative with respect to ζ gives:

$$p^* = \frac{1}{\zeta^* N}\sum_{k=1}^{N} \frac{a_k}{z/\zeta^* - a_k} = \frac{\mathcal{T}_{\mathbf{A}}\left(\frac{z}{\mathcal{R}_{\mathbf{B}}(p^*)}\right)}{\mathcal{R}_{\mathbf{B}}(p^*)}. \tag{2.106}$$

Hence, plugging (2.105) and (2.106) into (2.98), we get in the large N limit and then the limit $n \to 0$ by

$$G_{ij}(z)_{ij} = \frac{\delta_{ij}}{z - \mathcal{R}_{\mathbf{B}}(p^*)c_i}. \tag{2.107}$$

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

$$z g_{\mathbf{M}}(z) = Z g_{\mathbf{A}}(Z), \quad \text{with} \quad Z \equiv Z(z) = \frac{z}{\mathcal{R}_{\mathbf{B}}(p^*)}, \tag{2.108}$$

which can rewrite as

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

Let us define

$$\omega = \mathcal{T}_{\mathbf{M}}(z) = \mathcal{T}_{\mathbf{A}}(Z). \tag{2.109}$$

Using Eq. (2.106), this latter equation implies $p^* = \omega/\mathcal{R}_{\mathbf{B}}(p^*)$. Let us now show how to retrieve the free multiplicative convolution (2.81) from (2.108) in the large N limit. Indeed, let us rewrite (2.109) as

$$z\mathcal{T}_{\mathbf{M}}(z) = Z\mathcal{T}_{\mathbf{A}}(Z)\mathcal{R}_{\mathbf{B}}(p^*), \tag{2.110}$$

and it is trivial to see that using (2.109) 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 (2.23), this yields

$$\mathcal{S}_{\mathbf{M}}(\omega) = \mathcal{S}_{\mathbf{A}}(\omega)\frac{1}{\mathcal{R}_{\mathbf{B}}(p^*)}. \tag{2.111}$$

Using (2.25), we also have

$$\frac{1}{\mathcal{R}_{\mathbf{B}}(p^*)} = \mathcal{S}_{\mathbf{B}}(p^*\mathcal{R}_{\mathbf{B}}(p^*)), \tag{2.112}$$

But recalling that $p^* = \omega/\mathcal{R}_{\mathbf{B}}(p^*)$, we conclude from (2.105), (2.109) and (2.112) that

$$\frac{1}{\zeta^*} = \mathcal{R}_{\mathbf{B}}(p^*) = \mathcal{S}_{\mathbf{B}}(\mathcal{T}_{\mathbf{M}}(z)). \tag{2.113}$$

Going back to (2.111), we see that the spectral density of $\mathbf{M}$ is given by Voiculescu's free multiplication formula

$$\mathcal{S}_{\mathbf{M}}(\omega) = \mathcal{S}_{\mathbf{A}}(\omega)\mathcal{S}_{\mathbf{B}}(\omega), \tag{2.114}$$

confirming that the replica symmetry ansatz is indeed valid in this case. Finally, by plugging (2.113) into (2.107), we get the result (2.100).

3. Spectrum of large empirical covariance matrices

3.1. Sample covariance matrices

3.1.1. 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 review, 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

$$\mathbf{C}_{ij} = \mathbb{E}\big[y_i y_j\big], \quad i, j \in [\![1, N]\!] \tag{3.1}$$

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

$$\mathbf{C} = \sum_{i=1}^{N} \mu_i \mathbf{v}_i \mathbf{v}_i^*, \tag{3.2}$$

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

As illustrated in the introduction, the concept of covariance 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 a large loss of a diversified portfolio is dominated by the correlated moves of its different constituents (see Section 7.1 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 of *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 tth 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. (3.1) is then approximated by an average value over the whole sample data of size T, leading to the *sample* (or *empirical*) covariance matrix estimator:

$$\mathbf{E}_{ij} = \frac{1}{T}(\mathbf{Y}\mathbf{Y}^*)_{ij} = \frac{1}{T}\sum_{\tau=1}^{T} \mathbf{Y}_{it}\mathbf{Y}_{jt}. \tag{3.3}$$

In the statistical literature, this estimator is known as *Pearson* estimator and in the RMT community, the resulting matrix sometimes referred to as the Wishart Ensemble. Whereas the Wigner Ensemble has been the subject of a large amount of studies in physics [50], results on the Wishart Ensemble mostly come from mathematics & statistics [18,56,93,94], telecommunication [57] or the financial/econophysics literature [24,29,67], although some work in the physics literature also exists [95–98]—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 data-sets such that the sample size T *and* the number of variables N are both very large, specific issues arise when the observation ratio $q = N/T$ is of order unity. This setting is known in the literature as the high-dimensional limit or Kolmogorov regime (or more commonly called 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 precisely where tools from RMT can be helpful to make precise statements on the empirical covariance matrix (3.3).

A typical question would be to study the ESD of $\mathbf{E}$ in order to quantify its deviation from 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 [9], and is given by Eq. (2.38) 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 section. 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

$$\mathbf{Y} = \sqrt{\mathbf{C}}\mathbf{X}, \tag{3.4}$$

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

$$\mathbb{E}[X_{it}] = 0, \qquad \mathbb{E}[X_{it}^2] = \frac{1}{T}. \tag{3.5}$$

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 $\sqrt{T}X_{it}$ have a bounded 4th moment, in other words that the distribution cannot be extremely fat-tailed.

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

$$\mathbf{E} = \sum_{i=1}^{N} \lambda_i \mathbf{u}_i \mathbf{u}_i^*, \tag{3.6}$$

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

(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 component are labeled λ_- and λ_+ (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 λ_-, λ_+.

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

We end up this short introduction 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.

3.1.2. Zero-mean assumption

In real data-sets, sample vectors $\mathbf{y}_t$ usually have a non-zero mean (even if the true underlying distribution is of zero mean). One can 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:

$$\check{E}_{ij} = \frac{1}{T-1}\sum_{t=1}^{T}\Big(Y_{it} - \overline{Y_i}\Big)\Big(Y_{jt} - \overline{Y_j}\Big), \qquad \overline{Y_i} = \frac{1}{T}\sum_{\tau=1}^{T} Y_{i\tau} \tag{3.7}$$

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 2.3.2 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 [99, 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 with the corresponding $\mathbf{E}$ in the next sections.

3.1.3. Distribution of the data entries

The second remark deals with the distribution of the entries of the matrix $\mathbf{Y}$ given in Eq. (3.5). It is well-known for instance that financial returns are strongly non-Gaussian, with power-law tails [25], 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 of *robust* estimators [100,101] 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 [57,102–104]. In particular, the so-called *Maronna* robust M-estimator of $\mathbf{C}$ is the (unique) solution of the fixed point equation

$$\mathbf{M} := \frac{1}{T}\sum_{t=1}^{T} U\Big(\frac{1}{N}\mathbf{y}_t^*\mathbf{M}^{-1}\mathbf{y}_t\Big)\mathbf{y}_t\mathbf{y}_t^*, \tag{3.8}$$

where U is a non-increasing function. It was shown recently [105] that the matrix $\mathbf{M}$ converges to a matrix of the form encountered in Eq. (2.80) and thus different from $\mathbf{E}$. However, tractable formulae are scarce except for the multivariate Student distribution where $U(x) \sim x^{-1}$ [102,103,106,107]. In that case, we have from [108] that the LSD of $\mathbf{M}$ converges (almost surely) to that of standard Wishart matrix $\mathbf{E}$ 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 [102]). We postpone discussions about other class of distributions to Section 9.

3.2. Bulk statistics

3.2.1. Marčenko–Pastur equation

As we alluded to in the introduction, the fundamental tool to analyze the spectrum of large sample covariance matrices is the Marčenko–Pastur equation [18]. We actually have already encountered a special case of this equation in Section 2.2.3 where we consider the LSD of $\mathbf{E}$ under the null hypothesis $\mathbf{C} = \mathbf{I}_N$ (isotropic case). In this section, we allow the population correlation matrix $\mathbf{C}$ to be *anisotropic*, that is to say not proportional to the identity matrix. As we shall see, the final result is not as simple as Eq. (2.41) but many properties can be inferred from it.

The Marčenko–Pastur (MP) equation dates back to their seminal paper [18] 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 Section 7 for some examples or [93] and references therein). There are several ways to obtain this result, using e.g. recursion techniques [109], Feynman diagram expansion [67], replicas (see [22] or Section 2.4 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. (3.4) as

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

where the matrix $\mathbf{X}$ satisfies Eq. (3.5) and is independent from $\mathbf{C}$. The model falls into the model of free multiplication encountered in Section 2.3 since $\mathbf{E}$ is the free multiplicative convolution of $\mathbf{C}$ with a white Wishart kernel for $N \to \infty$ [110]. Therefore, the Stieltjes transform of $\mathbf{E}$ is exactly given by Eq. (2.84) that we specialize to

$$z\mathfrak{g}_{\mathbf{E}}(z) = Z(z)\mathfrak{g}_{\mathbf{C}}\left(Z(z)\right), \qquad \text{with} \qquad Z(z) \coloneqq z\mathcal{S}_{\mathcal{W}}(z\mathfrak{g}_{\mathbf{E}}(z) - 1). \tag{3.9}$$

Moreover, the $\mathcal{S}$-transform of $\mathcal{W}$ was obtained in Eq. (2.44), i.e. $\mathcal{S}_{\mathcal{W}}(z) = (1+qz)^{-1}$ for any $q > 0$. Thus, we can re-express $Z(z)$ as:

$$Z(z) = \frac{z}{1 - q + qz\mathfrak{g}_{\mathbf{E}}(z)}, \tag{3.10}$$

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. (3.9) 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 form as:

$$\mathfrak{g}_{\mathbf{E}}(z) = \int \frac{\rho_{\mathbf{C}}(\mu)\mathrm{d}\mu}{z - \mu(1 - q + qz\mathfrak{g}_{\mathbf{E}}(z))}. \tag{3.11}$$

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 Eq. (3.9).

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 [33] and, more recently, of Ledoit and Wolf [111] allows one to make progress in this direction with a numerical scheme that solves a discretized version of the inverse problem Eq. (3.11). 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 [112]:

$$q\frac{\partial Z(z,q)}{\partial q} = (Z(z,q) - z)\frac{\partial Z(z,q)}{\partial z}, \tag{3.12}$$

with initial condition $Z(z, q \to 0) = z + qz(1 - z\mathfrak{g}_{\mathbf{C}}(z))$. This representation can be given a direct interpretation but whether it is useful numerically or analytically remains to be seen.

3.2.2. 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. (2.21), it is easy to see that we can rewrite Eq. (3.9) as

$$\mathcal{T}_{\mathbf{E}}(z) = \mathcal{T}_{\mathbf{C}}(Z(z)), \qquad Z(z) = \frac{z}{1 + q\mathcal{T}_{\mathbf{E}}(z)}. \tag{3.13}$$

We know from Eq. (2.22) that the $\mathcal{T}$-transform can be expressed as power series for $z \to \infty$, hence we have

$$\mathcal{T}_{\mathbf{E}}(z) \underset{z\to\infty}{=} \sum_{k=1}^{\infty} \varphi(\mathbf{E}^k) z^{-k}, \tag{3.14}$$

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 $z \to \infty$

$$\mathcal{T}_{\mathbf{C}}(Z(z)) \underset{z\to\infty}{=} \sum_{k=1}^{\infty} \frac{\varphi(\mathbf{C}^k)}{z^k} \left(1 + q\sum_{\ell=1}^{\infty} \varphi(\mathbf{E}^\ell) z^{-\ell}\right)^k. \tag{3.15}$$

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. (3.13) which yields

$$\sum_{k=1}^{\infty} \frac{\varphi(\mathbf{E}^k)}{z^k} = \sum_{k=1}^{\infty} \frac{\varphi(\mathbf{C}^k)}{z^k} \left(1 + q\sum_{\ell=1}^{\infty} \varphi(\mathbf{E}^\ell) z^{-\ell}\right)^k, \tag{3.16}$$

which was first obtained in [67]. In particular, we infer from Eq. (3.16) that the first three moments of $\rho_{\mathbf{E}}$ satisfy

$$\begin{aligned} \varphi(\mathbf{E}) &= \varphi(\mathbf{C}) = 1 \\ \varphi(\mathbf{E}^2) &= \varphi(\mathbf{C}^2) + q \\ \varphi(\mathbf{E}^3) &= \varphi(\mathbf{C}^3) + 3q\varphi(\mathbf{C}^2) + q^2. \end{aligned} \tag{3.17}$$

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 wider (for $q > 0$) than the spectrum of the population covariance matrix $\mathbf{C}$. This is an alternative way to convince ourselves 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. (3.9) in terms of the $\mathcal{R}$-transform (see below for a derivation)

$$\omega\mathcal{R}_{\mathbf{E}}(\omega) = \zeta(\omega)\mathcal{R}_{\mathbf{C}}(\zeta(\omega)), \qquad \zeta(\omega) = \omega\big(1 + q\omega\mathcal{R}_{\mathbf{E}}(\omega)\big). \tag{3.18}$$

Using the cumulants expansion of the $\mathcal{R}$-transform, given in Eq. (2.19), we obtain for $\omega \to 0$

$$\omega\mathcal{R}_{\mathbf{E}}(\omega) = \sum_{\ell=1}^{\infty} \kappa_\ell(\mathbf{E})\omega^\ell, \tag{3.19}$$

and

$$\zeta(\omega)\mathcal{R}_{\mathbf{C}}(\zeta(\omega)) = \sum_{\ell=1}^{\infty} \kappa_\ell(\mathbf{C})\omega^\ell \left(1 + q\sum_{m=1}^{\infty} \kappa_m(\mathbf{E})\omega^m\right)^\ell. \tag{3.20}$$

By regrouping these last two equations into Eq. (3.18), the analogue of Eq. (3.16) in terms of free cumulants reads:

$$\sum_{\ell=1}^{\infty} \kappa_\ell(\mathbf{E})\omega^\ell = \sum_{\ell=1}^{\infty} \kappa_\ell(\mathbf{C})\omega^\ell \left(1 + q\sum_{m=1}^{\infty} \kappa_m(\mathbf{E})\omega^m\right)^\ell, \tag{3.21}$$

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

$$\mathfrak{g}(z) \underset{z\to 0}{=} -\sum_{k=1}^{\infty} \varphi(\mathbf{E}^{-k}) z^{k-1}. \tag{3.22}$$

This allows one to study the moment of the LSD of $\mathbf{E}^{-1}$ and this turns out to be an important quantity in many applications (see Section 7). Using Eq. (3.9), 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. (3.9) at $z \to 0$ and $q \in (0, 1)$:

$$\sum_{k=1}^{\infty} \varphi(\mathbf{E}^{-k}) z^k = \sum_{k=1}^{\infty} \varphi(\mathbf{C}^{-k}) \left(\frac{z}{1-q} \right)^k \left(\frac{1}{1 - \frac{q}{1-q} \sum_{\ell=1}^{\infty} \varphi(\mathbf{E}^{-\ell}) z^\ell} \right)^k, \tag{3.23}$$

that is a little bit more cumbersome than the moment generating expansion Eq. (3.16) or the cumulant expansion (3.21). Still, we get at leading order that

$$\varphi(\mathbf{E}^{-1}) = \frac{\varphi(\mathbf{C}^{-1})}{1-q}, \qquad \varphi(\mathbf{E}^{-2}) = \frac{\varphi(\mathbf{C}^{-2})}{(1-q)^2} + \frac{q\varphi(\mathbf{C}^{-1})^2}{(1-q)^3}. \tag{3.24}$$

We will see in Section 7.1 that the first relation (that can be found in [67]) has direct consequences for the out-of-sample risk of optimized portfolios.

Let us now give a formal derivation of Eq. (3.18). Let us define

$$\omega = \mathfrak{g}_{\mathbf{E}}(z), \qquad \zeta = \mathfrak{g}_{\mathbf{C}}(Z), \tag{3.25}$$

which allows us to rewrite Eq. (3.9) as

$$\omega \mathcal{B}_{\mathbf{E}}(\omega) = \zeta \mathcal{B}_{\mathbf{C}}(\zeta), \qquad Z \equiv \mathcal{B}_{\mathbf{C}}(\zeta) = \frac{\mathcal{B}_{\mathbf{E}}(\omega)}{1 - q + q\omega \mathcal{B}_{\mathbf{E}}(\omega)}. \tag{3.26}$$

Then, using the definition (2.16) of the $\mathcal{R}$-transform, we can rewrite this last equation as

$$\omega \mathcal{R}_{\mathbf{E}}(\omega) = \zeta \mathcal{R}_{\mathbf{C}}(\zeta), \qquad \mathcal{R}_{\mathbf{C}}(\zeta) + \frac{1}{\zeta} = \frac{\mathcal{R}_{\mathbf{E}}(\omega) + 1/\omega}{1 + q\omega \mathcal{R}_{\mathbf{E}}(\omega)}. \tag{3.27}$$

We deduce that

$$\mathcal{R}_{\mathbf{C}}(\zeta) = \frac{\mathcal{R}_{\mathbf{E}}(\omega) + 1/\omega}{1 + q\omega \mathcal{R}_{\mathbf{E}}(\omega)} - \frac{1}{\zeta}, \tag{3.28}$$

which yields

$$\omega \mathcal{R}_{\mathbf{E}}(\omega) = \zeta \left(\frac{\mathcal{R}_{\mathbf{E}}(\omega) + 1/\omega}{1 + q\omega \mathcal{R}_{\mathbf{E}}(\omega)} - \frac{1}{\zeta} \right). \tag{3.29}$$

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

$$\omega \mathcal{R}_{\mathbf{E}}(\omega) + 1 = \frac{\zeta}{\omega} \left(\frac{\omega \mathcal{R}_{\mathbf{E}}(\omega) + 1}{1 + q\omega \mathcal{R}_{\mathbf{E}}(\omega)} \right), \tag{3.30}$$

that is to say

$$\zeta \equiv \zeta(\omega) = \omega \big(1 + q\omega \mathcal{R}_{\mathbf{E}}(\omega) \big), \tag{3.31}$$

and Eq. (3.18) immediately follows by plugging this last equation into Eq. (3.28).

3.2.3. *Dual representation and edges of the spectrum*

Although a lot of information about the spectrum of $\mathbf{E}$ can be gathered from the Marčenko–Pastur equation (3.9), 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 one can answer some of these questions by using a dual representation of Eq. (3.9).

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

$$\mathbf{S} := \frac{1}{T} \mathbf{Y}^* \mathbf{Y} \equiv \mathbf{X}^* \mathbf{C} \mathbf{X}, \tag{3.32}$$

where we used Eq. (3.4) 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 $\mathbf{S}$:

$$\mathfrak{g}_{\mathbf{S}}(z) = \frac{1}{T} \left[\frac{T-N}{z} + N \mathfrak{g}_{\mathbf{E}}(z) \right] = \frac{1-q}{z} + q \mathfrak{g}_{\mathbf{E}}(z) = \frac{1}{Z(z)}. \tag{3.33}$$

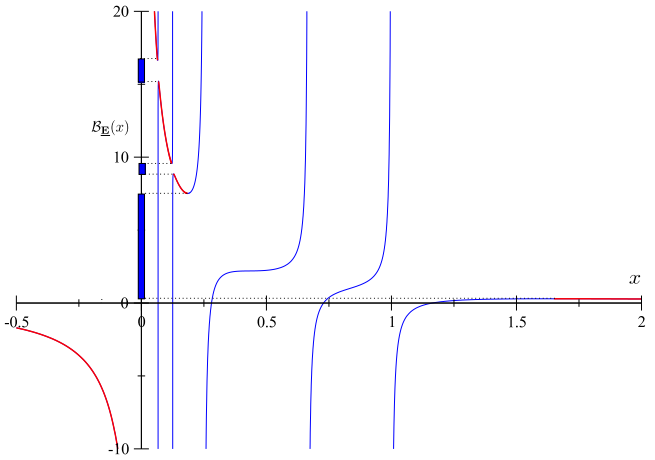


Fig. 6. The function $\mathcal{B}_{\underline{\mathbf{E}}}(x)$ with population eigenvalues density given by $0.002\,\delta_{15} + 0.002\,\delta_8 + 0.396\,\delta_3 + 0.3\,\delta_{1.5} + 0.3\,\delta_1$. Here $T = 1000, N = 500$ and we have 3 connected components. The vertical asymptotes are located at 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 $\mathfrak{g}_{\mathbf{S}}|_{\mathbb{R}\setminus\operatorname{supp}\rho_{\mathbf{S}}}$ is drawn in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The introduction of this dual representation of the empirical matrix allows one to get the following expression from Eq. (3.11):

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

After some manipulations, we can rewrite this last equation as

$$z = \frac{1}{\mathfrak{g}_{\mathbf{S}}(z)} + q\int \frac{\rho_{\mathbf{C}}(\mu)d\mu}{\mu^{-1} - \mathfrak{g}_{\mathbf{S}}(z)}. \tag{3.34}$$

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

$$\mathcal{B}_{\mathbf{S}}(\omega) := \frac{1}{\omega} + q\int \frac{\rho_{\mathbf{C}}(\mu)d\mu}{\mu^{-1} - \omega}, \tag{3.35}$$

and this is the dual representation of the Marčenko–Pastur equation (3.9). The analytic behavior of this last equation has been the subject of several studies, especially in [30]. In particular, it was proved that there exists a *unique* $\omega \in \mathbb{C}_+$ that solves Eq. (3.35). This yields the Stieltjes transform of **S** from which we re-obtain the Stieltjes transform of **E** using Eq. (3.33). We will see in the next section that the dual representation (3.35) 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 **E** can be inferred from Eq. (3.35). Within a one cut-assumption, the edges of the support of $\rho_{\mathbf{E}}$ are given by:

$$\lambda_{\pm}^{\mathbf{E}} = \mathcal{B}_{\mathbf{S}}(\omega_{\pm}) \quad \text{where } \omega_{\pm} \in \mathbb{R}^+ \text{ is such that } \mathcal{B}'_{\mathbf{S}}(\omega_{\pm}) = 0. \tag{3.36}$$

Indeed, knowing the spectral density of **S** allows us to get the spectral density of **E** since from Eq. (3.33) one gets:

$$\rho_{\mathbf{S}}(\lambda) = q\rho_{\mathbf{E}}(\lambda) + (1 - q)^+\delta_0, \tag{3.37}$$

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

$$\mathfrak{g}'_{\mathbf{S}}(z) = -\int \frac{\rho_{\mathbf{S}}(x)\mathrm{d}x}{(z - x)^2} < 0, \tag{3.38}$$

for any $z \notin \operatorname{supp}[\rho_{\mathbf{S}}]$, meaning that it is strictly decreasing outside of the support. We saw in Section 2.1.2 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. (3.36). If one assumes that there are a finite number r of (non-degenerate) spikes, we can readily generalize the above arguments and find that there will be $2(r + 1)$ critical points (see Fig. 6 for an illustration with two non-degenerate spikes).

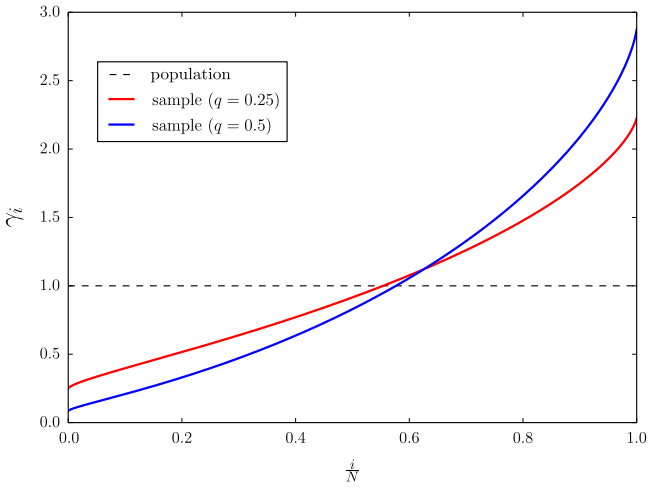


Fig. 7. Typical position of the sample eigenvalues under the Marčenko–Pastur law (2.42) with a finite observation ratio $q = 0.25$ (red line) and $q = 0.5$ (blue line). The dotted line corresponds to the locations of the population eigenvalues and we see a significant deviation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2.4. Solving Marčenko–Pastur equation

In this section, we investigate the direct problem of solving the Marčenko–Pastur equation (3.9) 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 **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 (3.11), 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 3.2.2 that the LSD of **E** will be significantly different from that of **C**. The influence of q can be well understood in the simple case where $\mathbf{C} = \mathbf{I}_N$. We know from Section 2.2.3 that this case is exactly solvable and the LSD of **E** is the well-known Marčenko–Pastur law (2.42), that we recall here:

$$\mathfrak{g}_{\mathbf{E}}(z) = \frac{z + q - 1 - \sqrt{z - \lambda_-^{\mathrm{mp}}}\sqrt{z - \lambda_+^{\mathrm{mp}}}}{2qz}, \qquad \lambda_\pm^{\mathrm{mp}} = (1 \pm \sqrt{q})^2. \tag{3.39}$$

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 variance of the sample eigenvalue distribution is order q, highlighting the systematic bias in the estimation of the eigenvalues using **E** when $q = \mathcal{O}(1)$. This effect can be visualized using the quantile representation of the spectral distribution. Indeed, it is known since [99,113] that the bulk eigenvalues $[\lambda_i]_{i \in [\![r+1,N]\!]}$ converge in the high-dimensional regime to their "quantile positions" $[\gamma_i]_{i \in [\![r+1,N]\!]}$. More precisely, this reads:

$$\lambda_i \approx \gamma_i, \quad \text{where} \quad \frac{i}{N} = \int^{\gamma_i} \rho_{\mathbf{E}}(\lambda)\mathrm{d}\lambda, \qquad i \geqslant r+1. \tag{3.40}$$

We plot the γ_i's of the Marčenko–Pastur law in Fig. 7 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 **E** is an untrustworthy estimator when the sample size is of the same order of magnitude as 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 **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 hyper-parameter $\kappa > 0$. From Section 2.2.4, we recall that

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

for $\kappa > 0$. Then, using the free multiplication formula (2.81), we have $\mathcal{S}_{\mathbf{E}}(\omega) = \mathcal{S}_{\mathbf{C}}(\omega)\mathcal{S}_{\mathcal{W}}(\omega)$ where $\mathcal{S}_{\mathcal{W}}(\omega)$ is given in (2.44), which yields a quadratic equation in $\mathcal{T}_{\mathbf{E}}(z)$. This implies that $\mathfrak{g}_{\mathbf{E}}$ reads:

$$\mathfrak{g}_{\mathbf{E}}(z) = \frac{z(1+\kappa) - \kappa(1-q) \pm \sqrt{(\kappa(1-q) - z(1+\kappa))^2 - z(z+2q\kappa)(2\kappa+1)}}{z(z+2q\kappa)}, \tag{3.41}$$

from which we can retrieve the edges of the support:

$$\lambda_\pm^{\mathrm{iw}} = \frac{1}{\kappa}\left[(1+q)\kappa + 1 \pm \sqrt{(2\kappa+1)(2q\kappa+1)}\right]. \tag{3.42}$$

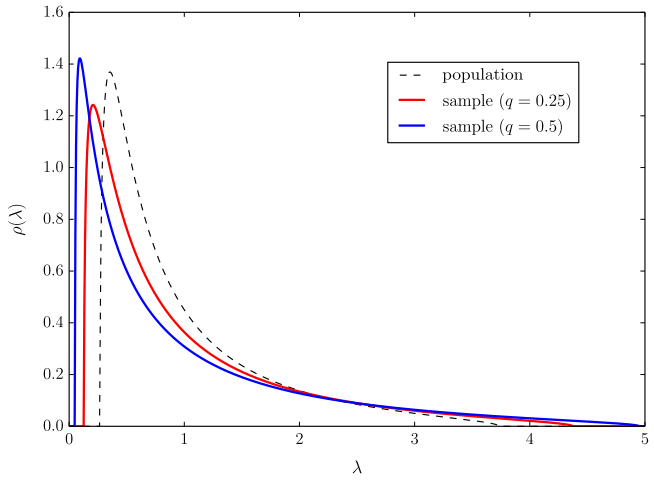


Fig. 8. Solution of the Marčenko–Pastur equation for the eigenvalue distribution of $\mathbf{E}$ when $\mathbf{C}$ is an inverse Wishart matrix with parameter $\kappa = 1.0$ for $q = 0.25$ (red line) and $q = 0.5$ (blue line). The black dotted line corresponds to the LSD $\rho_{\mathbf{C}}$.

One can check that the limit $\kappa \to \infty$ recovers the null hypothesis case $\mathbf{C} = \mathbf{I}_N$; the lower κ, the wider the spectrum of $\mathbf{C}$. We plot in Fig. 8 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 Section 5 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 (2.81) that

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

It is then easy to see from the definition (2.23) that $\mathcal{T}_{\mathbf{E}}(z) \equiv \omega(z)$ is solution of the cubic equation,

$$z(1+\omega(z))(1+q_0\omega(z))(1+q\omega(z)) - \omega(z) = 0, \tag{3.43}$$

from which we obtain $\mathfrak{g}_{\mathbf{E}}(z)$ thanks to (2.21) 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

$$\mathcal{R}_{\mathbf{C}}(\omega) = 1 + \sigma^2\omega, \tag{3.44}$$

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 (3.18), we find that $\mathfrak{g}_{\mathbf{E}}(z) = \omega$ is the solution of quartic equation:

$$\sigma^2\omega^2(1+q\omega\mathcal{R}_{\mathbf{E}}(\omega))^2 + \omega(1+q\omega\mathcal{R}_{\mathbf{E}}(\omega)) - \omega\mathcal{R}_{\mathbf{E}}(\omega) = 0, \tag{3.45}$$

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 (3.35) of Eq. (3.9) comes to be particularly useful. To solve the MP equation for a given z, we seek a certain $\mathfrak{g} \equiv \mathfrak{g}_{\mathbf{S}}$ such that[17]

$$z = \mathcal{B}_{\mathbf{S}}(\mathfrak{g}), \qquad \mathfrak{g} \in \mathbb{C}_+, \tag{3.46}$$

where the expression of $\mathcal{B}_{\mathbf{S}}$ in terms of $\rho_{\mathbf{C}}$ is explicit and given in Eq. (3.35). Numerically, the above equation is easily solved using a simple gradient descent algorithm, i.e. find $\mathfrak{g} \in \mathbb{C}_+$ such that

$$\begin{cases} \mathrm{Re}(z) = \mathrm{Re}\big[\mathcal{B}_{\mathbf{S}}(\mathfrak{g})\big] \\ \mathrm{Im}(z) = \mathrm{Im}\big[\mathcal{B}_{\mathbf{S}}(\mathfrak{g})\big]. \end{cases} \tag{3.47}$$

It then suffices to use Eq. (3.33) in order to get $\mathfrak{g}_{\mathbf{E}}(z)$ for any $z \in \mathbb{C}_-$. 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 ε an arbitrary small real positive

[17] Recall that $\mathbf{S}$ is the $T \times T$ equivalent of $\mathbf{E}$ defined in Eq. (3.32).

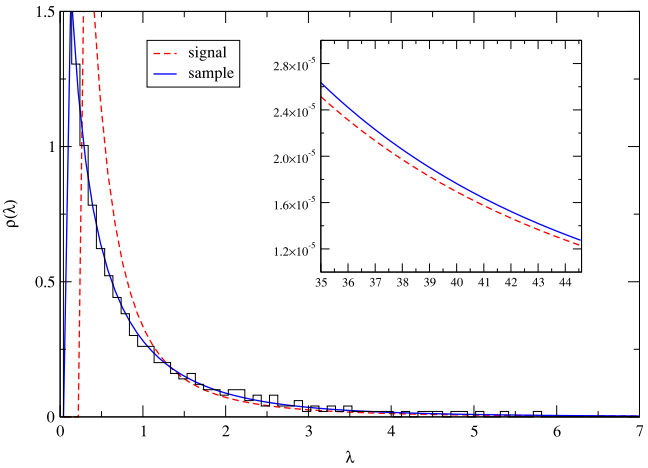


Fig. 9. Resolution of the Marčenko–Pastur equation when $\rho_{\mathbf{C}}$ is given a power law density with parameter $\lambda_0 = 0.3$ and a finite observation ratio $q = 0.5$ and $N = 500$. The dotted line corresponds to the LSD of $\mathbf{C}$ while the plain line corresponds to the LSD of $\mathbf{E}$. The histogram is the ESD when we compute $\mathbf{E}$ from the definition (3.3). The main figure covers the bulk of the eigenvalues while the inset zooms in the region of very large eigenvalues.

number into Eq. (3.47). Note that in the case where $\mathfrak{g}_{\mathbf{C}}$ is known, one can rewrite Eq. (3.35) as

$$\mathcal{B}_{\mathbf{S}}(x) = \frac{1}{x}\left[1 - q + \frac{q}{x}\mathfrak{g}_{\mathbf{C}}\left(\frac{1}{x}\right)\right], \tag{3.48}$$

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 parametrization is to assume a power-law distribution of the form [29]:

$$\rho_{\mathbf{C}}(\lambda) = \frac{sA}{(\lambda + \lambda_0)^{1+s}}\Theta(\lambda - \lambda_{\min}), \tag{3.49}$$

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 $\int \rho_{\mathbf{C}}(x)\mathrm{d}x = 1$ and $\int x\rho_{\mathbf{C}}(x)\mathrm{d}x = 1$. This leads to: $\lambda_{\min} = (1 - \lambda_0)/2$ and $A = (1 - \lambda_{\min})^2$. We restrict to $\lambda_0 > -11$ such that $\lambda_{\min} < 1$. From the density Eq. (3.49), one can perform the Stieltjes transform straightaway to find

$$\mathfrak{g}_{\mathbf{C}}(z) = \frac{1}{z + 1 - 2\lambda_0} + \frac{2(1 - \lambda_0)}{(z + 1 - 2\lambda_0)^2} + \frac{2(1 - \lambda_0)^2}{(z + 1 - 2\lambda_0)^3}\left[\log\left(\frac{\lambda_0 - z}{1 - \lambda_0}\right)\right], \tag{3.50}$$

which allows one to solve Eq. (3.48) for $\mathfrak{g}_{\mathbf{E}}(z)$ with only a few iterations. As we observe in Fig. 9, the theoretical value obtained from the numerical scheme (3.47) 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 the above case, where the spectrum of $\mathbf{E}$ embraces nearly all the positive real number line. Note that an ODE approach to solve Marčenko–Pastur equation has been proposed recently in [114]. While this is a bit more complicated to implement, numerical simulations in [114] show that it yields more robust results than the simple newton approach.

3.3. Edges and outliers statistics

As we alluded to several times above, the practical usefulness of the theoretical 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. 10 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 7.2 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, even 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 subsections 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.

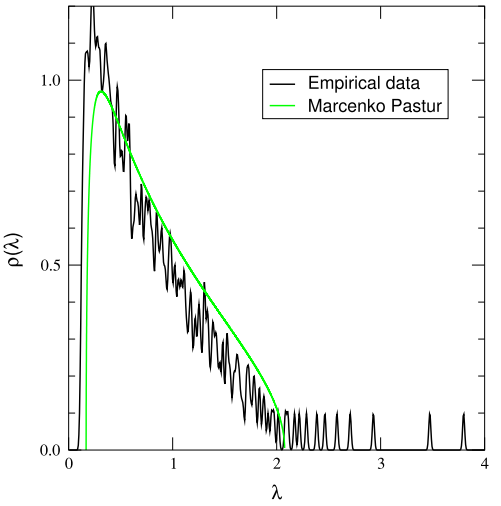


Fig. 10. Test of the null hypothesis on the empirical correlation matrix **E** using US stocks' data with $N = 406$ and $T = 1300$.

3.3.1. 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 [74,115]). 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 [116], culminating in the beautiful results by Tracy & Widom on the distribution of the *largest* eigenvalue of a random matrix [27]. 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 [117] on this topic). The derivation of the Tracy–Widom distribution mainly relies on Orthogonal polynomials that we will not discuss in this review (see e.g. [27,118]) but there also exists an alternative approach [119]. The link between this limiting law and the largest eigenvalue 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. [26,51,52,120–122] for details and references).

The Tracy–Widom result characterizes precisely the distance between the largest eigenvalue λ_1 of **E** and the upper edge of the spectrum that we denoted by λ_+. 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,

$$\mathcal{P}\left(\lambda_1 \leqslant \lambda_+ + \gamma N^{-2/3} u\right) = F_1(u), \tag{3.51}$$

where γ 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 of a finite fourth moment, see [115,74]).

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

$$\ln f_1(u) \propto -u^{3/2}, \quad (u \to +\infty); \qquad \ln f_1(u) \propto -|u|^3, \quad (u \to -\infty). \tag{3.52}$$

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 $\lambda_1 - \lambda_+ = \mathcal{O}(1)$) can also be obtained [51].

Note that the distribution of the smallest eigenvalue $\lambda_{\min}$ around the lower edge λ_- 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. [123].

3.3.2. 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. 10 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 "blow out" a finite number of eigenvalues of a given (spikeless) correlation matrix $\underline{\mathbf{C}}$, that we construct as:

$$\underline{\mathbf{C}} = \sum_{i=1}^{N} \underline{\mu}_i \mathbf{v}_i \mathbf{v}_i^*, \quad \text{where} \quad \underline{\mu}_i = \begin{cases} \mu_0 & \text{if } i \leqslant r \\ \mu_i & \text{if } i \geqslant r+1. \end{cases} \tag{3.53}$$

We choose the eigenvalue μ_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 $[\mu_i]_{i \leqslant r}$ is well separated from the bulk by assumption, we may parametrize each spike μ_i by a positive real number d_i for any $i \leqslant r$ as follows:

$$\mu_i = \mu_0(1+d_i) \equiv \mu_{r+1}(1+d_i), \qquad d_i > 0, \ i \leqslant r. \tag{3.54}$$

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

$$\mathbf{C} = \sum_{i=1}^{N} \mu_i \mathbf{v}_i \mathbf{v}_i^*, \quad \text{where} \quad \mu_i = \begin{cases} \mu_0(1+d_i) & \text{if } i \leqslant r \\ \mu_i & \text{if } i \geqslant r+1. \end{cases} \tag{3.55}$$

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

$$\mathbf{C} = \underline{\mathbf{C}}\big(\mathbf{I}_N + \mathbf{V}^{(r)} \mathbf{D} \mathbf{V}^{(r)*}\big), \tag{3.56}$$

where $\mathbf{V}^{(r)} := [\mathbf{v}_1, \ldots, \mathbf{v}_r] \in \mathbb{R}^{N \times r}$ and $\mathbf{D} := \mathrm{diag}(d_1, \ldots, d_r)$ 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 $T \times T$ "dual" matrix. As noticed in [39], 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 [39] 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:

$$\det(z\mathbf{I}_N - \mathbf{E}) = \det(z\mathbf{I}_N - \underline{\mathbf{E}}) \det(\mathbf{I}_N - d_1 (z\mathbf{I}_N - \underline{\mathbf{E}})^{-1} \mathbf{v}_1 \mathbf{v}_1^* \mathbf{E}). \tag{3.57}$$

We can conclude that λ_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 λ_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:

$$d_1 \left[\lambda_1 \langle \mathbf{v}_1, \mathbf{G}_{\underline{\mathbf{E}}}(\lambda_1) \mathbf{v}_1 \rangle - 1 \right] = 1, \tag{3.58}$$

where $\mathbf{G}_{\underline{\mathbf{E}}}$ is the resolvent of $\underline{\mathbf{E}}$. The difficult part of (3.58) 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[18] that $\mathbf{C}$ is Gaussian, which allows us to arbitrarily set $\mathbf{v}_1 = (1, 0, \ldots, 0)$. Then the equation we try to solve is:

$$\lambda_1 \mathbf{G}_{\underline{\mathbf{E}}}(\lambda_1)_{11} = d_1^{-1} + 1. \tag{3.59}$$

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. (4.6) (see (C.19) for an alternative derivation). The result reads

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

where we used the identity (3.33) and that $\underline{\mu}_1 \equiv \mu_{r+1}$ by construction of (3.56) in the last step. If λ_1 is not an eigenvalue of $\underline{\mathbf{E}}$, we find that Eq. (3.59) becomes in the LDL

$$\frac{1}{1 - \mu_{r+1}\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1)} = d_1^{-1} + 1, \tag{3.60}$$

which is equivalent to:

$$\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1) = \frac{1}{\mu_{r+1}(1+d_1)} \equiv \frac{1}{\mu_1}, \tag{3.61}$$

where we used (3.54) in the last step. Hence, we see that λ_1 is an outlier if it satisfies for large N:

$$\lambda_1 = \theta(\mu_1) := \mathcal{B}_{\underline{\mathbf{S}}}\left(\frac{1}{\mu_1}\right), \tag{3.62}$$

[18] The extension to non-Gaussian entries can be done using standard comparison techniques, see e.g. [113] for details.

This result is very general and can be extended for any outlier λ_i with $i \in [\![1, r]\!]$. Moreover, we see that for $N \to \infty$, the (random) outlier λ_1 converges to a deterministic function of μ_1. Hence, the function (3.62) depicts the "classical location" at which an outlier sticks and can therefore be interpreted as the analog of (3.40) for outliers. Note however that (3.62) requires the knowledge of the spikeless matrix $\underline{\mathbf{S}}$ (or $\underline{\mathbf{E}}$). In practice, one should make some assumptions to decide whether a given empirical eigenvalue should be considered as a spike.

The result (3.62) generalizes the result of Baik–Ben Arous–Péché for the spiked covariance matrix model [121]. 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 (2.42). In fact, in the limit $N \to \infty$, the spectra 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 (3.35) to find

$$\mathcal{B}_{\underline{\mathbf{S}}}(x) = \frac{1}{x} + \frac{q}{1-x}. \tag{3.63}$$

Applying this formula to Eq. (3.62) then leads to the so-called BBP phase transition

$$\begin{cases} \lambda_1 = \mu_1 + q\dfrac{\mu_1}{\mu_1 - 1} & \text{if } \mu_1 > 1 + \sqrt{q}; \\ \lambda_1 = \lambda_+ = (1+\sqrt{q})^2 & \text{if } \mu_1 \leqslant 1 + \sqrt{q}, \end{cases} \tag{3.64}$$

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 λ_+. All these isolated eigenvalues have Gaussian fluctuations of order $T^{-1/2}$ [121]. The typical fluctuation of order $T^{-1/2}$ is also true for an arbitrary $\mathbf{C}$ [39], and is much smaller than the uncertainty in the bulk of the distribution, of order $\sqrt{q}$. Note that a naive application of Eq. (3.9) to outliers would lead to a "mini-Wishart" distribution around the top eigenvalue, which is incorrect (the distribution is Gaussian) except if the top eigenvalue has a degeneracy proportional to N.

4. Statistics of the eigenvectors

We saw in the previous section that tools from RMT allow one to infer many properties of the (asymptotic) spectrum of $\mathbf{E}$, be it for the bulk or for more localized regions of the spectrum (edges and outliers). These results allow us to characterize in 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 $[\lambda_i]_{i\in[\![N]\!]}$ converges to a non-deterministic value, but this value is different from the corresponding "true" population eigenvalue $[\mu_i]_{i\in[\![N]\!]}$. 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) [97,98,113,124].

On the other hand, results concerning the eigenvectors are comparatively 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 $\mathbf{E} = \sqrt{\mathbf{C}}\mathcal{W}\sqrt{\mathbf{C}}$ and $\mathbf{E}' = \sqrt{\mathbf{C}}\mathcal{W}'\sqrt{\mathbf{C}}$ – that remain correlated through $\mathbf{C}$?

The aim of this section 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 2 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 (2.80)), but the same questions can be investigated for additive noise as well, see [40,113,125–127] and Appendix D.

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" as $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle$. 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 $[\mathbf{v}_i]_{i\in[\![N]\!]}$ and those of the sample matrix $[\mathbf{u}_i]_{i\in[\![N]\!]}$. The matrix of squared overlaps is defined as $\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2$, it forms a so-called bi-stochastic 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 section will be the resolvent (and not its normalized trace as in the previous section). Indeed, if we choose the $\mathbf{v}$'s to be our reference basis, we find from (2.6):

$$\langle \mathbf{v}, \mathbf{G}_{\mathbf{E}}(z)\mathbf{v} \rangle = \sum_{i=1}^{N} \frac{\langle \mathbf{v}, \mathbf{u}_i \rangle^2}{z - \lambda_i}, \tag{4.1}$$

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 $\mathbf{G_E}(z)$ of the form:

$$\langle \mathbf{v}, \mathbf{G_E}(z)\mathbf{v}' \rangle = \sum_{i=1}^{N} \frac{\langle \mathbf{v}, \mathbf{u}_i \rangle \langle \mathbf{u}_i, \mathbf{v}' \rangle}{z - \lambda_i}, \tag{4.2}$$

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

We see from Eqs. (4.1) and (4.2) 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 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 $N \to \infty$,[19] 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,

$$\Phi(\lambda_i, \mu_j) := N\mathbb{E}[\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2], \tag{4.3}$$

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. We emphasize that we consider the population eigenvectors to be deterministic throughout this section. Only the sample eigenvectors are random. 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

$$\Phi(\lambda_i, \tilde{\lambda}_j) := N\mathbb{E}[\langle \mathbf{u}_i, \tilde{\mathbf{u}}_j \rangle^2], \tag{4.4}$$

where $[\tilde{\lambda}_i]_{i\in[\![N]\!]}$ and $[\tilde{\mathbf{u}}_i]_{i\in[\![N]\!]}$ 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 $\mathbf{C}$).

We end this introduction with a short remark on the somewhat vague definitions (4.3) and (4.4). As explained above, we index the eigenvectors by their corresponding eigenvalues and this allows us to consider the continuous limit of (4.3). However, a more precise definition should be that $\Phi(\lambda, \mu) := \mathbb{E}\big[N^{-1} \sum_{i,j=1}^{N} \langle \mathbf{u}_i, \tilde{\mathbf{b}}_j \rangle^2 \delta(\lambda - \lambda_i)\delta(\mu - \mu_i)\big]$ but we keep the notation (4.3), with a slight abuse of notation, as it will be more convenient to separate the analysis between an outlier or bulk eigenvalue. We emphasize that this remark also holds for the overlaps (4.4) as well.

4.1. 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. (4.3) 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 now show that Eq. (4.3) 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. $\Phi(\lambda_i, \mu_j) \sim \mathcal{O}(1)$ (and not $\mathcal{O}(N)$) for any $i \in [\![r+1, N]\!]$ and $j \in [\![N]\!]$;

(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 of 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 true eigenvectors (on this point, see however [41]). Still, as we will see below, the squared 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 3 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. [36,40,128]. In fact, the angle between the sample spike eigenvectors with the parent spike can be computed exactly, see below.

[19] Recall that we have indexed the eigenvectors by their associated eigenvalue.

4.1.1. The bulk

Let us focus on the bulk eigenvectors first, i.e. eigenvectors associated to eigenvalues lying in the bulk of the spectral density when the dimension of the empirical correlation matrix grows to infinity. This question has been investigated very recently in [37,38] and we repeat the different arguments here. The first step is to characterize the asymptotic behavior of the resolvent of sample covariance matrices. This can be done by specializing Eq. (2.100) 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 (2.44), we know the $\mathcal{S}$-transform of white Wishart matrices explicitly so that one finds from Eq. (2.44), for $N \to \infty$:

$$z\mathbf{G}_{\mathbf{E}}(z)_{ij} = Z(z)\mathbf{G}_{\mathbf{C}}(Z(z))_{ij}, \quad \text{with} \quad Z = \frac{z}{1 - q + qz\mathfrak{g}_{\mathbf{E}}(z)}. \tag{4.5}$$

In the literature, such a limiting result is referred to as a "deterministic equivalent", as the RHS depends only on deterministic quantities,[20] and this is another evidence of the self-averaging property for large random matrices.

One should notice that (4.5) is a relation between resolvent matrices that generalizes the scalar Marčenko–Pastur equation (3.9) (which can be recovered by taking the trace on both sides of the equation). This relation first appeared in [67], obtained using a planar diagram expansion valid for Gaussian entries. A few years later, that result was proven rigorously in Ref. [113] in a much more general framework, highlighting again the *universal* nature of the resolvent of random matrices, down to the local scale.[21] Choosing to work in the basis where $\mathbf{C}$ is diagonal, Eq. (4.5) reduces to:

$$\mathbf{G}_{\mathbf{E}}(z)_{ij} = \frac{\delta_{ij}}{z - \mu_i(1 - q + qz\mathfrak{g}_{\mathbf{E}}(z))}. \tag{4.6}$$

This deterministic equivalent holds with fluctuations of order $N^{-1/2}$. This can be deduced e.g. from the Central Limit Theorem (CLT) (see Appendix C). Quite interestingly, an explicit upper bound for the error term is provided in [113]. In particular, the authors showed that Eq. (4.5) holds at a local scale $\eta = \widehat{\eta}N^{-1}$ with $\widehat{\eta} \gg 1$, with an error term bounded from above by:

$$\Psi(z) := \sqrt{q\frac{\operatorname{Im}\mathfrak{g}_{\mathbf{S}}(z)}{\widehat{\eta}}} + \frac{q}{\widehat{\eta}}, \tag{4.7}$$

provided that N is large enough. We give an illustration of this ergodic behavior in Fig. 11, and we see the agreement is excellent.

How can we compute the mean squared overlap using (4.5)? The idea is to derive an inversion formula similar to (2.11) for the full resolvent. More specifically, we start from (2.6) 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 (3.40) so that we obtain for $N \to \infty$ that

$$\langle \mathbf{v}_j, \mathbf{G}_{\mathbf{E}}(z)\mathbf{v}_j\rangle \underset{N\uparrow\infty}{\sim} \int \frac{\Phi(\lambda, \mu_j)\rho_{\mathbf{E}}(\lambda)}{\lambda_i - \lambda - \mathrm{i}\eta}\mathrm{d}\lambda \tag{4.8}$$

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 η around λ. Therefore, the final inversion formula is obtained using the Sokhotski–Plemelj identity as:

$$\Phi(\lambda_i, \mu_j) = \frac{1}{\pi\rho_{\mathbf{E}}(\lambda_i)} \lim_{\eta\to 0^+} \operatorname{Im}\langle \mathbf{v}_j, \mathbf{G}_{\mathbf{E}}(\lambda_i - \mathrm{i}\eta)\mathbf{v}_j\rangle, \tag{4.9}$$

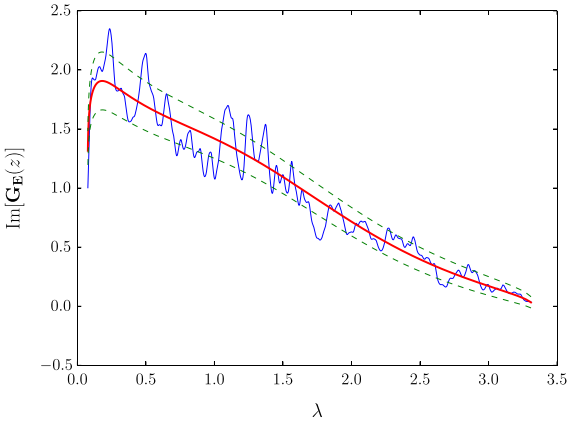
where the assumption that λ_i lies in the bulk of the spectrum is crucial here. This last identity thus allows us to compute the squared overlap $\Phi(\lambda_i, \mu_j)$ from the full resolvent $\mathbf{G}_{\mathbf{E}}$, for any i in the bulk ($i \geqslant r + 1$) and a fixed $j \in [\![1, N]\!]$. Specializing to the explicit form of $\mathbf{G}_{\mathbf{E}}(z)$ given in Eq. (4.6), we finally obtain a beautiful explicit result for the (rescaled) average squared overlap:

$$\Phi(\lambda_i, \mu_j) = \frac{q\mu_j\lambda_i}{(\mu_j(1 - q) - \lambda_i + q\mu_j\lambda_i\mathfrak{h}_{\mathbf{E}}(\lambda_i))^2 + q^2\mu_j^2\lambda_i^2\pi^2\rho_{\mathbf{E}}^2(\lambda_i)}, \tag{4.10}$$

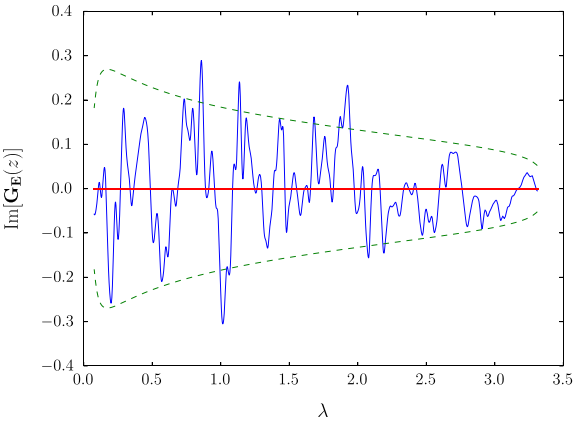
with $i \in [\![r + 1, N]\!], j \in [\![1, N]\!]$ and $\mathfrak{h}_{\mathbf{E}}(\lambda_i)$ denotes the real part of the Stieltjes transform $\mathfrak{g}_{\mathbf{E}}$ (see Eq. (2.9)). This relation is exact in the limit $N \to \infty$ and was first derived by Ledoit and Péché in [37]. We emphasize again that this expression remains correct even if μ_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 at large N, and therefore non-outlier sample eigenvectors retain very little information about their corresponding true eigenvectors. This implies that

[20] Recall that $\mathfrak{g}_{\mathbf{E}}(z)$ is the *limiting* Stieltjes transform.

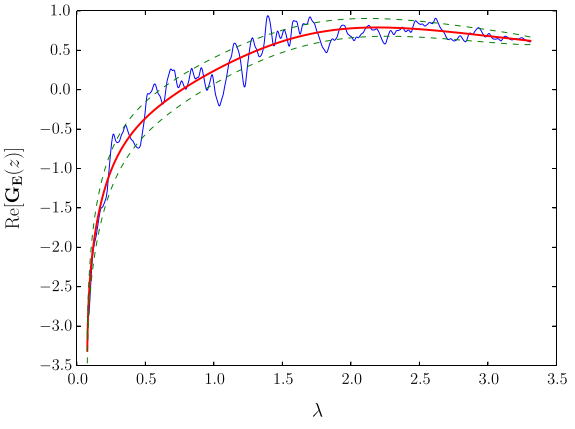
[21] Note that the Gaussian assumption is not needed either within the Replica method presented in Section 2.



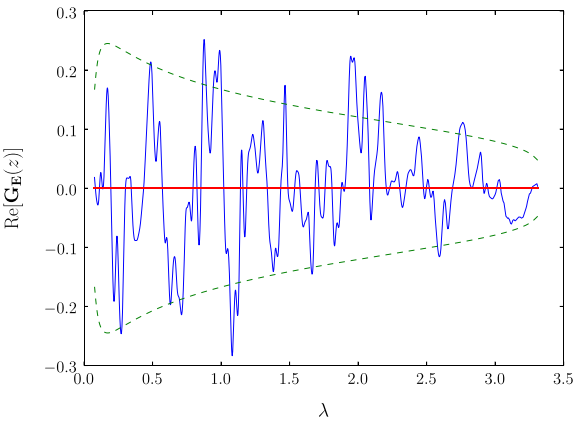
(a) Diagonal entry of $\mathrm{Im}[\mathbf{G}_{\mathbf{E}}(z)]$ with $i = j = 1000$.



(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 $j = 1001$.

Fig. 11. Illustration of Eq. (4.6). 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. (4.5). The green dotted corresponds to the confidence interval whose formula is given by Eq. (4.7). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

any bulk eigenvector is a extremely poor estimator of the true one in the high-dimensional regime. We provide in Fig. 12 an illustration of Eq. (4.10) 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. (4.10). 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.

4.1.2. Outliers

By construction, the spiked correlation model of Section 3.3.2 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 [129], 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 (4.9). 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 $\underline{\mathbf{C}} = \mathbf{I}_N$ [128,93]. 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 [39].

From Eq. (3.62) we saw that each outlier eigenvalues $[\lambda_i]_{i\in[\![1,r]\!]}$ of $\mathbf{E}$ converges to a deterministic limit $\theta(\mu_i)$, where μ_i is the corresponding population spike and θ 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 [39] for details). Then, defining Γ_i to be the boundary of the

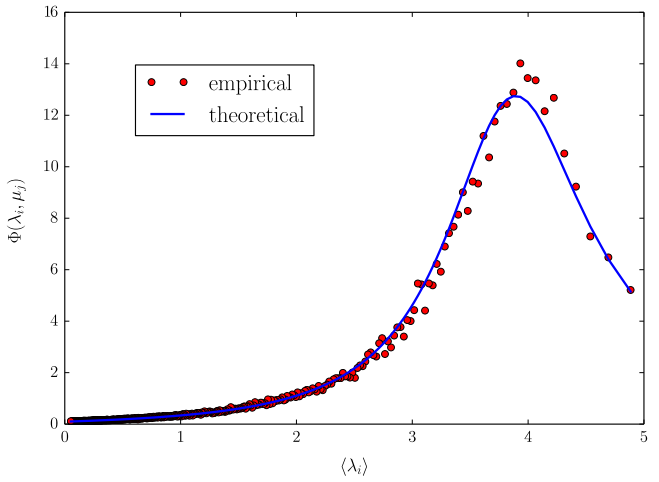


Fig. 12. Rescaled mean squared overlaps $\Phi(\lambda_i, \mu_j)$ as a function of λ_i. We choose $\mathbf{C}$ as an inverse-Wishart matrix with parameter $\kappa = 1.0$ and set $N = 500$, $q = 0.5$. The empirical average (red points) comes from 500 independent realizations of $\mathbf{E}$. The theoretical prediction (blue line) is given by Eq. (4.10). The peak of the mean squared overlap is in the vicinity of $\lambda_i \approx \mu_j \approx 4$.

closed disc D_i, we can obtain the squared overlap for outlier eigenvectors using Cauchy's integral formula

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = \frac{1}{2\pi \, \mathrm{i}} \oint_{\Gamma_i} \langle \mathbf{v}_j, \mathbf{G}_{\mathbf{E}}(z) \mathbf{v}_j \rangle \mathrm{d}z, \tag{4.11}$$

for $i, j \in [\![1, r]\!]$. We emphasize there is no expectation value in Eq. (4.11) (compare to our definition of the overlap in Eq. (4.3)). 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 (3.56) 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 Appendix B for a reminder):

$$\mathbf{V}^{(r)*} \mathbf{G}_{\mathbf{E}}(z) \mathbf{V}^{(r)} = -\frac{1}{z} \left[\mathbf{D}^{-1} - \frac{\sqrt{\mathbf{I}_N + \mathbf{D}}}{\mathbf{D}} \left(\mathbf{D}^{-1} + \mathbf{I}_N - z \mathbf{V}^{(r)*} \mathbf{G}_{\underline{\mathbf{E}}} \mathbf{V}^{(r)} \right)^{-1} \frac{\sqrt{\mathbf{I}_N + \mathbf{D}}}{\mathbf{D}} \right]. \tag{4.12}$$

This identity has been used in several studies that deal with related problems [99,39] and references therein. Its derivation only needs linear algebra arguments and can be found in Section 4.1.3. With this identity, we see that the statistics of the outliers can be expressed through the spikeless matrix $\underline{\mathbf{E}}$. In particular, the integrand of (4.11) 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 (4.5). By plugging (4.5) into (4.12), one obtains

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = -\frac{1}{2\pi \, \mathrm{i}} \oint_{\theta(\Gamma_i)} \frac{1}{z} \left[\frac{1}{d_j} - \frac{1 + d_j}{d_j^2} \frac{1}{d_j^{-1} + 1 - z \langle \mathbf{v}_j, \mathbf{G}_{\mathbf{E}_0}(z) \mathbf{v}_j \rangle} \right] dz. \tag{4.13}$$

Then, using Eq. (3.58) and Cauchy's theorem, one eventually finds [39]

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = \delta_{ij} \mu_i \frac{\theta'(\mu_i)}{\theta(\mu_i)} + \mathcal{O}(N^{-1/2}) = \delta_{ij} \mu_i \frac{\theta'(\mu_i)}{\lambda_i} + \mathcal{O}(N^{-1/2}), \tag{4.14}$$

for any $i, j \in [\![1, r]\!]$ and where we used (3.62) 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. (4.14) 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 (4.14) is to reconsider the spiked covariance matrix model introduced in the previous section. Let us assume for simplicity a single spike ($r = 1$) and from Eq. (3.63), 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 (4.14) yields

$$\langle \mathbf{u}_1, \mathbf{v}_1 \rangle^2 = \frac{\mu_1}{\theta(\mu_1)} \left(1 - \frac{q}{(\mu_1 - 1)^2} \right) + \mathcal{O}(T^{-1/2}), \tag{4.15}$$

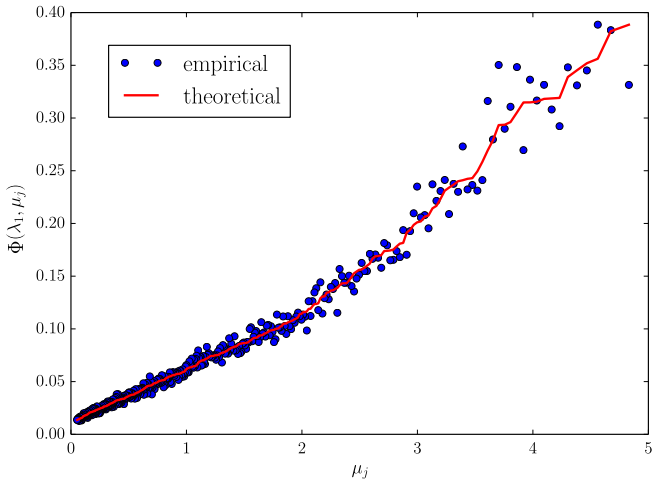


Fig. 13. Rescaled mean squared overlap $\Phi(\lambda_1, \mu_j)$ as a function of μ_j for $j > 1$. We chose the spikeless population matrix $\underline{\mathbf{C}}$ to be an Inverse-Wishart matrix with parameter $\kappa = 1.0$ and $N = 500$. We add a rank one perturbation such that $\lambda_1 \approx 10$ 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. (4.16).

which is the expected result [36,40,99,115,41]. 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 (3.64).

The same analysis can be applied for the overlap between the sample spikes and the population bulk eigenvalues $j > r$. The details can be found in [39] and the final result reads

$$\Phi(\lambda_i, \mu_j) = q \frac{\mu_j}{\lambda_i (1 - \mu_j/\mu_i)^2}, \qquad i \in [\![1, r]\!],\ j \in [\![r+1, N]\!]. \tag{4.16}$$

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. (4.16) in Fig. 13 as a function of the population eigenvalues μ_i with $i > 2$ in the case where $r = 1$: 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 is excellent.

4.1.3. *Derivation of the identity (4.12)*

The derivation of the identity (4.12) 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. (3.56):

$$\begin{aligned} \sqrt{\underline{\mathbf{C}}}\,\mathbf{C}^{-1}\sqrt{\underline{\mathbf{C}}} - \mathbf{I}_N &= (\mathbf{I}_N + \mathbf{V}\mathbf{D}\mathbf{V}^*)^{-1} - \mathbf{I}_N \\ &= -(\mathbf{I}_N + \mathbf{V}\mathbf{D}\mathbf{V}^*)^{-1}\mathbf{V}\mathbf{D}\mathbf{V}^* \\ &= -\mathbf{V}\mathbf{D}(\mathbf{I}_r + \mathbf{D})^{-1}\mathbf{V}^* \end{aligned} \tag{4.17}$$

where we used the resolvent identity (4.32) in the second line. This allows us to get (omitting the argument z)

$$\begin{aligned} \underline{\mathbf{C}}^{-1/2}\mathbf{C}^{1/2}\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{C}^{1/2}\underline{\mathbf{C}}^{-1/2} &= \underline{\mathbf{C}}^{-1/2}\left(z\mathbf{C}^{-1} - \mathbf{X}\mathbf{X}^*\right)^{-1}\underline{\mathbf{C}}^{-1/2} \\ &= \left(z(\underline{\mathbf{C}}^{1/2}\mathbf{C}^{-1}\underline{\mathbf{C}}^{1/2} - \mathbf{I}_N) + z\mathbf{I}_N - \underline{\mathbf{E}}\right)^{-1} \\ &= \left(-z\mathbf{V}\mathbf{D}(I + \mathbf{D})^{-1}\mathbf{V}^* + \mathbf{G}_{\underline{\mathbf{E}}}^{-1}\right)^{-1}, \end{aligned} \tag{4.18}$$

where we invoked the previous identity Eq. (4.17) in the last step. From (B.8), 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 $\mathbf{C} \equiv \mathbf{V}^*$:

$$\underline{\mathbf{C}}^{-1/2}\mathbf{C}^{1/2}\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{C}^{1/2}\underline{\mathbf{C}}^{-1/2} = \mathbf{G}_{\underline{\mathbf{E}}} + z\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V}\Big(\mathbf{D}^{-1} + \mathbf{I}_r - z\mathbf{V}^*\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V}\Big)^{-1}\mathbf{V}^*\mathbf{G}_{\underline{\mathbf{E}}}. \tag{4.19}$$

From there, one has

$$(\mathbf{I}_N + \mathbf{D})^{1/2}\mathbf{V}^*\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V}(\mathbf{I}_N + \mathbf{D})^{1/2} = \mathbf{V}^*\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V} + z\mathbf{V}^*\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V}\Big(\mathbf{D}^{-1} + \mathbf{I}_r - \mathbf{V}^*\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V}\Big)^{-1}\mathbf{V}^*\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V}. \tag{4.20}$$

We then use the identity

$$\mathbf{A} - \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A} = \mathbf{B} - \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}, \tag{4.21}$$

with $\mathbf{A} = V^*\mathbf{G}_{\underline{\mathbf{E}}}V$ and $\mathbf{B} = -(\mathbf{D}^{-1}+\mathbf{I}_r)/z$ to obtain

$$(\mathbf{I}_r+\mathbf{D})^{1/2}\mathbf{V}^*\mathbf{G}_{\mathbf{E}}\mathbf{V}(\mathbf{I}_r+\mathbf{D})^{1/2} = -\frac{1}{z}\left[\frac{\mathbf{I}_r+\mathbf{D}}{\mathbf{D}} + \frac{\mathbf{I}_r+\mathbf{D}}{\mathbf{D}}\left(-(\mathbf{D}^{-1}+\mathbf{I}_r) + z\mathbf{V}^*\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V}\right)^{-1}\frac{\mathbf{I}_r+\mathbf{D}}{\mathbf{D}}\right]. \tag{4.22}$$

By rearranging the terms, we finally get

$$\mathbf{V}^*\mathbf{G}_{\mathbf{E}}\mathbf{V} = -\frac{1}{z}\left[\mathbf{D}^{-1} - \frac{\sqrt{\mathbf{I}_r+\mathbf{D}}}{\mathbf{D}}\left(\mathbf{D}^{-1}+\mathbf{I}_r - z\mathbf{V}^*\mathbf{G}_{\underline{\mathbf{E}}}\mathbf{V}\right)^{-1}\frac{\sqrt{\mathbf{I}_r+\mathbf{D}}}{\mathbf{D}}\right], \tag{4.23}$$

which is precisely Eq. (4.12).

4.2. Overlaps between the eigenvectors of correlated sample covariance matrices

We now consider the second problem of this section, 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

$$\mathbf{E} := \sqrt{\mathbf{C}}\boldsymbol{\mathcal{W}}\sqrt{\mathbf{C}}, \qquad \tilde{\mathbf{E}} := \sqrt{\mathbf{C}}\tilde{\boldsymbol{\mathcal{W}}}\sqrt{\mathbf{C}}, \tag{4.24}$$

where $\boldsymbol{\mathcal{W}}$ and $\tilde{\boldsymbol{\mathcal{W}}}$ are two independent white Wishart matrix with parameter q and q' respectively. As in Section 4.1, 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. (4.4) 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 (4.4) significantly depart from the trivial null hypothesis as soon as the population $\mathbf{C}$ has a non-trivial 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 [127] and we shall only give here the main steps. For the sake of clarity, we use the notations $\tilde{\lambda}_1 \geqslant \tilde{\lambda}_2 \geqslant \cdots \geqslant \tilde{\lambda}_N$ 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 (4.4) as it is usually done in RMT. To that end, we define the bivariate complex function

$$\psi(z,\tilde{z}) := \left\langle \frac{1}{N}\operatorname{Tr}\left[(z-\mathbf{E})^{-1}(\tilde{z}-\tilde{\mathbf{E}})^{-1}\right]\right\rangle_{\mathcal{P}}, \tag{4.25}$$

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

$$\psi(z,\tilde{z}) = \left\langle \frac{1}{N}\sum_{i,j=1}^{N}\frac{1}{z-\lambda_i}\frac{1}{\tilde{z}-\tilde{\lambda}_j}\langle \mathbf{u}_i, \tilde{\mathbf{u}}_j\rangle^2\right\rangle_{\mathcal{P}}, \tag{4.26}$$

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 3.2.1) so that the sample eigenvalues become deterministic in the large N limit. Hence, we obtain after taking the continuous limit

$$\psi(z,\tilde{z}) \sim \int\int \frac{\rho(\lambda)}{z-\lambda}\frac{\tilde{\rho}(\tilde{\lambda})}{\tilde{z}-\tilde{\lambda}}\Phi(\lambda,\tilde{\lambda})\mathrm{d}\lambda\mathrm{d}\tilde{\lambda}, \tag{4.27}$$

where ρ and $\tilde{\rho}$ are respectively the spectral density of $\mathbf{E}$ and $\tilde{\mathbf{E}}$, and Φ denotes the mean squared overlap defined in (4.4). Then, it suffices to compute

$$\psi(x-\mathrm{i}\eta, y\pm\mathrm{i}\eta) \sim \int\int \frac{(x-\lambda+\mathrm{i}\eta)}{(x-\lambda)^2+\eta^2}\frac{(y-\tilde{\lambda}\mp\mathrm{i}\eta)}{(y-\tilde{\lambda})^2+\eta^2}\rho(\lambda)\tilde{\rho}(\tilde{\lambda})\Phi(\lambda,\tilde{\lambda})\mathrm{d}\lambda\mathrm{d}\tilde{\lambda} \tag{4.28}$$

from which, one deduces that

$$\mathrm{Re}\big[\psi(x-\mathrm{i}\eta, y+\mathrm{i}\eta) - \psi(x-\mathrm{i}\eta, y-\mathrm{i}\eta)\big] \sim 2\int\int \frac{\eta\rho(\lambda)}{(x-\lambda)^2+\eta^2}\frac{\eta\tilde{\rho}(\tilde{\lambda})}{(y-\tilde{\lambda})^2+\eta^2}\Phi(\lambda,\tilde{\lambda})\mathrm{d}\lambda\mathrm{d}\tilde{\lambda}. \tag{4.29}$$

Finally, the inversion formula follows from Sokhotski–Plemelj identity

$$\lim_{\eta\to 0^+}\mathrm{Re}\big[\psi(x-\mathrm{i}\eta, y+\mathrm{i}\eta) - \psi(x-\mathrm{i}\eta, y-\mathrm{i}\eta)\big] \sim 2\pi^2\rho(x)\tilde{\rho}(y)\Phi(x,y). \tag{4.30}$$

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 (4.30) allows us to study the mean squared overlap (4.4) through the asymptotic behavior of the bivariate function $\psi(z,\tilde{z})$. Moreover, since we are able to control each entry of the resolvent of $\mathbf{E}$ and $\tilde{\mathbf{E}}$ (see Eq. (4.5)), the evaluation of Eq. (4.25) is immediate and leads to

$$\psi(z,\tilde{z}) \sim \frac{1}{z\tilde{z}}\frac{1}{N}\mathrm{Tr}\big[Z(z)(Z(z)-\mathbf{C})^{-1}\tilde{Z}(\tilde{z})(\tilde{Z}(\tilde{z})-\mathbf{C})^{-1}\big], \tag{4.31}$$

where $Z(z)$ is defined in (4.5) 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

$$\Big(Z(z)-\mathbf{C}\Big)^{-1}\Big(\tilde{Z}(\tilde{z})-\mathbf{C}\Big)^{-1} = \frac{1}{\tilde{Z}(\tilde{z})-Z(z)}\Big[\Big(Z(z)-\mathbf{C}\Big)^{-1} - \Big(\tilde{Z}(\tilde{z})-\mathbf{C}\Big)^{-1}\Big] \tag{4.32}$$

to obtain

$$\psi(z,\tilde{z}) \sim \frac{Z(z)\tilde{Z}(\tilde{z})}{z\tilde{z}}\frac{1}{\tilde{Z}(\tilde{z})-Z(z)}\frac{1}{N}\mathrm{Tr}\Big[\Big(Z(z)-\mathbf{C}\Big)^{-1} - \Big(\tilde{Z}(\tilde{z})-\mathbf{C}\Big)^{-1}\Big]. \tag{4.33}$$

From this last equation and using Marčenko–Pastur equation (3.9), we finally conclude that

$$\psi(z,\tilde{z}) \sim \frac{1}{\tilde{Z}(\tilde{z})-Z(z)}\left[\frac{\tilde{Z}(\tilde{z})}{\tilde{z}}\mathfrak{g}_{\mathbf{E}}(z) - \frac{Z(z)}{z}\mathfrak{g}_{\tilde{\mathbf{E}}}(\tilde{z})\right]. \tag{4.34}$$

One notices that Eq. (4.34) 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. (4.30) in order to retrieve the mean squared overlap (4.4). Before stating the main result of this section, we first rewrite (4.34) 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 $\mathbf{X}\mathbf{X}^* = \mathcal{W}$ and Eq. (3.33). Similarly, we define $\tilde{\mathbf{S}} = T^{-1}\tilde{\mathbf{X}}^*\mathbf{C}\tilde{\mathbf{X}}$ with $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^* = \tilde{\mathcal{W}}$. Using (3.33) and omitting the argument z and $\tilde{z}$, we can rewrite (4.34) as

$$\psi(z,\tilde{z}) \sim \frac{1}{q\tilde{q}z\tilde{z}}\left[\frac{(\tilde{q}z-q\tilde{z})\mathfrak{g}_{\tilde{\mathbf{S}}}^2}{\mathfrak{g}_{\mathbf{S}}-\mathfrak{g}_{\tilde{\mathbf{S}}}} + \frac{(q-\tilde{q})\mathfrak{g}_{\tilde{\mathbf{S}}}}{\mathfrak{g}_{\mathbf{S}}-\mathfrak{g}_{\tilde{\mathbf{S}}}}\right] + \frac{\mathfrak{g}_{\mathbf{S}}+\mathfrak{g}_{\tilde{\mathbf{S}}}}{q\tilde{z}} - \frac{1-q}{qz\tilde{z}}. \tag{4.35}$$

We see from (4.30) 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

$$m_0(\lambda) \equiv \lim_{\eta\to 0^+}\mathfrak{g}_{\mathbf{S}}(\lambda-\mathrm{i}\eta) = m_R(\lambda) + \mathrm{i}m_I(\lambda) \tag{4.36}$$

with

$$m_R(\lambda) = q\mathfrak{h}_{\mathbf{E}}(\lambda) + \frac{1-q}{\lambda}, \qquad m_I(\lambda) = q\rho_{\mathbf{E}}(\lambda) + (1-q)\delta_0, \tag{4.37}$$

where $\mathfrak{h}_{\mathbf{E}}$ is the Hilbert transform of $\rho_{\mathbf{E}}$. Note that this relation follows from Eq. (3.9). 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. (4.4) for any $\lambda\in\mathrm{supp}\,\varrho$ and $\tilde{\lambda}\in\tilde{\varrho}$ is given by (see [127] for a detailed derivation)

$$\Phi_{q,\tilde{q}}(\lambda,\tilde{\lambda}) = \frac{2(\tilde{q}\lambda-q\tilde{\lambda})\big[m_R|\tilde{m}_0|^2 - \tilde{m}_R|m_0|^2\big] + (\tilde{q}-q)\big[|\tilde{m}_0|^2 - |m_0|^2\big]}{\lambda\tilde{\lambda}\Big[(m_R-\tilde{m}_R)^2+(m_I+\tilde{m}_I)^2\Big]\Big[(m_R-\tilde{m}_R)^2+(m_I-\tilde{m}_I)^2\Big]}. \tag{4.38}$$

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 $\mathbf{C}$. One can easily check that $\tilde{m}_R = 1/\mu$ and $\tilde{m}_I = 0$. Hence, we deduce from (4.38) that

$$\Phi_{q,\tilde{q}=0}(\lambda,\mu) = \frac{q}{\lambda\mu\big[(m_R-1/\mu)^2+m_I^2\big]} = \frac{q\mu}{\lambda|1-\mu m_0(\lambda)|^2}, \tag{4.39}$$

which is another way to write (4.10) after applying the formula (3.33) in the limit $\eta \to 0^+$. It therefore shows that the result (4.38) generalizes Eq. (4.10) 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. (4.38) can be somewhat simplified to:

$$\Phi(\lambda, \tilde{\lambda}) = \frac{q(\lambda - \tilde{\lambda})\big(m_R(\lambda)|m_0(\tilde{\lambda})|^2 - m_R(\tilde{\lambda})|m_0(\lambda)|^2\big)}{\lambda\tilde{\lambda}\Big[(m_R - \tilde{m}_R)^2 + (m_I + \tilde{m}_I)^2\Big]\Big[(m_R - \tilde{m}_R)^2 + (m_I - \tilde{m}_I)^2\Big]}, \tag{4.40}$$

that becomes when $\tilde{\lambda} = \lambda$ [127],

$$\Phi(\lambda, \lambda) = \frac{q}{2\lambda^2} \frac{|m_0(\lambda)|^4 \partial_\lambda \big[m_R(\lambda)/|m_0(\lambda)|^2\big]}{m_I^2(\lambda)|\partial_\lambda m_0(\lambda)|^2}. \tag{4.41}$$

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 λ when they both come from the same population matrix $\mathbf{C}$. Any statistically significant deviation between this predicted overlap and empirical results can be interpreted as a violation of the hypothesis that the "true" population matrices corresponding to $\mathbf{E}$ and $\tilde{\mathbf{E}}$ are in fact different. This is extremely interesting from the point of view of applications, in particular to financial data where nothing ensures that $\mathbf{C}$ is time independent.

Now that we have all these theoretical results, let us now give some applications of the formula (4.40) as they will highlight that we can indeed find genuine information about the spectrum of $\mathbf{C}$ from the mean squared overlap (4.4). 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 2.2.3, 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 Eqs. (2.41) and (3.33) that $\mathfrak{g}_{\mathbf{S}}$ is given by

$$\mathfrak{g}_{\mathbf{S}}(z) = \frac{z + q - 1 - \mathrm{i}\sqrt{4zq - (z + q - 1)^2}}{2z} \tag{4.42}$$

for any $z \in \mathbb{C}_-$. It is easy to see using the definition (4.36) that we have

$$m_R(\lambda) = \frac{\lambda + q - 1}{2z}, \qquad m_I(\lambda) = \frac{\sqrt{4\lambda q - (\lambda + q - 1)^2}}{2\lambda}. \tag{4.43}$$

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. (4.41), we eventually get

$$\Phi_{q,q}(\lambda, \lambda) = 1, \tag{4.44}$$

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 (2.58) with $\kappa > 0$ defined in Eq. (2.54). As we saw in the previous section, the Stieltjes transform $\mathfrak{g}_{\mathbf{E}}(z)$ is explicit in this case (see Eq. (3.41)). Going back to Eq. (4.41), one can readily obtain from Eq. (3.41),

$$m_R(\lambda) = \frac{\lambda(1 + q\kappa) + q\kappa(1 - q)}{\lambda(\lambda + 2q\kappa)}, \qquad m_I(\lambda) = q\frac{\sqrt{\lambda - \lambda_-^{\mathrm{iw}}}\sqrt{\lambda_+^{\mathrm{iw}} - \lambda}}{\lambda(\lambda + 2q\kappa)}, \tag{4.45}$$

with $\lambda \in [\lambda_-^{\mathrm{iw}}, \lambda_+^{\mathrm{iw}}]$ where $\lambda_\pm^{\mathrm{iw}}$ is defined in (3.42). Plugging these expressions into Eq. (4.41) and after elementary computations, one finds

$$\Phi_{q,q}(\lambda, \lambda) = \frac{(1 + q\kappa)(\lambda + 2q\kappa)^2}{2q\kappa\big[2\lambda(1 + \kappa(1 + q)) - \lambda^2\kappa + \kappa(-1 + 2q(1 + q\kappa))\big]}. \tag{4.46}$$

The immediate consequence of this last formula is that in the presence of *anisotropic* correlations, the mean squared overlap (4.4) 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 [127]

$$\Phi(\lambda, \tilde{\lambda}) \underset{\kappa \to \infty}{\sim} \left[1 + \frac{(\lambda - 1)(\tilde{\lambda} - 1)}{2q^2\kappa} + \mathcal{O}(\kappa^{-2})\right], \tag{4.47}$$

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 $(2\kappa)^{-1} \to 0^+$ [127]. In the general case, we provide a numerical illustration

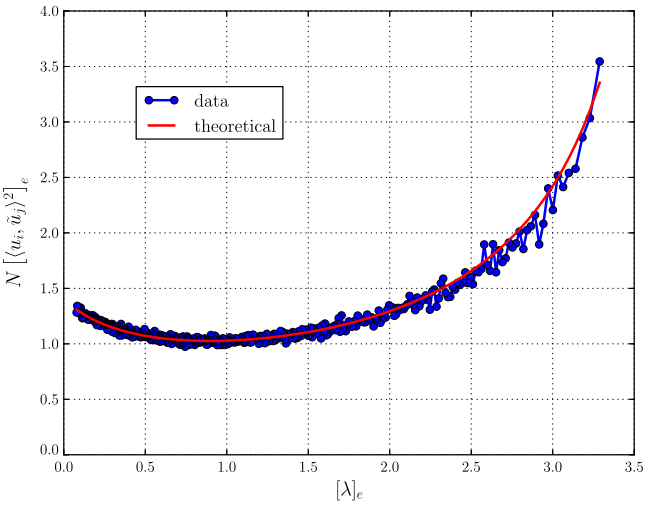


Fig. 14. Evaluation of $N\mathbb{E}\langle \mathbf{u}_i, \tilde{\mathbf{u}}_i\rangle^2$ with $N = 500$ and $q = \tilde{q} = 0.5$. The population matrix $\mathbf{C}$ is given by an Inverse-Wishart with parameter κ and the sample covariance matrices $\mathbf{S}$ and $\tilde{\mathbf{S}}$ are generated from a multivariate Gaussian distribution. The empirical average (blue points) is taken over 200 realizations and the theoretical prediction Eq. (4.41) (red line) is evaluated for all $[\lambda_i]_e$.

of this last statement in Fig. 14 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 (4.46) with the empirical average $[\langle \mathbf{u}_i, \tilde{\mathbf{u}}_i\rangle^2]_e$ taken over 200 realizations of $\mathbf{E}$ and we see that the agreement is again excellent. We therefore conclude that a possible application of (4.38) is to estimate directly the statistical texture of $\mathbf{C}$ using only sample eigenvectors: see Section 7 for an interesting example.

We now present an alternative derivation of $\Phi_{q,\tilde{q}}$ that uses the result of Section 4.1. 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 $\mathbf{V} \coloneqq [\mathbf{v}_1, \dots, \mathbf{v}_N]$. Hence, we may always write

$$\langle \mathbf{u}_i, \tilde{\mathbf{u}}_j\rangle = \left\langle \mathbf{u}_i\,,\ \left(\sum_{k=1}^{N} \mathbf{v}_k\mathbf{v}_k^*\right)\tilde{\mathbf{u}}_j \right\rangle = \sum_{k=1}^{N} \langle \mathbf{u}_i, \mathbf{v}_k\rangle\langle \mathbf{v}_k, \tilde{\mathbf{u}}_j\rangle \tag{4.48}$$

Using the results of Section 4.1, we rename the overlaps $\langle \mathbf{u}_i, \mathbf{v}_k\rangle = \sqrt{\Phi_q(\lambda_i, \mu_k)/N} \times \varepsilon(\lambda_i, \mu_k)$ where $\Phi_q(\lambda, \mu)$ is defined in (4.3) and $\varepsilon(\lambda, \mu)$ are random variables of unit variance. Hence, we have

$$\langle \mathbf{u}_i, \tilde{\mathbf{u}}_j\rangle = \frac{1}{N}\sum_{k=1}^{N} \sqrt{\Phi_q(\lambda_i, \mu_k)\Phi_{\tilde{q}}(\tilde{\lambda}_j, \mu_k)}\ \varepsilon(\lambda_i, \mu_k)\varepsilon(\tilde{\lambda}_j, \mu_k). \tag{4.49}$$

As noticed in [127], by averaging over the noise and making an "ergodic hypothesis" [130] – 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:

$$\Phi_{q,\tilde{q}}(\lambda_i, \tilde{\lambda}_j) = \frac{1}{N}\sum_{k=1}^{N} \Phi_q(\lambda_i, \mu_k)\Phi_{\tilde{q}}(\tilde{\lambda}_j, \mu_k). \tag{4.50}$$

It turns out that this expression is completely general and exactly equivalent to Eq. (4.40) if we replace the overlaps function Φ by (4.10). However, whereas this expression still contains some explicit dependence on the structure of the pure matrix $\mathbf{C}$, it has completely disappeared in Eq. (4.40). An interesting application of the formula (4.50) is when the spectrum of $\mathbf{E}$ (and $\tilde{\mathbf{E}}$) contains a finite number of outliers. Using the results (4.14) and (4.16) yields in the LDL and for $i \leqslant r$:

$$\Phi_{q,\tilde{q}}(\lambda_i, \tilde{\lambda}_i) \approx \mu_1^2 \frac{\theta'(\mu_1)\tilde{\theta}'(\mu_1)}{\theta(\mu_1)\tilde{\theta}(\mu_1)}, \tag{4.51}$$

where we recall that the function θ is defined in (3.62) and we define $\tilde{\theta}$ accordingly by replacing q with $\tilde{q}$. Note that we can express (4.51) in terms of observable variables by noticing that

$$\mu_1 = \frac{1}{\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1)}, \qquad \theta'(\mu_1) = \frac{-1}{\mathfrak{g}'_{\underline{\mathbf{S}}}(\theta(\mu_1))\mu_i^2}, \tag{4.52}$$

that we plug into (4.51) to conclude that

$$\Phi_{q,\tilde{q}}(\lambda_1,\tilde{\lambda}_1) \approx \frac{\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1)}{\lambda_1 \mathfrak{g}'_{\underline{\mathbf{S}}}(\lambda_1)} \frac{\mathfrak{g}_{\underline{\tilde{\mathbf{S}}}}(\lambda_1)}{\tilde{\lambda}_1 \mathfrak{g}'_{\underline{\tilde{\mathbf{S}}}}(\lambda_1)}. \tag{4.53}$$

This expression becomes even simpler when $q = \tilde{q}$ as it becomes

$$\Phi_{q,q}(\lambda_1,\tilde{\lambda}_1) \approx \left(\frac{\mathfrak{g}_{\underline{\mathbf{S}}}(\lambda_1)}{\lambda_1 \mathfrak{g}'_{\underline{\mathbf{S}}}(\lambda_1)}\right)^2. \tag{4.54}$$

One further deduces from (4.14) and (4.16) that for $i \leqslant r$, $\Phi_{q,\tilde{q}}(\lambda_i,\tilde{\lambda}_j) \sim \mathcal{O}(N^{-1})$ for any $j \neq i$.

5. Bayesian random matrix theory

We saw in the previous sections 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 $\rho_{\mathbf{E}}$ deviates significantly from the true spectrum whenever $q = \mathcal{O}(1)$. There has been many attempts in the literature to correct this "curse of dimensionality" using either heuristics or decision theoretic arguments (see Section 7.2 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 section, 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 we estimate "accurately" the matrix $\mathbf{C}$ in the high-dimensional regime knowing that the eigenvalues are systematically biased and the eigenvectors nearly completely unknown?

The aim of the present section 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. As explained in the introduction to this review, this theory has enjoyed much 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 Bayes (see e.g. [131] 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 section is to construct a Bayesian estimator for $\Xi(\mathbf{E})$. We therefore organize this section as follows. In the first part, we recall some basic results on Bayesian inference and introduce the estimator that will interest us. We then re-consider the famous "linear shrinkage" estimator, mentioned in Eq. (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 sections is applied to derive an optimal estimator for $\mathbf{C}$, which will turn out to be more efficient that all past attempts—see Section 8.

5.1. Bayes optimal inference: some basic results

5.1.1. Posterior and joint probability distributions

Bayesian theory allows one to answer, at least 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

$$\mathcal{P}(\mathbf{C}|\mathbf{Y}) = \frac{\mathcal{P}(\mathbf{Y}|\mathbf{C})\mathcal{P}(\mathbf{C})}{\mathcal{P}(\mathbf{Y})} \tag{5.1}$$

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, modeling 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 $\mathbf{C}$.
- $\mathcal{P}(\mathbf{Y})$ 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

$$\mathcal{P}(\mathbf{Y}) = \int \mathcal{D}\mathbf{C}\mathcal{P}(\mathbf{C})\mathcal{P}(\mathbf{Y}|\mathbf{C}). \tag{5.2}$$

Furthermore, we shall often use the concept of *joint* probability distribution defined by

$$\mathcal{P}(\mathbf{C}, \mathbf{Y}) = \mathcal{P}(\mathbf{Y}|\mathbf{C})\mathcal{P}(\mathbf{C}). \tag{5.3}$$

Thus, the two crucial inputs in a Bayesian model are 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.
$$\mathcal{P}(\mathbf{C}, \mathbf{Y}) = \mathcal{P}(\mathbf{Y}|\mathbf{C})\mathcal{P}(\mathbf{C}). \tag{5.4}$$
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 section.

5.1.2. 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 the estimator is from the true quantity $\mathbf{C}$. In general, this loss function is assumed to be a non-negative 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:

$$R^{\text{Bayes}}(\mathcal{L}(\mathbf{C}, \Xi(\mathbf{Y}))) := \left\langle \mathcal{L}(\mathbf{C}, \Xi(\mathbf{Y})) \right\rangle_{\mathcal{P}(\mathbf{C},\mathbf{Y})}, \tag{5.5}$$

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.,

$$\mathcal{L}^{L_2}(\mathbf{C}, \Xi(\mathbf{Y})) = \text{Tr}\left[(\mathbf{C} - \Xi(\mathbf{Y}))(\mathbf{C} - \Xi(\mathbf{Y}))^*\right]. \tag{5.6}$$

By using the fact that covariance matrices are symmetric and then applying Bayes rule, we see that

$$\begin{aligned} R^{\text{Bayes}} &= \left\langle \left\langle \text{Tr}\left[(\mathbf{C} - \Xi(\mathbf{Y}))^2\right] \right\rangle_{\mathcal{P}(\mathbf{Y}|\mathbf{C})} \right\rangle_{\mathcal{P}(\mathbf{C})} \\ &= \left\langle \left\langle \text{Tr}\left[(\mathbf{C} - \Xi(\mathbf{Y}))^2\right] \right\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{Y})} \right\rangle_{\mathcal{P}(\mathbf{Y})}, \end{aligned} \tag{5.7}$$

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 (5.6) is given by the *minimum mean squared error* (MMSE) condition, i.e.

$$\Xi^{\text{MMSE}} \equiv \Xi^{\text{MMSE}}(\mathbf{Y}) := \underset{\Xi(\mathbf{Y}) \in \mathcal{M}_N(\mathbf{Y})}{\text{argmin}} \left\langle \mathcal{L}^{L_2}(\mathbf{C}, \Xi(\mathbf{Y})) \right\rangle_{\mathcal{P}(\mathbf{C},\mathbf{Y})}, \tag{5.8}$$

Expanding (5.7), it is readily seen that the MMSE estimator is given by the posterior mean:

$$\Xi^{\mathrm{MMSE}} = \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{Y})}. \tag{5.9}$$

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.

5.2. 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,

$$\mathbb{P}(\mathbf{Y}|\mathbf{C}) = \frac{1}{(2\pi)^{\frac{NT}{2}} \det(\mathbf{C})^{\frac{T}{2}}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \sum_{i,j=1}^{N} Y_{it} \mathbf{C}^{-1}_{i,j} Y_{jt} \right\}. \tag{5.10}$$

It is easy to see that this is of the Boltzmann type, as in Eq. (2.1). More precisely, using the cyclic property of the trace operator one gets

$$\sum_{t=1}^{T} \sum_{i,j=1}^{N} Y_{it} \mathbf{C}^{-1}_{ij} Y_{jt} = \mathrm{Tr}\left[\mathbf{Y}\mathbf{C}^{-1}\mathbf{Y}^{*}\right] = T\mathrm{Tr}\left[\mathbf{E}\mathbf{C}^{-1}\right].$$

Thus, the N-variate Gaussian likelihood function can be written as

$$\mathcal{P}(\mathbf{Y}|\mathbf{C}) = \frac{1}{(2\pi)^{\frac{NT}{2}}} \exp\left\{ -\frac{T}{2} \mathrm{Tr}\big[\log(\mathbf{C}) + \mathbf{E}\mathbf{C}^{-1}\big] \right\} \equiv \mathcal{P}(\mathbf{E}|\mathbf{C}), \tag{5.11}$$

where we used Jacobi's formula $\det(\mathbf{A}) = \exp[\mathrm{Tr}\log\mathbf{A}]$ 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

$$\Xi^{\mathrm{MMSE}} \equiv \Xi^{\mathrm{MMSE}}(\mathbf{E}) := \langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}. \tag{5.12}$$

After a little thought, this set-up agrees perfectly with the framework developed in Sections 3 and 4. 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 3.2.1). Differently said, the Marčenko–Pastur equation (3.9) 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.

5.3. 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 (5.12). 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 distributions.

As an illustration, let us consider a warm-up 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

$$\mathcal{P}(\boldsymbol{\mu}|\mathbf{y}) = \mathcal{N}_N\left(\frac{\tau^2}{\tau^2+\sigma^2}\mathbf{y}, \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\mathbf{I}_N \right). \tag{5.13}$$

Therefore, the Bayes MMSE (5.9) of $\boldsymbol{\mu}$ is given by

$$\langle \boldsymbol{\mu} \rangle_{\mathcal{P}(\boldsymbol{\mu}|\mathbf{y})} = \left(1 - \frac{\sigma^2}{\sigma^2+\tau^2} \right)\mathbf{y}, \tag{5.14}$$

that is – loosely speaking – the celebrated James–Stein estimator [13]. 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. (5.11). Again, we will see that conjugate prior approach yields a very interesting result. Using the potential theory formalism introduced in (2.1) and in Section 2.2, it is easy to see from Eq. (5.11) that the potential function associated to a Gaussian likelihood function reads

$$V_q(\mathbf{E}, \mathbf{C}) = \frac{1}{2q}\left[\log(\mathbf{C}) + \mathbf{E}\mathbf{C}^{-1}\right], \tag{5.15}$$

that is clearly the Inverse-Wishart distribution encountered in (2.58) in the presence of an external field $\mathbf{E}$. Hence, let us introduce an inverse-Wishart ensemble with two hyper-parameters $\{\gamma, \kappa\}$ as a prior for $\mathbf{C}$[22]:

$$\mathcal{P}(\mathbf{C}) = Z \exp\left\{-N\mathrm{Tr}\left[\gamma \log \mathbf{C} + \kappa \mathbf{C}^{-1}\right]\right\},$$

with Z a normalization constant that depends on γ, κ and N. For simplicity, we impose that $\langle \mathbf{C} \rangle_{\mathcal{P}(\mathbf{C})} = \mathbf{I}_N$ and 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 (5.11), we find that the posterior distribution is also an inverse-Wishart distribution of the form:

$$\mathcal{P}(\mathbf{C}|\mathbf{E}) \propto \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[(T + \nu + N + 1)\log \mathbf{C} + T(2q\kappa \mathbf{I}_N + \mathbf{E})\mathbf{C}^{-1}\right]\right\}, \tag{5.16}$$

where we defined $\nu := N(2\kappa + 1) - 1$. As a consequence, we expect the Bayes estimator to be explicit like the James–Stein estimator (5.14) and the final result for Ξ^{MMSE} is obtained from (2.59):

$$\Xi^{\mathrm{MMSE}} = \frac{T}{T + \nu - N - 1}(2q\kappa \mathbf{I}_N + \mathbf{E}). \tag{5.17}$$

This estimator is known as the *linear shrinkage* estimator, first obtained in [16],

$$\Xi^{\mathrm{lin}} := \frac{T}{T + \nu - N - 1}(2q\kappa \mathbf{I}_N + \mathbf{E}) \approx \frac{1}{1 + 2q\kappa}\mathbf{E} + \frac{2q\kappa}{1 + 2q\kappa}\mathbf{I}_N + \mathcal{O}(T^{-1}), \tag{5.18}$$

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:

$$\Xi^{\mathrm{lin}} = \alpha_s \mathbf{E} + (1 - \alpha_s)\mathbf{I}_N \quad \text{where} \quad \alpha_s := \frac{1}{1 + 2q\kappa} \in [0, 1], \ \kappa > 0. \tag{5.19}$$

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 α_s. We give a simple illustration of how this estimator transforms the eigenvalues in Fig. 15. In particular, we see that small eigenvalues are lifted upwards while the top ones are pulled 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 κ (or directly α_s) in order to use this estimator in practice? In [16], Haff promoted an empirical Bayes approach similar to the work of James and Stein [13]. In the high-dimensional regime, Ledoit & Wolf [17] noticed that this approach may suffer from the fact that classical estimators become unreliable and consequently proposed a consistent estimator of α_s. There also exist more straightforward methods to estimate the parameter κ directly from the data, using RMT tools. We summarize all these approaches in Section 7.2.1.

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:

$$\mathcal{P}(\mathbf{C}) = Z \exp\left\{-N\mathrm{Tr}\left[\gamma \log \mathbf{C} + \kappa \mathbf{C}_0 \mathbf{C}^{-1}\right]\right\},$$

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:

$$\Xi^{\mathrm{lin}} = \alpha_s \mathbf{E} + (1 - \alpha_s)\mathbf{C}_0 \qquad \alpha_s \in [0, 1]. \tag{5.20}$$

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 are constant. This can also be seen as a spike correlation model, as was shown in (3.56), with $\underline{\mathbf{C}} = \mathbf{I}_N$, $r = 1$, $v_1 = (1, 1, \ldots, 1)$ and $d_1 = (N - 1)\rho$.

[22] More precisely, it is an inverse Wishart distribution $\mathcal{IW}_N(N, N(2\gamma - 1) - 1, 2N\kappa \mathbf{I}_N)$ defined in Eq. (2.58).

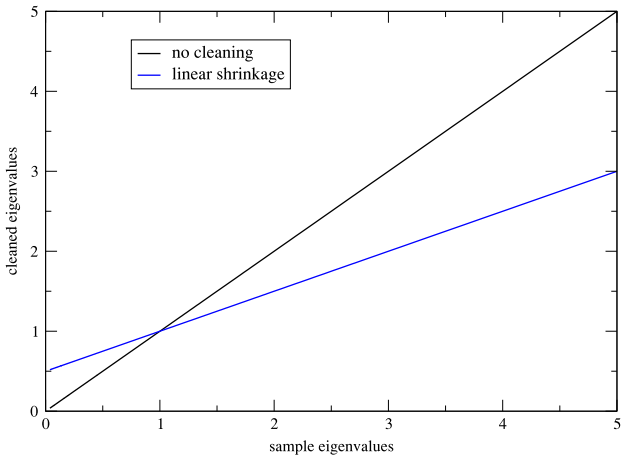


Fig. 15. Impact of the linear shrinkage (5.19) with $\alpha_s = 0.5$ on the eigenvalues (blue line), compared to the sample eigenvalues (black line). We see that the small eigenvalues are shifted upward and the large ones are pulled downward. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We now present the empirical Bayes approach through the "non-observable" James–Stein estimator (5.14). 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 (5.14), it is not hard to see that the evidence $\mathcal{P}(\mathbf{y})$, defined in (5.2), is given by

$$\mathcal{P}(\mathbf{y}) \sim \mathcal{N}_N(0, (\sigma^2 + \tau^2)\mathbf{I}_N). \tag{5.21}$$

Recall from (5.14) that our aim is to estimate the ratio $\sigma^2/(\sigma^2 + \tau^2)$ where σ^2 is known. To that end, we notice from (5.21) that

$$\|\mathbf{y}\|_2^2 \sim (\sigma^2 + \tau^2)\chi_N^2, \tag{5.22}$$

where $\|\cdot\|_2$ is the $\mathbb{L}_2$ norm and χ_N^2 is the chi-square distribution with N degrees of freedom. Therefore, we can conclude by maximum likelihood estimation that

$$\frac{\sigma^2 \times \max(N-2, 0)}{\|\mathbf{y}\|_2^2} \approx \frac{\sigma^2}{\sigma^2 + \tau^2}, \tag{5.23}$$

which yields an estimator of the unobservable term in Eq. (5.14). Hence, if we plug this sample estimate into (5.14), it yields the celebrated James–Stein estimator:

$$\hat{\boldsymbol{\mu}}_{\mathrm{JS}} = \left(1 - \frac{\sigma^2 \times \max(N-2, 0)}{\|\mathbf{y}\|_2^2}\right)\mathbf{y}, \tag{5.24}$$

that provides an improvement upon the maximum likelihood estimator of the mean of a Gaussian population whenever $N \geqslant 3$.

5.4. 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 κ. 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 section, 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. (2.1), i.e.

$$\mathcal{P}(\mathbf{C}) \propto \exp[-N \operatorname{Tr} V_0(\mathbf{C})] \tag{5.25}$$

where V_0 denotes the potential function. Therefore, it is easy to see that $\mathbf{C} \overset{\text{law}}{=} \boldsymbol{\Omega}\mathbf{C}\boldsymbol{\Omega}^*$ for any $N \times N$ orthogonal matrix $\boldsymbol{\Omega} \in \mathbf{O}(N)$. In other words, the eigenbasis of $\mathbf{C}$ is not biased in any specific direction. Moreover, using the Gaussian likelihood function (5.11), the posterior distribution reads:

$$\mathcal{P}(\mathbf{C}|\mathbf{E}) = \frac{1}{Z}\exp\Big[-N \operatorname{Tr} \mathcal{V}(\mathbf{C}, \mathbf{E})\Big], \qquad \mathcal{V}(\mathbf{C}, \mathbf{E}) := V_q(\mathbf{C}, \mathbf{E}) + V_0(\mathbf{C}), \tag{5.26}$$

where V_q is defined in Eq. (5.15). As a result, one can derive the identity:

$$\mathcal{P}(\mathbf{C}|\mathbf{E}) = \mathcal{P}(\boldsymbol{\Omega}\mathbf{C}\boldsymbol{\Omega}^*|\boldsymbol{\Omega}\mathbf{E}\boldsymbol{\Omega}^*), \tag{5.27}$$

Therefore, the Bayes MMSE estimator Eq. (5.9) obeys the following property:

$$\begin{aligned}\langle \mathbf{C}\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})} &= \int \boldsymbol{\Omega}\mathbf{C}'\boldsymbol{\Omega}^*\mathcal{P}(\boldsymbol{\Omega}\mathbf{C}'\boldsymbol{\Omega}^*|\mathbf{E})\mathcal{D}\mathbf{C}' \\ &= \boldsymbol{\Omega}\left[\int \mathbf{C}'\mathcal{P}(\mathbf{C}'|\boldsymbol{\Omega}^*\mathbf{E}\boldsymbol{\Omega})\mathcal{D}\mathbf{C}'\right]\boldsymbol{\Omega}^* \equiv \boldsymbol{\Omega}\langle \mathbf{C}\rangle_{\mathcal{P}(\mathbf{C}|\boldsymbol{\Omega}^*\mathbf{E}\boldsymbol{\Omega})}\boldsymbol{\Omega}^*\end{aligned} \tag{5.28}$$

where we changed variables $\mathbf{C} \to \boldsymbol{\Omega}\mathbf{C}'\boldsymbol{\Omega}^*$ and used Eq. (5.27) in the last step. Now we can always choose $\boldsymbol{\Omega} = \mathbf{U}$ such that $\mathbf{U}^*\mathbf{E}\mathbf{U}$ is diagonal. In this case, it is not difficult to convince oneself using symmetry arguments that $\langle \mathbf{C}\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{U}^*\mathbf{E}\mathbf{U})}$ 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 [132] and references therein:

$$\Xi^{\text{MMSE}} = \mathbf{U}\boldsymbol{\Gamma}(\Lambda)\mathbf{U}^*, \tag{5.29}$$

where $\mathbf{U} \in \mathbb{R}^{N\times N}$ is the eigenvectors of $\mathbf{E}$ and $\boldsymbol{\Gamma}(\boldsymbol{\Lambda}) = \text{diag}(\gamma_1(\boldsymbol{\Lambda}), \ldots, \gamma_N(\boldsymbol{\Lambda}))$ is a $N \times N$ diagonal matrix whose entries are functions of the sample eigenvalues $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$. 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 (5.19), 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 $\boldsymbol{\Gamma}(\boldsymbol{\Lambda})$, let us motivate the assumption of rotational invariance for the prior distribution of $\mathbf{C}$. Suppose that we have no prior information on possible privileged directions in the N-dimensional space that would allow one to bias the eigenvectors of the estimator Ξ^{MMSE} in these special directions. In this case, it makes sense that the only reasonable eigenbasis for our estimator Ξ^{MMSE} must be that the (noisy) observation $\mathbf{E}$ at our disposal. Any estimator satisfying Eq. (5.28) 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 some 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 [39,41] and Section 9 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 eigen decomposition (5.29) of the estimator Ξ^{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, one can see that the following identity holds:

$$\frac{1}{N}\text{Tr}\left[(z\mathbf{I}_N - \mathbf{E})^{-1}\langle \mathbf{C}\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{\gamma_i}{z - \lambda_i}, \tag{5.30}$$

which will allow us to extract the γ_i we are looking for, i.e. determine the optimal shrinkage function of the Bayes estimator (5.29). 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:

$$\begin{aligned}\text{Tr}\left[(z\mathbf{I}_N - \mathbf{E})^{-1}\langle \mathbf{C}\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}\right] &= \left\langle \text{Tr}\left[(z\mathbf{I}_N - \mathbf{E})^{-1}\langle \mathbf{C}\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}\right]\right\rangle_{\mathcal{P}(\mathbf{E})}, \\ &= \left\langle\left\langle \text{Tr}\left[(z\mathbf{I}_N - \mathbf{E})^{-1}\mathbf{C}\right]\right\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}\right\rangle_{\mathcal{P}(\mathbf{E})}.\end{aligned} \tag{5.31}$$

Using Bayes formula (5.1), we rewrite this last equation as

$$\begin{aligned}\text{Tr}\left[(z\mathbf{I}_N - \mathbf{E})^{-1}\langle \mathbf{C}\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}\right] &= \left\langle\left\langle \text{Tr}\left[(z\mathbf{I}_N - \mathbf{E})^{-1}\mathbf{C}\right]\right\rangle_{\mathcal{P}(\mathbf{E}|\mathbf{C})}\right\rangle_{\mathcal{P}(\mathbf{C})}, \\ &= \left\langle \text{Tr}\left[\left\langle(z\mathbf{I}_N - \mathbf{E})^{-1}\right\rangle_{\mathcal{P}(\mathbf{E}|\mathbf{C})}\mathbf{C}\right]\right\rangle_{\mathcal{P}(\mathbf{C})}.\end{aligned} \tag{5.32}$$

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 Sections 3 and 4. Therefore, since the eigenvalues $[\lambda_i]_i$ become deterministic in the limit $N \to \infty$ (see Section 3), we deduce that for large N

$$\frac{1}{N}\text{Tr}\left[(z\mathbf{I}_N - \mathbf{E})^{-1}\langle \mathbf{C}\rangle_{\mathcal{P}(\mathbf{C}|\mathbf{E})}\right] \approx \int \frac{\rho_{\mathbf{E}}(\lambda)\text{d}\lambda}{z - \lambda}\left\langle\sum_{j=1}^{N}\mu_j\,\Phi(\lambda, \mu_j)\right\rangle_{\mathbf{C}}, \tag{5.33}$$

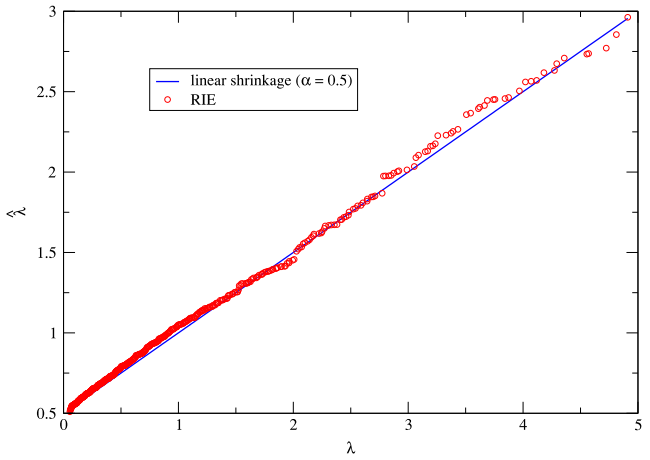


Fig. 16. Comparison of our analytical RI-Bayes estimator (5.34) (red dots) with the theoretical result Eq. (5.19) (blue line) when the prior distribution is an inverse Wishart (2.58). The parameters are $N = 500$, $q = 0.5$ and $\alpha_s = 0.5$.

where $\Phi(\lambda, \mu)$ is the mean squared overlap defined in Eq. (4.3). By comparing Eqs. (5.30) and (5.33), we can readily conclude that

$$\gamma(\Lambda) \equiv \gamma(\lambda) = \left\langle \sum_{j=1}^{N} \mu_j \Phi(\lambda, \mu_j) \right\rangle_{\mathbf{C}} \sim \int \mu \Phi(\lambda, \mu) \rho_{\mathbf{C}}(\mu) \mathrm{d}\mu, \tag{5.34}$$

where we used again an "ergodic hypothesis" [130] 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 γ of the Bayes estimator (5.29) that depends on the mean squared overlap, studied in Section 4, and the prior spectral density $\rho_{\mathbf{C}}$. Said differently the final result Eq. (5.34) is explicit but still seems to depend on the prior we choose for **C**. In fact, as we shall see in the next section, Eq. (5.34) can be estimated from the knowledge of **E** itself, i.e. without making any 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 **E** is enough to determine the correct prior ensemble to which **C** must belong.

We end this section with a self-consistency check in order to illustrate the result (5.34). As alluded to above, the nonlinear shrinkage function (5.34) generalizes the linear shrinkage (5.19). To highlight this, we assume that **C** is an isotropic Inverse Wishart matrices, such that the prior spectral density $\rho_{\mathbf{C}}$ is given by Eq. (2.53). We plot in Fig. 16 the eigenvalues we obtain using our Bayes estimator (5.19) (red dots) coming from a single realization of **E** with **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 and at the same time, of the RI-Bayes estimator (5.34) in this particular case. In Section 6.4.2, we will show explicitly that Eq. (5.33) reproduces Eq. (5.19) when **C** is an isotropic Inverse Wishart matrix.

6. Optimal rotational invariant estimator for general covariance matrices

6.1. Oracle estimator

In the previous section, we introduced a Bayesian framework to build an estimator of the population correlation matrix **C** using the data **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 to this topic. It was used successfully in many contexts as a simple way to provide robustness against the noise in high dimensional settings (see e.g. [11,16] or [133] 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 section, 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 **C** – hence the name "oracle" – but that one decides to create an estimator of **C** that is constrained to have a predetermined eigenbasis **U**. (In practice, this eigenbasis will be that of the sample correlation matrix **E**). What is the best one can do to estimate the true matrix **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 expressible 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 $\mathbf{U} = [\mathbf{u}_i]_{i\in[\![1,N]\!]}$. The optimal estimator of $\mathbf{C}$ in $\mathcal{M}(\mathbf{U})$ in the L_2 sense is given by:

$$\Xi^{\text{ora.}} = \underset{\Xi\in\mathcal{M}(\mathbf{U})}{\operatorname{argmin}} \left\| \Xi - \mathbf{C} \right\|_{L_2}^2. \tag{6.1}$$

It is trivial to find that the solution of this quadratic optimization problem, as:

$$\Xi^{\text{ora.}} = \sum_{i=1}^{N} \xi_i^{\text{ora.}} \mathbf{u}_i \mathbf{u}_i^*, \qquad \xi_i^{\text{ora.}} = \langle \mathbf{u}_i, \mathbf{C}\mathbf{u}_i \rangle. \tag{6.2}$$

This provides the best possible estimator of $\mathbf{C}$ given that we are "stuck" with the eigenbasis $[\mathbf{u}_i]_{i\in[\![1,N]\!]}$. The meaning of this estimator is better understood if we rewrite it a function of the eigenvectors of $\mathbf{C}$, to wit:

$$\xi_i^{\text{ora.}} = \sum_{j=1}^{N} \mu_j \langle \mathbf{u}_i, \mathbf{v}_j \rangle^2. \tag{6.3}$$

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 (6.2) explicitly uses the fact that the estimator lies in a wrong basis.

Coming back to our estimation of $\mathbf{C}$ 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 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 $[\xi_i^{\text{ora.}}]_{i\in[\![1,N]\!]}$ are indeed equivalent to the RI-Bayes MMSE formula (5.34), except that in Eq. (6.2), the population matrix $\mathbf{C}$ is a (deterministic) general covariance matrix. The equivalence between Bayes estimator (5.34) and unconditional estimator is not that surprising in the large N limit and has been mentioned in different contexts [133,134].

6.2. Explicit form of the optimal RIE

For practical purposes, the oracle estimator (6.2) 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 of "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 first for bulk eigenvalues, then for outliers — with the further surprise that the final expression is exactly the same in the two cases.

6.2.1. 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 [37]. More recently, this oracle estimator was considered in a more general framework [38] (including the case of additive noise models, see Appendix D) 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. (4.3) can be established.

More precisely, let us fix $i \geqslant r + 1$,[23] we expect that in the limit of large dimension, the squared 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, for large N, we have that for any $i > r$,

$$\xi_i^{\text{ora.}} = \sum_{j=1}^{N} \mu_j \, \Phi(\lambda_i, \mu_j) \approx \frac{1}{N\pi \, \rho_{\mathbf{E}}(\lambda_i)} \lim_{\eta\to 0^+} \operatorname{Im} \left[\sum_{j=1}^{N} \mu_j \, (z_i \mathbf{I}_N - \mathbf{E})^{-1}_{jj} \right], \tag{6.4}$$

where we have used the result Eq. (4.9) with $z_i = \lambda_i - \mathrm{i}\eta$. One finds using the Marčenko–Pastur relation (3.11) 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 $[\xi_i^{\text{ora.}}]_{i\in[\![r,N]\!]}$:

$$\xi_i^{\text{ora.}} \sim \hat{\xi}(\lambda_i) \qquad \text{with} \qquad \hat{\xi}(\lambda) := \frac{\lambda}{\left| 1 - q + q\lambda \lim\limits_{\eta\to 0^+} \mathfrak{g}_{\mathbf{E}}(\lambda - \mathrm{i}\eta) \right|^2}, \tag{6.5}$$

[23] Recall that the largest r eigenvalues are assumed to be outliers.

where $|\cdot|$ denotes the complex modulus. We notice that the RHS of this last equation does not involve the matrix $\mathbf{C}$ anymore and depends only on deterministic quantities. This is the "miracle" of the large N limit we alluded to above: the *a priori* non-observable oracle estimator converges to a deterministic quantity that may be estimated directly from the data.

6.2.2. 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 Section 3, this is not the case for outliers. Hence, the method of [37] and [38] are not valid anymore. Surprisingly, though, the final result happens to be identical to Eq. (6.5)! This has been established recently in [39] and the starting point of the method is to rewrite the oracle solution as

$$\xi_i^{\text{ora.}} = \sum_{j=1}^{r} \mu_j \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 + \sum_{j=r+1}^{N} \mu_j \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2, \tag{6.6}$$

from which we conclude, using also the results of Section 4, 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 $\mathcal{O}((N - r) \times 1/N) \sim \mathcal{O}(1)$.

We begin with the easy term which is the first one in the RHS of Eq. (6.6). Indeed, recall from Eq. (4.14) 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

$$\sum_{j=1}^{r} \mu_j \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 \sim \mu_i^2 \frac{\theta'(\mu_i)}{\theta(\mu_i)} \tag{6.7}$$

where we used Eq. (3.62) in the last step. The second term in Eq. (6.6) 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 4 and the result is given in Eq. (4.16) that we recall here 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 $r \ll N$ [39]

$$\sum_{j=r+1}^{N} \mu_j \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 \sim \frac{\mu_i^2}{\theta(\mu_i)} \frac{1}{T} \sum_{j=1}^{N} \frac{\mu_j^2}{(\mu_i - \mu_j)^2}, \tag{6.8}$$

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 (3.35). Indeed, by setting $z = \theta(\mu_i)$ with $i \leqslant r$ and θ defined in Eq. (3.62), Eq. (3.35), becomes

$$\theta(\mu_i) = \mu_i + \frac{1}{T} \sum_{j=1}^{N} \frac{1}{\mu_j^{-1} - \mu_i^{-1}} \tag{6.9}$$

and by taking the derivative with respect to μ_i, this yields

$$\frac{1}{T} \sum_{j=1}^{N} \frac{\mu_j^2}{(\mu_i - \mu_j)^2} = 1 - \theta'(\mu_i), \tag{6.10}$$

for any $i \leqslant r$. By plugging this identity into Eq. (6.8), we then obtain

$$\sum_{j=r+1}^{N} \mu_j \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 \sim \frac{\mu_i^2}{\theta(\mu_i)} \big(1 - \theta'(\mu_i)\big), \tag{6.11}$$

for any $i \leqslant r$. All in all, we see by plugging Eqs. (6.7) and (6.11) into Eq. (6.6) that we finally get

$$\xi_i^{\text{ora.}} \sim \frac{\mu_i^2}{\theta(\mu_i)}, \tag{6.12}$$

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. (3.62), we can rewrite the RHS of Eq. (6.12) as a function of the sample eigenvalues. Firstly, one notices that $\theta(\mu_i) = \lambda_i$ for $N \to \infty$ thanks to Eq. (3.62). Moreover, we can also invert Eq. (3.62) 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. (3.33) in the last step. Therefore, we deduce that in the high dimensional limit, we can rewrite Eq. (6.12) as

$$\xi_i^{\text{ora.}} \sim \frac{\lambda_i}{\left|1 - q + q\lambda_i \mathfrak{g}_{\underline{\mathbf{E}}}(\lambda_i)\right|^2}. \tag{6.13}$$

We see that the result is similar to the result for the bulk eigenvalues except that for outliers, we need the Stieltjes 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 [129] that we can replace it by the Stieltjes transform of $\mathbf{E}$ so that we finally conclude that for any outlier $i \leqslant r$,

$$\xi_i^{\text{ora.}} \sim \hat{\xi}(\lambda_i), \tag{6.14}$$

where the optimal shrinkage function $\hat{\xi}$ is defined in (6.5). We see that the outliers of oracle estimator also converge to a deterministic function which is exactly the same than for bulk eigenvalues (6.5) 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 section. 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. (6.5) or (6.14) and only requires the knowledge of the Stieltjes transform of $\mathbf{E}$, which is observable—see below.

6.3. Some properties of the "cleaned" eigenvalues

Even though the optimal nonlinear shrinkage function (6.26) seems relatively simple, it is not immediately clear what is the effect induced by the transformation $\lambda_i \to \hat{\xi}(\lambda_i)$. 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 consider the moments of the spectrum of $\Xi^{\text{ora.}}$. From Eq. (6.3) we immediately derive that:

$$\text{Tr}\Xi^{\text{ora.}} = \sum_{j=1} \mu_j \mathbf{v}_j^* \left(\sum_{i=1} \mathbf{u}_i \mathbf{u}_i^* \right) \mathbf{v}_j = \text{Tr}\mathbf{C}, \tag{6.15}$$

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, if 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 square matrix with non-negative 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 are 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

$$\text{Tr}(\Xi^{\text{ora.}})^2 \leqslant \text{Tr}\mathbf{C}^2 \leqslant \text{Tr}\mathbf{E}^2, \tag{6.16}$$

where the last inequality comes from Eq. (3.17). 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 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 top (resp. small) eigenvalues compared to their "true" locations μ_i for any $i \in [\![1, N]\!]$, except for the trivial case $\mathbf{C} = \mathbf{I}_N$. As a consequence, estimating the population eigenvalues $[\mu_i]_{i \in [\![1,N]\!]}$ is *not* what one should do to obtain an optimal estimator

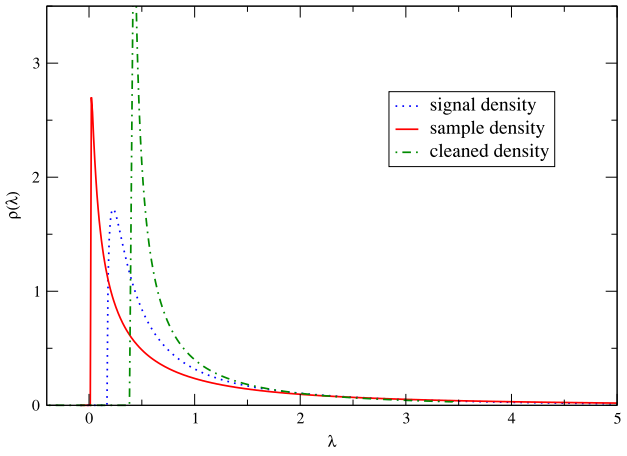


Fig. 17. Evaluation of the eigenvalue density of the signal, sample and cleaned density for $q = 0.5$ when the prior is an inverse Wishart of parameter $\kappa = 1$. We see that the cleaned density is the narrowest one, while the sample is the widest, as expected.

of $\mathbf{C}$ when there is only partial information about its eigenvectors. We provide an illustration in Fig. 17 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. (6.5) and (6.14) 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 supp $\rho_{\mathbf{E}}$ and let us assume $q < 1$ so that $\mathbf{E}$ has no exact zero mode.[24] We know since Section 6.2.2 that the estimator (6.5) 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. (3.23) that $\lambda \mathfrak{g}_{\mathbf{E}}(\lambda) = \mathcal{O}(\lambda)$ for $\lambda \to 0^+$. This allows us to conclude from Eq. (6.5) that for very small outliers,

$$\lim_{\lambda \to 0^+} \hat{\xi}(\lambda) = \frac{\lambda}{(1-q)^2} + \mathcal{O}(\lambda^2), \tag{6.17}$$

which is in agreement with Eq. (6.16): small eigenvalues are enhanced for $q \in (0, 1)$.

The other asymptotic limit $\lambda \to \infty$ is also useful since it gives us the behavior of the nonlinear shrinkage function $\hat{\xi}$ for large outliers. In that case, we know from Eq. (3.16) that $\lim_{\lambda \uparrow \infty} \lambda \mathfrak{g}_{\mathbf{E}}(\lambda) \sim 1 + \lambda^{-1} \varphi(\mathbf{E})$, where φ denotes the normalized trace operator (2.61). Therefore, we conclude that

$$\lim_{\lambda \to \infty} \hat{\xi}(\lambda) \approx \frac{\lambda}{\left(1 + q\lambda^{-1}\varphi(\mathbf{E}) + \mathcal{O}(\lambda^{-2})\right)^2} \sim \lambda - 2q\varphi(\mathbf{E}) + \mathcal{O}(\lambda^{-1}), \tag{6.18}$$

and if we use that $\operatorname{Tr} \mathbf{E} = \operatorname{Tr} \mathbf{C} = N$, we simply obtain

$$\lim_{\lambda \to \infty} \hat{\xi}(\lambda) \approx \lambda - 2q + \mathcal{O}(\lambda^{-1}). \tag{6.19}$$

It is interesting to compare this with the well-known "Baik–Ben Arous–Péché" (BBP) result on large outliers [121], which reads (see Eq. (3.64)) $\lambda \approx \mu + q$ for $\lambda \to \infty$. As a result, we deduce from Eq. (6.19) that $\hat{\xi}(\lambda) \approx \mu - q$ and we therefore find the following ordering relation

$$\hat{\xi}(\lambda) < \mu < \lambda, \tag{6.20}$$

for an isolated and large eigenvalues λ and for $q > 0$. Again, this result is in agreement with Eq. (6.16): large eigenvalues should be reduced for any $q > 0$, even below the "true" value of the outlier μ. More generally, the non-linear shrinkage function $\hat{\xi}$ interpolates smoothly between $\lambda/(1-q)^2$ for small λ's to $\lambda - 2q$ for large λ'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 (6.5) is monotonic with respect to the sample eigenvalues.

[24] Recall that we assume $\mathbf{C}$ to be positive definite for the sake of simplicity.

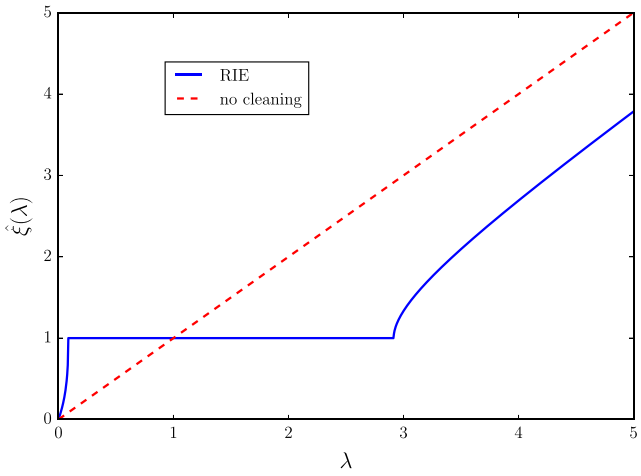


Fig. 18. Evaluation of the optimal RIE's eigenvalues for $\mathbf{C} = I_N$ as a function of the sample eigenvalues $[\lambda_i]_{i\in[\![1,N]\!]}$ for $q = 1/2$. The nonlinear shrinkage function is plotted with the plain blue line. We see that for $\lambda > (1+\sqrt{q})^2$, a phase transition occurs and the corresponding "cleaned" eigenvalues converge for large λ to $\lambda - 2q$ (the red dotted line shifted down by $2q = 1$). Note the square-root singularity of the estimator as one gets close to the edge of the spectrum. There is a similar phase transition for outliers $\lambda < (1-\sqrt{q})^2$ (see Fig. 20).

6.4. Some analytical examples

The general properties of the oracle shrinkage procedure described above can be given more 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.

6.4.1. 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 [121], that leads to a non-trivial shrinkage formula.

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 λ_i with $i \leqslant r$. Hence, the estimator is easily obtained by plugging the Stieltjes transform (2.41) into Eq. (6.5), with a result shown in Fig. 18.

For bulk eigenvalues, the computation can be done more explicitly. First, using Eq. (2.41), 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 $\lambda \in \left[(1-\sqrt{q})^2, (1+\sqrt{q})^2\right]$, 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

$$\lim_{\eta\to 0}\left|1 - q + q\lambda\mathfrak{g}_{\mathbf{E}}(\lambda - \mathrm{i}\eta)\right|^2 = \frac{(z+1-q)^2 + \left(4\lambda q - (\lambda + q - 1)^2\right)}{4},$$

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

$$\hat{\xi}(\lambda) = 1, \qquad \lambda \in \left[(1-\sqrt{q})^2, (1+\sqrt{q})^2\right]. \tag{6.21}$$

We provide an illustration of this phase transition in Fig. 18 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. (6.19).

6.4.2. Revisiting the linear shrinkage

In Section 5, 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 κ. We want to revisit this result within the framework of the present section, and we will see that in the presence of extra spikes, the optimal shrinkage function (6.5) again shows a phase

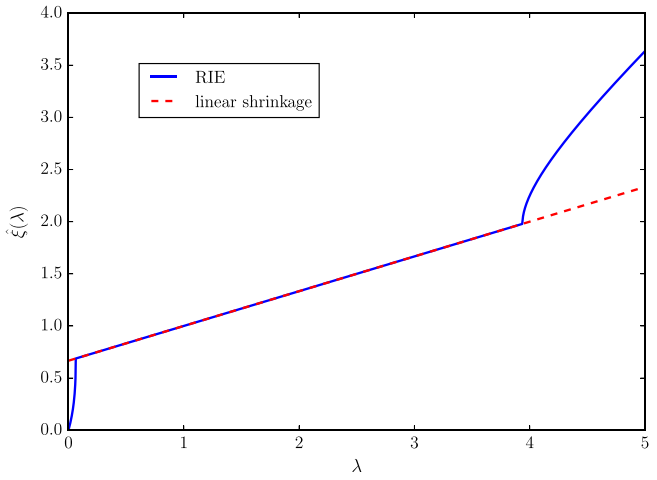


Fig. 19. Evaluation of the optimal RIE's eigenvalues for an Inverse Wishart prior with $\kappa = 2$ as a function of the sample eigenvalues $[\lambda_i]_{i\in[\![1,N]\!]}$. The matrix $\mathbf{E}$ is generated using Wishart matrix with parameter $N = 500$ and $q = 0.5$. The nonlinear shrinkage function is plotted with the plain blue line and it coincides with the estimator Eq. (5.19) (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 Fig. 20).

transition phenomenon and therefore differs from the linear estimator Eq. (5.19) 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 numerical result is immediately obtained from Eqs. (6.5) and (3.41). For the bulk component, the square root term in Eq. (3.41) becomes imaginary. Hence, setting $z = \lambda - \mathrm{i}\eta$ into Eq. (3.41) with $\lambda \in [\lambda_-^{\text{iw}}, \lambda_+^{\text{iw}}]$ and $\lambda_\pm^{\text{iw}}$, defined in Eq. (3.42), one obtains

$$\left|1 - q + q\lambda \lim_{\eta\to 0^+} \mathfrak{g}_{\mathbf{E}}(\lambda - \mathrm{i}\eta)\right|^2 = \frac{\left[\lambda(1+q\kappa) + \kappa q(1-q)\right]^2 + q^2\left[2\lambda\kappa(\kappa(1+q)+1) - \kappa^2(1-q)^2 - \lambda^2\kappa^2\right]}{(\lambda + 2q\kappa)^2},$$

with $\kappa > 0$. This can be rewritten after expanding the square as

$$\left|1 - q + q\lambda \lim_{\eta\to 0^+} \mathfrak{g}_{\mathbf{E}}(\lambda - \mathrm{i}\eta)\right|^2 = \frac{\lambda(1+2q\kappa)}{(\lambda+2q\kappa)}. \tag{6.22}$$

By plugging this last equation into Eq. (6.5) gives for any $\lambda \in [\lambda_-^{\text{iw}}, \lambda_+^{\text{iw}}]$

$$\xi^{\text{ora.}}(\lambda) = \frac{\lambda + 2q\kappa}{1 + 2q\kappa}, \tag{6.23}$$

and if we recall the definition $\alpha_s = 1/(1+2q\kappa) \in [0, 1]$ of Eq. (5.19), we retrieve exactly the linear shrinkage estimator (5.19),

$$\xi^{\text{ora.}}(\lambda) \sim \alpha_s\lambda + (1-\alpha_s), \qquad \lambda \in [\lambda_-^{\text{iw}}, \lambda_+^{\text{iw}}]. \tag{6.24}$$

This last result illustrates in a particular case the genuine link between the optimal RIE $\Xi^{\text{ora.}}$ and Bayes optimal inference techniques in the LDL. In particular, we show that for an isotropic Inverse Wishart matrix, the estimator $\Xi^{\text{ora.}}$ 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 Fig. 19 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 [38] where the case of an additive noise is also considered—see Appendix D, yielding a generalization of the well-known Wiener's signal-to-noise ratio optimal estimator [135].

We also illustrate in Fig. 20 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 (6.17) is pretty accurate. This prediction becomes less and less effective as λ moves closer to the left edge.

6.5. 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 must be solved numerically. In that case, the main question is to estimate the function $\mathfrak{g}_{\mathbf{E}}(z)$ without imposing any "prior"

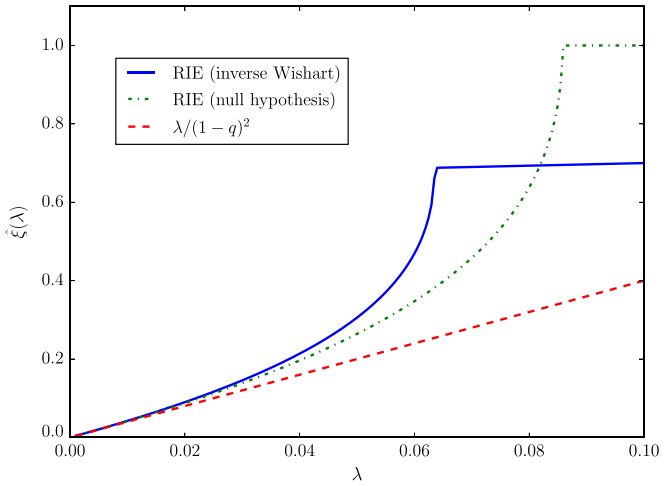


Fig. 20. Comparison of the prediction Eq. (6.17) (red dashed line) compared to the analytical solution of the null hypothesis (6.21) (green dash-dotted line) and the Inverse Wishart prior (6.24) with parameter $\kappa = 2$ (blue plain line). In both cases, wet set $q = 0.5$. The asymptotic prediction (6.17) becomes less and less accurate as λ moves closer to the left edge and the analytic solution (blue line) depicts a phase transition.

on $\mathbf{C}$. Indeed, even though the function $\xi^{\text{ora.}}$ only depends on observable quantities, we still need to estimate the function $\mathfrak{g}_{\mathbf{E}}(z)$ using only a finite (and random) set of sample eigenvalues.

This question has been addressed recently in [39], where apart from extending the result of [37] to outliers (as reviewed above), the mathematical technique used in [39] provides a derivation of Eq. (6.5) at a *local* scale and for any large but finite N. As alluded to in Section 4, 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 [39] can be summarized as follows: the limiting Stieltjes transform $\mathfrak{g}_{\mathbf{E}}(z)$ can be replaced by its discrete form

$$\mathfrak{g}_{\mathbf{E}}^N(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z-\lambda_i}, \tag{6.25}$$

with *high probability* (see e.g. [113] 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 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,

$$\left|\xi_i^{\text{ora.}} - \hat{\xi}_i^N\right| \leqslant \frac{K}{\sqrt{T}}, \qquad \hat{\xi}_i^N \equiv \hat{\xi}^N(\lambda_i) := \frac{\lambda_i}{\left|1-q+qz_i\mathfrak{g}_{\mathbf{E}}^N(z_i)\right|^2}, \tag{6.26}$$

provided that λ_i is not near zero [39]. We see that Eq. (6.26) 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 (6.26) 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",

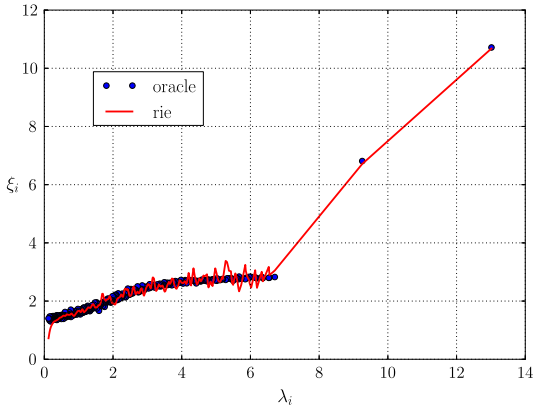
$$\rho_{\mathbf{C}} = 0.002\delta_{15} + 0.002\delta_8 + 0.396\delta_3 + 0.3\delta_{1.5} + 0.3\delta_1. \tag{6.27}$$

(ii) Deformed GOE, i.e $\mathbf{C} = I_N + \text{GOE}$ (of width $\sigma = 0.2$) with extra spikes located at $\{3, 3.5, 4.5, 6\}$.

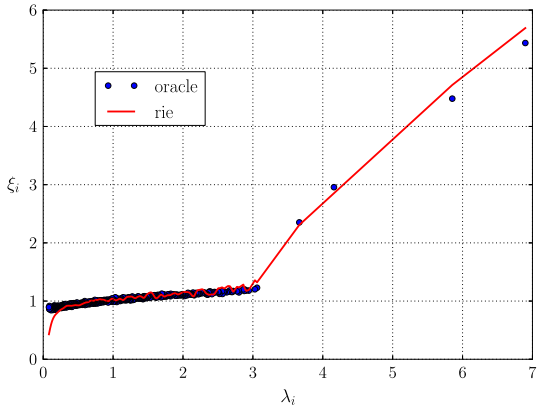
(iii) Toeplitz matrix with entries $\mathbf{C}_{ij} = 0.6^{|i-j|}$ with spikes located at $\{7, 8, 10, 11\}$;

(iv) Power-law distributed eigenvalues (see [29] and Section 3) with $\lambda_0 = -0.6$ (or $\lambda_{\min} = 0.8$). Using a large N proxy for the classical positions of the μ_i, one gets [29]:

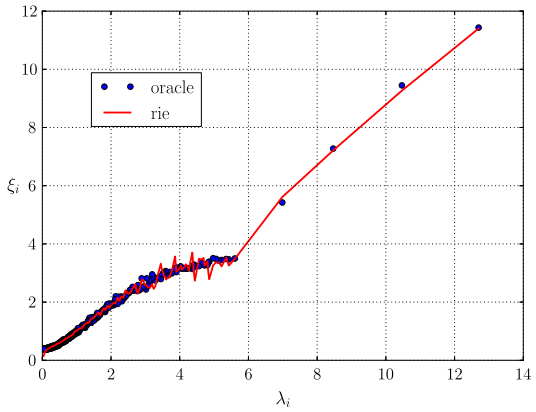
$$\mu_i = -\lambda_0 + \frac{(1+\lambda_0)}{2}\sqrt{\frac{N}{i}} \qquad i \in [\![1, N]\!]. \tag{6.28}$$



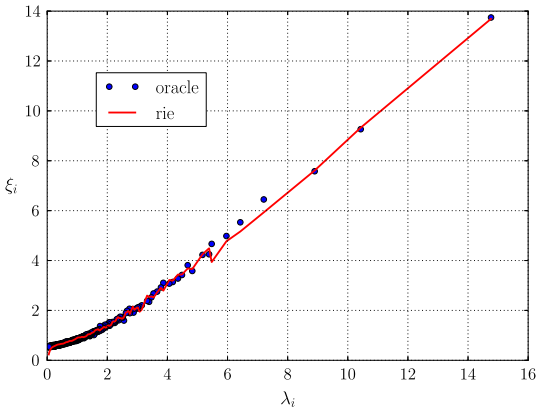
(a) Multiple sources (case (i)).



(b) Deformed GOE (case (ii)).



(c) Toeplitz (case (iii)).



(d) Power law (case (iv)).

Fig. 21. Comparison of numerically estimated oracle estimator (6.26) (red line) with the exact oracle RIE estimator (6.2) (blue points) for the four cases presented at the beginning of Section 6.5 with $N = 500$ and $T = 1000$. The results come from a single realization of $\mathbf{E}$ using a multivariate Gaussian measurement process.

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 $\mathbf{C}$ remains bounded. We plot the results obtained with the estimator Eq. (6.26) and the oracle estimator Eq. (6.2) in Fig. 21.

Overall, the estimator (6.26) 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 (6.26) matches again well with the oracle estimator. We nevertheless notice that the small eigenvalues are *systematically* underestimated by the empirical optimal RIE (6.26). This effect will be investigated in more details in Section 8.

As a further check, we provide here a numerical test of the "optimal" scale η. As explained above, it was shown in [39] that the value $\eta = N^{-1/2}$ gives the upper bound in (6.26). However, one might wonder if this value is indeed optimal with real (or synthetic) data. To test this, we study the estimator (6.26) as a function of η and compute the corresponding mean squared error with respect to the oracle estimator $\Xi^{\text{ora.}}$ for $\eta = \alpha N^{-1/2}$ and $\alpha \in [0.01, 50]$. For each $\mathbf{C}$, we evaluate the error for 100 different realizations of $\mathbf{E}$ using a multivariate Gaussian process. The results are reported in Fig. 22. The optimal value of $\alpha \approx 1.5$ for all the examples except when $\mathbf{C}$ is a Toeplitz matrix (yellow dots) where the optimal value of $\alpha \approx 8.4$.

6.6. Extension to the free multiplicative model

As highlighted in [38], the evaluation of the optimal RIE for bulk eigenvalues can be extended to more general multiplicative random matrix models (for additive noise models, see Appendix D). In particular, it is possible to derive (formally) the optimal nonlinear shrinkage function (6.5) for the bulk eigenvalues of the measurement model (2.80) which generalizes the case of sample covariance matrices (see Section 3.2.1).

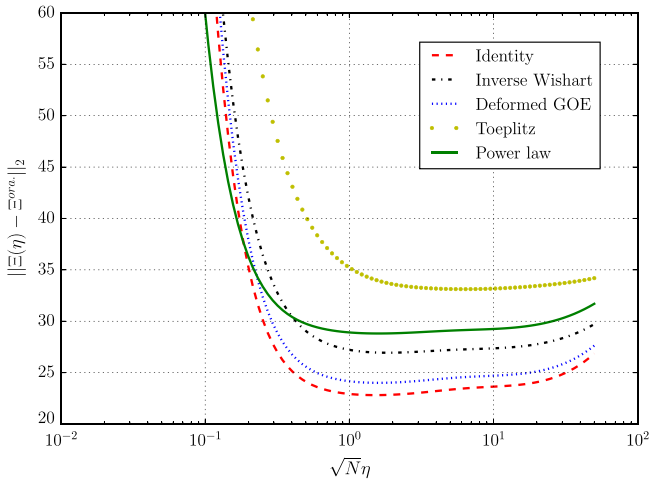


Fig. 22. Mean squared difference between the optimal estimator (6.26) and the oracle estimator. The estimator (6.26) is now studied as a function of η. The x-axis (in logarithm scale) shows the value of $\alpha = \sqrt{N}\eta$ for the sake of clarity. We consider five different examples for $\mathbf{C}$ (same configuration as in Fig. 21 and the identity matrix). For each example, we generate 100 independent realizations of $\mathbf{E}$ with $N = 500$ and $T = 1000$.

To that end, let us define $\mathbf{M} \coloneqq \mathbf{C}^{1/2}\boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^*\mathbf{C}^{1/2}$ where $\mathbf{B}$ is a $N \times N$ symmetric rotational invariant noise term and $\boldsymbol{\Omega}$ is a $N \times N$ random rotation matrix that is distributed according to the Haar measure. One can easily check from Eq. (2.100) that

$$\mathrm{Tr}\left[\mathbf{G}_{\mathbf{M}}(z)\mathbf{C}\right] = N(z\mathfrak{g}_{\mathbf{M}}(z) - 1)\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1). \tag{6.29}$$

Using the analyticity of the $\mathcal{S}$-transform, we define the function $\gamma_{\mathbf{B}}$ and $\omega_{\mathbf{B}}$ such that:

$$\lim_{z\to\lambda-\mathrm{i}0^+} \mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1) \coloneqq \gamma_{\mathbf{B}}(\lambda) + \mathrm{i}\pi\,\rho_{\mathbf{M}}(\lambda)\omega_{\mathbf{B}}(\lambda), \tag{6.30}$$

and as a result, the optimal RIE for bulk eigenvalues of the free multiplicative noise model (2.80) may be inferred from (6.4):

$$\xi_i^{\mathrm{ora.}} \sim F_2(\lambda_i); \qquad F_2(\lambda) = \lambda\gamma_{\mathbf{B}}(\lambda) + (\lambda\mathfrak{h}_{\mathbf{M}}(\lambda) - 1)\omega_{\mathbf{B}}(\lambda). \tag{6.31}$$

Note that one retrieves the estimator (6.5) by plugging Eqs. (2.44) and (6.30) into Eq. (6.31). We omit details, which can be found in [38], and we conclude that the formula (6.31) indeed generalizes Eq. (6.5). 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. (6.31) (see Section 9 for some relevant applications of this model).

We emphasize in passing that we may also derive the mean squared overlap (4.3) in the bulk of the distribution using Eq. (2.100). To that end, we invoke the relation (4.9) and Eq. (2.100) to obtain [38]:

$$\Phi(\lambda, \mu) = \frac{\mu\beta_m(\lambda)}{(\lambda - \mu\alpha_m(\lambda))^2 + \pi^2\mu^2\beta_m(\lambda)^2\rho_{\mathbf{M}}(\lambda)^2}, \tag{6.32}$$

where we defined the functions α_m and β_m as

$$\begin{cases} \alpha_m(\lambda) \coloneqq \lim\limits_{z\to\lambda-i0^+} \mathrm{Re}\left[\dfrac{1}{\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1)}\right] \\ \beta_m(\lambda) \coloneqq \lim\limits_{z\to\lambda-i0^+} \mathrm{Im}\left[\dfrac{1}{\mathcal{S}_{\mathbf{B}}(z\mathfrak{g}_{\mathbf{M}}(z) - 1)}\right] \dfrac{1}{\pi\,\rho_{\mathbf{M}}(\lambda)}, \end{cases} \tag{6.33}$$

and the subscript m stands for "multiplication".

We conclude this technical section by mentioning one 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 (6.31) remains *universal* (as we believe it is) in the sense that the cleaning formula for bulk eigenvalues and outliers is identical. The block matrix representation (C.8) might be useful in that respect.

7. Application: Markowitz portfolio theory and previous "cleaning" schemes

7.1. Markowitz optimal portfolio theory

For the reader not familiar with Markowitz's optimal portfolio theory [5], 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

$$\langle Y_{it} Y_{jt'} \rangle = \mathbf{C}_{ij}\delta_{tt'}, \tag{7.1}$$

where the absence of correlations in the time direction is 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

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^N} \dfrac{1}{2}\mathbf{w}^*\mathbf{C}\mathbf{w} \\ \text{s.t. } \mathbf{w}^*\mathbf{g} \geq \mathcal{G} \end{cases} \tag{7.2}$$

where $\mathbf{g}$ is a N-dimensional vector of predictors (assumed to be deterministic and given by, e.g. in depth analysis of economic data) and $\mathcal{G}$ is the expected gain. This mathematical problem can be easily solved by introducing a Lagrangian multiplier γ to rewrite this constrained optimization problem as an unconstrained one[25]:

$$\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{2}\mathbf{w}^*\mathbf{C}\mathbf{w} - \gamma\mathbf{w}^*\mathbf{g}. \tag{7.3}$$

Assuming that $\mathbf{C}$ is invertible, it is not hard to find the optimal solution and the value of γ such that overall expected return is exactly $\mathcal{G}$. It is given by

$$\mathbf{w}_\mathbf{C} = \mathcal{G}\frac{\mathbf{C}^{-1}\mathbf{g}}{\mathbf{g}^*\mathbf{C}^{-1}\mathbf{g}}, \tag{7.4}$$

that requires the knowledge of both $\mathbf{C}$ and $\mathbf{g}$, which are *a priori* unknown. As mentioned above, forming expectations of future returns is the job of the investor or of the financial analyst, based on his/her information and anticipations, so we assume that $\mathbf{g}$ is given. Even if these predictions were completely wrong, it would still make 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. (7.4). 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?[26] If one knew the population correlation matrix, $\mathbf{C}$, the *true* optimal risk associated $\mathbf{w}_\mathbf{C}$ would be given by

$$\mathcal{R}^2_{\text{true}} := \langle \mathbf{w}_\mathbf{C}, \mathbf{C}\mathbf{w}_\mathbf{C}\rangle = \frac{\mathcal{G}^2}{\mathbf{g}^*\mathbf{C}^{-1}\mathbf{g}}. \tag{7.5}$$

However, the optimal strategy (7.4) 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?

7.1.1. Predicted and realized risk

One very naive way to use the Markowitz optimal portfolio is to apply (7.4) using the empirical matrix $\mathbf{E}$ instead of $\mathbf{C}$. Recalling the results of Sections 3 and 4, 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:

$$\mathbf{w}_\mathbf{E} = \mathcal{G}\frac{\mathbf{E}^{-1}\mathbf{g}}{\mathbf{g}^*\mathbf{E}^{-1}\mathbf{g}}, \tag{7.6}$$

and the minimum risk associated to this portfolio is thus given by

$$\mathcal{R}^2_{\text{in}} = \langle \mathbf{w}_\mathbf{E}, \mathbf{E}\,\mathbf{w}_\mathbf{E}\rangle = \frac{\mathcal{G}^2}{\mathbf{g}^*\mathbf{E}^{-1}\mathbf{g}}, \tag{7.7}$$

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

$$\mathbb{E}[\mathbf{g}^*\mathbf{E}^{-1}\mathbf{g}] \geqslant \mathbf{g}^*\mathbb{E}\big[\mathbf{E}\big]^{-1}\mathbf{g} = \mathbf{g}^*\mathbf{C}^{-1}\mathbf{g} \tag{7.8}$$

25 One can check that the so-called Karush–Kuhn–Tucker conditions are satisfied.

26 An equivalent risk measure is the volatility which is simply the square root of the variance of the portfolio strategy.

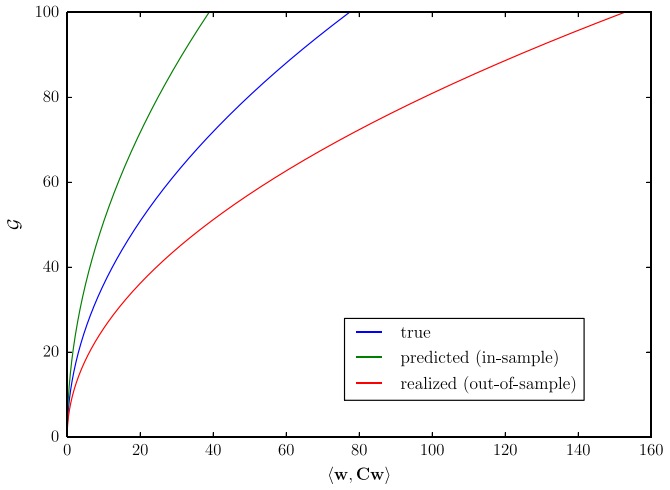


Fig. 23. Efficient frontier associated to the mean–variance optimal portfolio (7.4) for $\mathbf{g} = (1, \ldots, 1)^*$ and $\mathbf{C}$ a shifted GOE around the identity matrix, with $\sigma = 0.2$ and for $q = 0.5$. The blue line depicts the expected gain as a function of the *true* optimal risk (7.5) in percentage. The green line gives the predicted (in-sample) risk while the red line gives the realized (out-of-sample) risk, which is well above the true risk. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 by $\mathbf{E}'$ the empirical matrix of this out-of-sample period; the *out-of-sample* risk is then naturally defined by:

$$\mathcal{R}^2_{\text{out}} = \langle \mathbf{w}_{\mathbf{E}}, \mathbf{E}'\, \mathbf{w}_{\mathbf{E}} \rangle = \frac{\mathcal{G}^2 \mathbf{g}^\dagger \mathbf{E}^{-1} \mathbf{E}' \mathbf{E}^{-1} \mathbf{g}}{(\mathbf{g}^\dagger \mathbf{E}^{-1} \mathbf{g})^2}. \tag{7.9}$$

For large matrices, we expect the result to be self-averaging and given by its expectation. Since the noise in $\mathbf{w}_{\mathbf{E}}$ can be assumed to be independent from that in $\mathbf{E}'$, we get for large N [136]:

$$w^*_{\mathbf{E}} \mathbf{E}' w_{\mathbf{E}} \approx w^*_{\mathbf{E}} \mathbf{C} w_{\mathbf{E}} \tag{7.10}$$

and one readily obtains, from the fact that Eq. (7.5) is the minimum possible risk, the following inequality: $\mathcal{R}^2_{\text{true}} \leqslant \mathcal{R}^2_{\text{out}}$. We plot in Fig. 23 an illustration of these inequalities using the so-called efficient frontier where we assumed that $\mathbf{g} = (1, \ldots, 1)^*$. For a given $\mathbf{C}$ (here a shifted GOE around the identity matrix, with $\sigma = 0.2$), we build $\mathbf{w}_{\mathbf{C}}$ and $\mathbf{w}_{\mathbf{E}}$ and compare Eqs. (7.5), (7.7) and (7.9) for $q = 0.5$. We see that using $\mathbf{w}_{\mathbf{E}}$ is clearly overoptimistic and can potentially lead to disastrous results in practice. We refer to [137] for a rigorous study of this problem. We emphasize that this conclusion holds for different risk measures as well [6,7].

7.1.2. 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

$$\mathbf{g} \sim \mathcal{N}_N(0, \mathbf{I}_N), \tag{7.11}$$

but the result holds for any vector $\mathbf{g}$ whose direction is independent of $\mathbf{C}$ or $\mathbf{E}$, such that $\mathbf{g}$ is normalized as $\mathbf{g}^*\mathbf{g} = N$, i.e. each component of $\mathbf{g}$ is of order unity. We emphasize that these assumptions are not necessarily realistic (predictors can be biased along the principal components of $\mathbf{C}$) but allow us to quantify more precisely the relation between the in/true/out of sample risk. The suboptimal returns that follow the use "bad" predictors $\mathbf{g}$ is outside of the scope of this review. Let $\mathbf{M}$ be a positive definite matrix which is independent from the vector $\mathbf{g}$, then we have in the large N limit,

$$\frac{\mathbf{g}^*\mathbf{M}\mathbf{g}}{N} = \frac{1}{N} \operatorname{Tr}[\mathbf{g}\mathbf{g}^*\mathbf{M}] \underset{\text{freeness}}{=} \frac{\mathbf{g}^*\mathbf{g}}{N} \varphi(\mathbf{M}) \tag{7.12}$$

where we recall that φ is the normalized trace operator. Thus, from our assumption (7.11) we easily deduce,

$$\frac{\mathbf{g}^*\mathbf{M}\mathbf{g}}{N} - \varphi(\mathbf{M}) \underset{N\to\infty}{\to} 0. \tag{7.13}$$

Now setting $\mathbf{M} = \{\mathbf{E}^{-1},\ \mathbf{C}^{-1}\}$, we apply Eq. (7.13) to Eqs. (7.7), (7.5) and (7.9) respectively, to find

$$\begin{aligned}
\mathcal{R}^2_{\text{in}} &\to \frac{\mathcal{G}^2}{N\varphi(\mathbf{E}^{-1})},\\
\mathcal{R}^2_{\text{true}} &\to \frac{\mathcal{G}^2}{N\varphi(\mathbf{C}^{-1})},\\
\mathcal{R}^2_{\text{out}} &\to \frac{\mathcal{G}^2\varphi(\mathbf{E}^{-1}\mathbf{C}\mathbf{E}^{-1})}{N\varphi^2(\mathbf{E}^{-1})},
\end{aligned} \tag{7.14}$$

where we recall that φ is the normalized trace operator defined in Eq. (2.61). Let us focus on the first two terms above. For $q < 1$, we have shown above that in the high-dimensional regime one has $\varphi(\mathbf{C}^{-1}) = (1-q)\varphi(\mathbf{E}^{-1})$ – see Eq. (3.24). As a result, we have, for $N \to \infty$

$$\mathcal{R}^2_{\text{in}} = (1-q)\mathcal{R}^2_{\text{true}}. \tag{7.15}$$

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 (7.15).

Next we would like to find the same type of relation for the "out-of-sample" risk. We recall that under the framework of Section 3, we may always rewrite $\mathbf{E} = \mathbf{C}^{1/2}\mathcal{W}\mathbf{C}^{1/2}$ where $\mathcal{W}$ is a white Wishart matrix of parameter q independent from $\mathbf{C}$. Hence, we have for the out-of-sample risk

$$\mathcal{R}^2_{\text{out}} = \frac{\mathcal{G}^2\varphi(\mathbf{C}^{-1}\mathcal{W}^{-2})}{N\varphi^2(\mathbf{E}^{-1})}$$

when $N \to \infty$. Then, the trick is to notice that in the limit of large matrices, $\mathcal{W}$ and $\mathbf{C}$ are *asymptotically free*. This allows us to conclude from the freeness relation (2.64) that

$$\varphi(\mathbf{C}^{-1}\mathcal{W}^{-2}) = \varphi(\mathbf{C}^{-1})\,\varphi(\mathcal{W}^{-2}), \tag{7.16}$$

Hence, using the asymptotic relation (3.24), we find:

$$\mathcal{R}^2_{\text{out}} = \mathcal{G}^2(1-q)^2\frac{\varphi(\mathcal{W}^{-2})}{N\varphi(\mathbf{C}^{-1})}. \tag{7.17}$$

Finally, one can readily compute $\varphi(\mathcal{W}^{-2})$ by performing the large $z \to 0$ expansion of the Stieltjes transform of the Marčenko–Pastur density given Eq. in (3.24) by replacing $\mathbf{C}$ with $\mathbf{I}_N$, that is to say $\varphi(\mathcal{W}^{-2}) = (1-q)^{-3}$ for $q < 1$. We finally get:

$$\mathcal{R}^2_{\text{out}} = \frac{\mathcal{R}^2_{\text{true}}}{1-q}. \tag{7.18}$$

All in all, we obtained the following asymptotic relations:

$$\frac{\mathcal{R}^2_{\text{in}}}{1-q} = \mathcal{R}^2_{\text{true}} = (1-q)\mathcal{R}^2_{\text{out}}, \tag{7.19}$$

which holds for a completely general $\mathbf{C}$. Note that similar results have been obtained in a slightly different context in [136] for $\mathbf{C} = I_N$ and later in [138,139]. Hence, if one invests with the "naive" weights $\mathbf{w}_{\mathbf{E}}$, 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.

7.1.3. Out-of-sample risk minimization

We insisted throughout the last section that the relevant quantity to control in portfolio management is the realized, out-of-sample risk. It is also clear from Eq. (7.19) that using the sample estimate $\mathbf{E}$ is a very bad idea and hence, it is natural to ask: which estimator of $\mathbf{C}$ should one use to minimize the out-of-sample risk? The Markowitz formula (7.4) 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[27]

$$\begin{cases} \min_{\mathbf{w}} \mathbb{E}\left[\dfrac{1}{T_{\text{out}}} \left(\displaystyle\sum_{t'=t+1}^{t+T_{\text{out}}} \langle \mathbf{w}, \mathbf{r}_{t'} \rangle \right)^2 \middle| \mathcal{F}(t) \right], \\ \text{s.t. } \mathbf{w}^* \mathbf{g} \geq \mathcal{G}, \end{cases} \tag{7.20}$$

where $\mathcal{F}(t)$ is all the information available at time t (the investment data), T_{out} is the out-of-sample period, and $\mathbf{r}$ is the vector of returns of the N stocks in our portfolio. Assuming i.i.d 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:

$$\begin{aligned} \mathbb{E}\left[\frac{1}{T_{\text{out}}} \left(\sum_{t'=t+1}^{t+T_{\text{out}}} \langle \mathbf{w}, \mathbf{r}_{t'} \rangle \right)^2 \middle| \mathcal{F}(t) \right] &= \left\langle \mathbf{w}, \frac{1}{T_{\text{out}}} \sum_{t'} \mathbb{E}\left[\mathbf{r}_t \mathbf{r}_t^* \middle| \mathcal{F}(t) \right] \mathbf{w} \right\rangle, \\ &= \left\langle \mathbf{w}, \mathbb{E}\left[\mathbf{C} \middle| \mathcal{F}(t) \right] \mathbf{w} \right\rangle. \end{aligned} \tag{7.21}$$

Recalling the results from Section 5, 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[28] (see Eq. (5.11)). Therefore, using the result Eq. (5.12), 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 (7.9). Let us show this explicitly in the context of rotationally invariant estimators, that we considered in Sections 5 and 6. Let us define our RIE as

$$\Xi = \sum_{i=1}^{N} \xi(\lambda_i) \mathbf{u}_i \mathbf{u}_i^*,$$

where we recall that $[\mathbf{u}_i]_{i \in [\![1,N]\!]}$ 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 (7.13) is still valid, such that the realized risk associated to the portfolio $\mathbf{w}_\Xi$ reads for $N \to \infty$:

$$\mathcal{R}_{\text{out}}^2(\Xi) = \mathcal{G}^2 \frac{\text{Tr}\left(\Xi^{-1} \mathbf{C} \Xi^{-1} \right)}{\left(\text{Tr}\, \Xi^{-1} \right)^2} \tag{7.22}$$

using the spectral decomposition of Ξ, we can rewrite the numerator as

$$\text{Tr}\left(\Xi^{-1} \mathbf{C} \Xi^{-1} \right) = \sum_{i=1}^{N} \frac{\langle \mathbf{u}_i, \mathbf{C}\mathbf{u}_i \rangle}{\xi^2(\lambda_i)}. \tag{7.23}$$

On the other hand, one can rewrite the denominator of Eq. (7.22) as

$$\left(\text{Tr}\, \Xi^{-1} \right)^2 = \left(\sum_{i=1}^{N} \frac{1}{\xi(\lambda_i)} \right)^2. \tag{7.24}$$

Regrouping these last two equations allows us to rewrite Eq. (7.22) as

$$\mathcal{R}_{\text{out}}^2(\Xi) = \mathcal{G}^2 \sum_{i=1}^{N} \frac{\langle \mathbf{u}_i, \mathbf{C}\mathbf{u}_i \rangle}{\xi^2(\lambda_i)} \left(\sum_{i=1}^{N} \frac{1}{\xi(\lambda_i)} \right)^{-2}. \tag{7.25}$$

Our aim is to find the optimal shrinkage function $\xi(\lambda_j)$ associated to the sample eigenvalues $[\lambda_j]_{j \in [\![1,N]\!]}$, such that the out-of-sample risk is minimized. This can be done by solving, for a given j, the following first order condition:

$$\frac{\partial \mathcal{R}_{\text{out}}^2(\Xi)}{\partial \xi(\lambda_j)} = 0. \tag{7.26}$$

[27] 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.

[28] We expect this result to hold also for the multivariate Student, see Section 3.1.3.

By performing the derivative with respect to $\xi(\lambda_j)$ in (7.25), one obtains

$$-2\frac{\langle \mathbf{u}_j, \mathbf{C}\mathbf{u}_j\rangle \xi'(\lambda_j)}{\xi^3(\lambda_j)}\left(\sum_{i=1}^N \frac{1}{\xi(\lambda_i)}\right)^{-2} + 2\frac{\xi'(\lambda_j)}{\xi^2(\lambda_j)}\left(\sum_{i=1}^N \frac{\langle \mathbf{u}_i, \mathbf{C}\mathbf{u}_i\rangle}{\xi^2(\lambda_i)}\right)\left(\sum_{i=1}^N \frac{1}{\xi(\lambda_i)}\right)^{-3} = 0, \tag{7.27}$$

and one can check that the solution is precisely given by

$$\xi(\lambda_j) = \langle \mathbf{u}_j, \mathbf{C}\mathbf{u}_j\rangle := \xi_j^{\text{ora.}}, \tag{7.28}$$

which is the oracle estimator that we have studied in Sections 5 and 6. Note that this result has been obtained in [140] where the authors also showed that this estimator maximizes the Sharpe ratio, i.e., the expected return of the strategy divided by its volatility.

As a conclusion, the optimal RIE (6.5) 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

$$\mathcal{R}^2_{\text{out}}(\Xi^{\text{ora.}}) = \frac{\mathcal{G}^2}{\text{Tr}\left[(\Xi^{\text{ora.}})^{-1}\right]}, \tag{7.29}$$

where we used the notable property that for any $n \in \mathbb{Z}$:

$$\text{Tr}[(\Xi^{\text{ora.}})^n \mathbf{C}] = \text{Tr}[(\Xi^{\text{ora.}})^{n+1}], \tag{7.30}$$

which directly follows from the general formula (6.2).

7.1.4. *Optimal in and out-of-sample risk for an inverse Wishart prior*

In this section, we specialize the result (7.29) 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. (2.55) by $z \to 0$ that

$$\varphi(\mathbf{C}^{-1}) = -\mathfrak{g}_{\mathbf{C}}(0) = 1 + \frac{1}{2\kappa}, \tag{7.31}$$

so that we get from Eq. (7.14) that in the large N limit:

$$\mathcal{R}^2_{\text{true}} = \frac{\mathcal{G}^2}{N}\frac{2\kappa}{1+2\kappa}. \tag{7.32}$$

Next, we see from Eq. (7.29) 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:

$$\varphi((\Xi^{\text{ora.}})^{-1}) = -(1+2q\kappa)\mathfrak{g}_{\mathbf{E}}(-2q\kappa) = 1 + \frac{1}{2\kappa(1+q(1+2\kappa))}, \tag{7.33}$$

and therefore we have from Eq. (7.29)

$$\mathcal{R}^2_{\text{out}}(\Xi^{\text{ora.}}) = \frac{\mathcal{G}^2}{N}\frac{2\kappa(1+q(1+2\kappa))}{1+2\kappa(1+q(1+2\kappa))}, \tag{7.34}$$

from which it is clear from Eqs. (7.34) and (7.32) that for any $\kappa > 0$:

$$\frac{\mathcal{R}^2_{\text{out}}(\Xi^{\text{ora.}})}{\mathcal{R}^2_{\text{true}}} = 1 + q\frac{2\kappa}{1+2\kappa(1+q(1+2\kappa))} \geqslant 1, \tag{7.35}$$

where the last inequality becomes an equality 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

$$\mathcal{R}^2_{\text{in}}(\Xi^{\text{ora.}}) = \mathcal{G}^2\frac{\text{Tr}\left[(\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}\right]}{N\varphi^2((\Xi^{\text{ora.}})^{-1})}, \tag{7.36}$$

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 (6.24), we can once again find a closed formula. As above, we relegate the derivation at the end of this section and the result reads:

$$\varphi\left((\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}\right) = -(1-z)^2\left[\mathfrak{g}_{\mathbf{E}}(z) + z\mathfrak{g}'_{\mathbf{E}}(z)\right]\Big|_{z=-2q\kappa} = \frac{(1+2\kappa)(1+2q\kappa)^3}{2\kappa(1+q(1+2\kappa))^3}. \tag{7.37}$$

Hence by plugging Eqs. (7.37) and (7.33) into Eq. (7.36), we obtain

$$\mathcal{R}^2_{\text{in}}(\Xi^{\text{ora.}}) = \frac{\mathcal{G}^2}{N}\frac{2\kappa(1+2q\kappa)}{(1+2\kappa)(1+q(1+2\kappa))}, \tag{7.38}$$

and we therefore deduce with Eq. (7.32) that for any $\kappa > 0$:

$$\frac{\mathcal{R}^2_{\text{in}}(\Xi^{\text{ora.}})}{\mathcal{R}^2_{\text{true}}} = 1 - \frac{q}{1+q(1+2\kappa)} \leqslant 1, \tag{7.39}$$

where the inequality becomes an equality for $q = 0$ as above.

Finally, one may easily check from Eqs. (7.19), (7.35) and (7.39), that

$$\mathcal{R}^2_{\text{in}}(\Xi^{\text{ora.}}) - \mathcal{R}^2_{\text{in}}(\mathbf{E}) \geqslant 0, \qquad \mathcal{R}^2_{\text{out}}(\Xi^{\text{ora.}}) - \mathcal{R}^2_{\text{out}}(\mathbf{E}) \leqslant 0, \tag{7.40}$$

showing explicitly 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 (7.33) and (7.37). We begin with Eq. (7.33) and we use that the eigenvalues of the oracle estimator converge to Eq. (6.24) when $N \to \infty$. $\mathbf{C}$ is assumed to be an inverse Wishart of parameter $\kappa > 0$. Hence, one has

$$\varphi((\Xi^{\text{ora.}})^{-1}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\alpha_s(\lambda_i-1)} = \frac{1}{\alpha_s}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\frac{1-\alpha_s}{\alpha_s}+\lambda_i}, \tag{7.41}$$

and using Eq. (5.19), 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

$$\varphi((\Xi^{\text{ora.}})^{-1}) \sim (1+2q\kappa)\mathfrak{g}_{\mathbf{E}}(-2q\kappa), \tag{7.42}$$

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. (7.33) 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 (3.41). Nonetheless, it seems that Eq. (3.41) 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. (3.41) as a power of ε to find:

$$\mathfrak{g}_{\mathbf{E}}(z) = \frac{q-z}{z(1+q-z)} + \mathcal{O}(\varepsilon),$$

meaning that for $\varepsilon = 0$, we obtain

$$\mathfrak{g}_{\mathbf{E}}(-2q\kappa) = -\frac{1+2\kappa}{2\kappa(1+q(1+2\kappa))}. \tag{7.43}$$

It is then easy to deduce Eq. (7.33) from this last equation and Eq. (7.42).

The computation of Eq. (7.37) is a bit more tedious but very similar to the derivation of the previous section. Indeed, using that $(\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}$ share the same eigenbasis, we have thanks to Eq. (6.24):

$$\varphi((\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i}{(1+\alpha_s(\lambda_i-1))^2}, \tag{7.44}$$

which gives after some simple manipulations:

$$\varphi((\Xi^{\text{ora.}})^{-1}\mathbf{E}(\Xi^{\text{ora.}})^{-1}) = \frac{1}{\alpha_s}\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{1+\alpha_s(\lambda_i-1)} - \frac{1-\alpha_s}{(1+\alpha_s(\lambda_i-1))^2}\right]. \tag{7.45}$$

Defining $z = -2q\kappa < 0$, one can deduce the first equality of Eq. (7.37) using the same identification with the Stieltjes transform (and its derivative with respect to z) as above. The derivative of Eq. (3.41) reads:

$$\mathfrak{g}'_{\mathbf{E}}(z) = \frac{1}{z^2(z+2q\kappa)^2}\left[z(2\kappa q+z)\left(1+\kappa - \frac{\kappa(\kappa(q-z+1)+1)}{\sqrt{\kappa^2(z+q-1)^2-2\kappa z(1+2\kappa)}}\right) - 2(q\kappa+z)\beta(z)\right], \tag{7.46}$$

where $\beta(z)$ is defined by

$$\beta(z) := z(1+\kappa) - \kappa(1-q) + \sqrt{\kappa^2(z+q-1)^2 - 2\kappa z(1+2\kappa)}, \tag{7.47}$$

which is the denominator of Eq. (3.41). We omit further details as the proof of the second equality of Eq. (7.37) relies on a Taylor expansion around $-2q\kappa$ in the same spirit than in the previous section. This regularizes the Stieltjes transform and its derivative and one eventually obtains:

$$-2q\kappa\, \mathfrak{g}'_{\mathbf{E}}(-2q\kappa) = \frac{q(1+2\kappa)\big[q+2(1+\kappa+2q\kappa(1+\kappa))\big]}{2\kappa(1+q(1+2\kappa))^3} \tag{7.48}$$

and we find the desired result by plugging this last equation into Eq. (7.37).

7.2. 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 e.g. 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 [18]. 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. [29,141,93] for complementary sources of information.

We shall present four different classes of estimators. The first one is the **linear shrinkage** method. This estimator has been studied in details in Sections 5 and 6 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 [28,24] 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 3 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 3). 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 [33].

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 section.

7.2.1. Linear shrinkage

We recall that the linear shrinkage is given by

$$\Xi^{\text{lin}} = \alpha_s \mathbf{E} + (1-\alpha_s)\mathbf{I}_N, \qquad \alpha \in [0,1]. \tag{7.49}$$

As discussed in Section 5, this estimator has a long history in high-dimensional statistics [16,17] 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 [17] or in [133] in a more RMT oriented standpoint. It is easy to see that Ξ^{lin} shares the same eigenbasis as the sample estimator $\mathbf{E}$, and is thus a rotationally invariant estimator with

$$\Xi^{\text{lin}} = \sum_{i=1}^{N} \xi^{\text{lin}} \mathbf{u}_i \mathbf{u}_i^*, \qquad \xi^{\text{lin}} = 1 + \alpha_s(\lambda_i - 1). \tag{7.50}$$

We already emphasized that this estimator exhibits all the expected features: the small eigenvalues are shifted upwards (compared to the sample eigenvalues) while the top eigenvalues are pulled downwards (see Fig. 24). As alluded to above, this estimator has been fully investigated in [17]. Most notably, the authors were able to determine an asymptotic optimal formula to estimate α_s directly from the data. Keeping the notations of Section 3, our dataset is $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T) \in \mathbb{R}^{N\times T}$ and we assume that $\mathbb{E}[Y_{it}] = 0$ and $\mathbb{E}[Y_{it}^2] = T^{-1}$ for all $i \in [\![1, N]\!]$. Defining:

$$\begin{aligned} \beta &:= \frac{1}{N}\operatorname{Tr}\left[(\mathbf{E}-\mathbf{I}_N)(\mathbf{E}-\mathbf{I}_N)^*\right] \\ \gamma &:= \max\left(\beta, \frac{1}{T^2}\sum_{k=1}^{T}\frac{1}{N}\operatorname{Tr}\left[(\mathbf{y}_k\mathbf{y}_k^* - \mathbf{E})(\mathbf{y}_k\mathbf{y}_k^* - \mathbf{E})^*\right]\right), \end{aligned} \tag{7.51}$$

then

$$\widehat{\alpha}_s = 1 - \frac{\beta}{\gamma}, \tag{7.52}$$

is a consistent estimator of α_s in the high-dimensional regime [17].

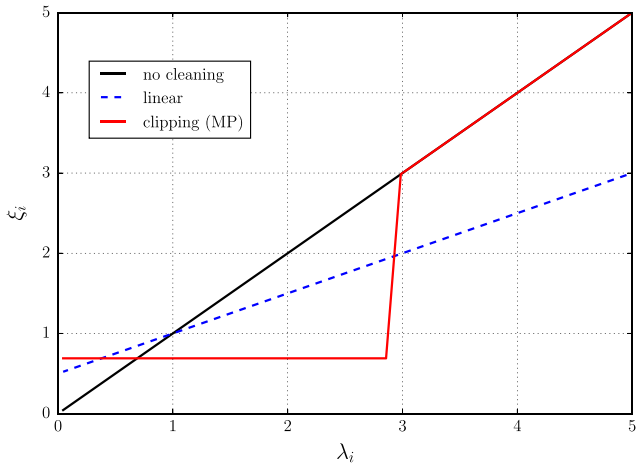


Fig. 24. Impact on sample eigenvalues of the eigenvalues clipping (7.55) (red plain line) with a threshold given by $(1+\sqrt{q})^2$ with $q = 0.5$ and the linear shrinkage (7.50) (blue dashed line) with intensity $\alpha_s = 0.5$. We see that the lowest eigenvalues are shifted upward.

Using tools from RMT, and more precisely the result of Sections 3 and 4, we can find another consistent estimators of α_s which uses the fact that linear shrinkage implicitly assumes the underlying correlation matrix to be an Inverse-Wishart matrix with parameter κ, from which α_s is deduced as $\alpha_s = (1+2q\kappa)^{-1}$. The value of κ can be extracted from the data using the relation (valid for $q < 1$):

$$\mathfrak{g}_{\mathbf{C}}(0) = (1-q)\mathfrak{g}_{\mathbf{E}}(0) = 1 + 2\kappa \tag{7.53}$$

where the last equality can be deduced from (2.55) and (3.24). Therefore, we obtain a simple estimate for κ from the trace of $\mathbf{E}^{-1}$ as:

$$\kappa = \frac{1}{2}\left((1-q)\frac{\operatorname{Tr}\mathbf{E}^{-1}}{N} - 1\right). \tag{7.54}$$

However, this estimate is only reliable when κ 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 κ). A more robust alternative to estimate κ is the "two-sample" test introduced in Section 4.2, see Eqs (4.40) and [127].

7.2.2. Eigenvalues clipping

This method is perhaps the first RMT-based estimator for large covariance matrices. It has been investigated in several papers [23,28,24] 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 Fig. 10). An alternative interpretation would be that outliers are true factors while the others are meaningless.

In a recent paper [99], 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 (3.56), 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 threshold between noise and signal.

Endowed with a simple rule to isolate the signal eigenvalues, how should one clean the noisy ones? Laloux et al. [28] 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:

$$\Xi^{\text{clip.}} := \sum_{i=1}^{N} \xi_i^{\text{c}} \mathbf{u}_i \mathbf{u}_i^*, \qquad \xi_i^{\text{clip.}} = \begin{cases} \lambda_i & \text{if } \lambda_i \geqslant (1+\sqrt{q})^2 \\ \bar{\lambda} & \text{otherwise,} \end{cases} \tag{7.55}$$

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 Fig. 24 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_{eff} that allows one to fit best the empirical spectral density [28]. This effect can be induced either by small temporal autocorrelation in the time series [86,142,143] 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 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 α_c plays the role of the upper bound λ_+ 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 α_c is not based on any theoretical rule.

Another concern about this method is that we know from Section 6.3 that the optimal estimator of the large outliers is *not* their bare empirical value λ_i. Rather, one should shift them downwards even when far from the bulk, 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. (7.55) (see [144] for a related discussion).

7.2.3. 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 (3.9). More formally, we seek the set of true eigenvalues $[\mu_j]_{j\in[\![1,N]\!]}$ that solve Eq. (3.9) 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 3.2.1, 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 Section 5). More specifically, we assume that $\mathbf{C}$ belongs to a rotationally invariant ensemble – so that 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

$$\lambda_i \to \widehat{\mu}_i \quad \text{such that} \quad \frac{i}{N} = \int_{\widehat{\mu}_i}^{\infty} \widehat{\rho}_{\mathbf{C}}(x)\mathrm{d}x. \tag{7.56}$$

Note that under the transformation (7.56), 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 (3.49) 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 [29,145]. In that case, the parametric substitution turns out to be explicit in the limit of large dimension. Moreover, the estimation of the unique parameter λ_0 in this model can be done using e.g. maximum likelihood, as we can compute exactly $\rho_{\mathbf{E}}$ on $\mathbb{R}^+$ using (3.50) and (3.35). This then yields a parameter $\widehat{\lambda}_0$ and hence $\widehat{\rho}_{\mathbf{C}}$ as well. As a result, the substitution procedure (7.56) becomes for $N \to \infty$ [29]:

$$\mu_i = -\widehat{\lambda}_0 + \frac{(1+\widehat{\lambda}_0)}{2}\sqrt{\frac{N}{i}} \qquad i \in [\![1,N]\!]. \tag{7.57}$$

We present such a procedure in Fig. 25 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 [33] 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:

$$\left\{ z_j = \frac{1}{\mathfrak{g}_{\mathbf{S}}(z_j)} \left[1 - q + q \int \frac{\rho_{\mathbf{C}}(\mu)d\mu}{1-\mu\,\mathfrak{g}_{\mathbf{S}}(z_j)} \right], \quad \text{with} \quad z_j = \lambda_j - i\eta \right\}_{j=1}^{N}$$

that follows from Eq. (3.35) and where we recall that $\mathbf{S}$ is the $T \times T$ dual matrix of $\mathbf{E}$ defined in (3.32). The main assumption of this method is to decompose the density of states $\rho_{\mathbf{C}}$ as a weighted sum of Dirac masses:

$$\rho_{\mathbf{C}}(\mu) = \sum_{k=1}^{N} \widehat{w}_k \delta(\mu - \mu_k), \quad \text{such that} \sum_{k=1}^{N} \widehat{w}_k = 1 \quad \text{and } \widehat{w}_k \geq 0, \ \forall\, k \in [\![1,N]\!]. \tag{7.58}$$

Note that this decomposition simply uses 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 μ_j and their corresponding weights $\widehat{w}_j$ so that the parametrization looks inextricably complex. In [33], the author suggested to fix the positions $[\mu_j]_{j\in[\![1,N]\!]}$ *a priori* such that we are left with the weights $[\widehat{w}_j]_{j\in[\![1,N]\!]}$

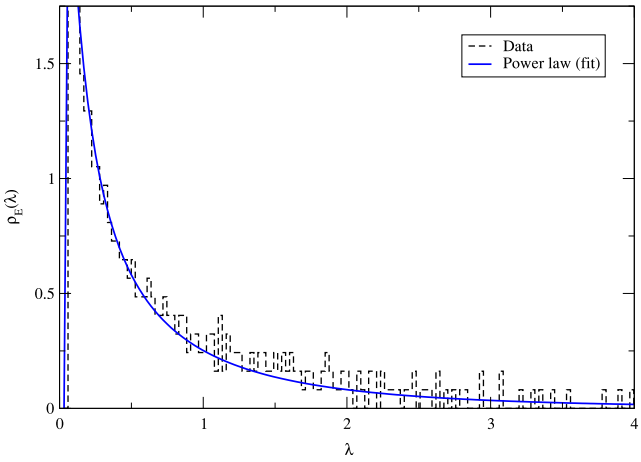


Fig. 25. Fit of the power law distribution (3.49) on the sample eigenvalues of the 450 most liquid assets of the S&P index from 2006 to 2010 using the Marčenko–Pastur equation (3.9). The fit has been performed using a maximum likelihood procedure and yields $\alpha \approx 0.3$. The black dashed histogram represents the empirical spectral density.

as the only unknown variables in the problem. Within this framework, the author then proposed to obtain the optimal weights through the following optimization program:

$$[\widehat{w}_j]_{j\in[\![1,N]\!]} = \begin{cases} \underset{\{w_i\}_{i=1}^N}{\text{argmin}}\ \mathcal{L}\left(\left\{\frac{1}{\mathfrak{g}_{\mathbf{S}}(z_j)}\left[1-q+q\sum_{k=1}^{N}\frac{w_k}{1-\mu_k\,\mathfrak{g}_{\mathbf{S}}(z_j)}\right]-z_j\right\}_{j=1}^{N}\right) \\ \text{subject to } \sum_{k=1}^{N} w_k = 1, \quad \text{and } w_k \geq 0 \quad \forall\, k \in [\![1,N]\!], \end{cases} \tag{7.59}$$

where $\mathcal{L}$ is a certain loss function and $z_j = \lambda_j - i\eta$. 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 $[\mu_j]_{j\in[\![1,N]\!]}$ that have to be chosen.

In the large N limit, the first approximation is fairly accurate (see Section 7). 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 (7.59) for each λ_j, then the consistency of the algorithm has been showed in [33] 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 $[\widehat{w}_j]_{j\in[\![1,N]\!]}$, the cleaning procedure is immediate

$$\lambda_i \to \widehat{\mu}_i \quad \text{where} \quad \widehat{\mu}_i = \min\left\{x \in \mathbb{R}^+ : \sum_{k=1}^{N}\widehat{w}_k\Theta(\mu_k - x) \geq \frac{i}{N}\right\} \tag{7.60}$$

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 $[\mu_j]_{j\in[\![1,N]\!]}$ has therefore been proposed in [111]. This alternative method, called QuEST, turns out to be much more robust numerically (see [146] and Section 8 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 find an estimator of the true eigenvalues $\widehat{\mu}_i$ for all $i = 1, \dots, N$. As a result, the eigenvalue substitution estimator is then obtained as

$$\Xi^{\text{sub}} = \sum_{k=1}^{N}\widehat{\mu}_k\mathbf{u}_k\mathbf{u}_k^*. \tag{7.61}$$

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 Section 4. Therefore,

for covariance matrices estimation, it is not advised to use the substitution (7.61) since this is not the optimal solution. However, it can be used to compute the optimal RIE (6.5) and we refer to Section 8.1.3 for more details.

7.3. Factor models

The main idea behind linear factor models is quite simple: the (normalized) data Y_{it} is represented as a linear combination of M common factors F

$$Y_{it} = \sum_{k=1}^{M} \beta_{ik} f_{kt} + \varepsilon_{it} \tag{7.62}$$

where the β_{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 ε_{it} is the idiosyncratic part of Y_{it} (or the residual in Statistics), assumed to be of zero mean. The model (7.62) in matrix form reads

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{F} + \boldsymbol{\varepsilon}, \tag{7.63}$$

which is known as *Generalized Linear Model* [147]. It is often assumed that the residuals are i.i.d. across i with t fixed (see e.g. [148] for an application in Finance). It is not hard to see that the covariance matrix under the model (7.62), is given by

$$\mathbf{C} = \beta \Sigma_F \beta^* + \Sigma_\varepsilon \tag{7.64}$$

where Σ_F is the covariance matrix of size $M \times M$ of the factors F – which can be chosen, without loss of generality, to be proportional to the identity matrix – and Σ_ε is the $N \times N$ covariance matrix of the residuals ε, which is simply the identity in the simplest framework. Within the linear decomposition (7.62), 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

$$\mathbf{E} = \frac{1}{T}(\boldsymbol{\beta}\mathbf{F} + \boldsymbol{\varepsilon})(\boldsymbol{\beta}\mathbf{F} + \boldsymbol{\varepsilon})^*, \tag{7.65}$$

becomes less 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 $\boldsymbol{\varepsilon}$. 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 [149], the author assumes that the empirical estimator of Σ_ε 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 (7.65) cannot exceed

$$\lambda_+^{\text{eff}}(q) := (1 + \sqrt{q})^2 + \delta(q, N) \tag{7.66}$$

where the last term δ is a suitably defined constant as to reflect the width of the Tracy–Widom tail, i.e. $\delta(q, N) \sim N^{-2/3}$ [149]. If however one observes that the largest sample eigenvalue λ_1 exceeds λ_+^{eff}, then a true factor probably exists. In that case, the procedure suggested in [149] is to extract the corresponding largest component from the data:

$$Y_{it}^{(1)} = Y_{it} - \beta_{1t} f_{1t},$$

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 [150] to the case where the empirical estimator of the Σ_ε is an *anisotropic* Wishart matrix for which one has several results concerning the spectrum (see Section 3). The procedure is similar to the one above: the author proposed an algorithm to detect outliers for this anisotropic Wishart matrix using the results of Ref. [151]. We refer to [150] for more details. We can therefore see that RMT allows one to derive some rigorously 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) [152] assumes a unique factor corresponds to the market portfolio, or its extension to three factors model by Fama–French [153] (see [154] for further more recent extensions). In that case, one can simplify the problem to the estimation of the β by assuming that the factors f_k and the residuals ε_i are linearly uncorrelated:

$$\langle f_k f_l \rangle = \delta_{kl}, \quad \langle \varepsilon_i \varepsilon_j \rangle = \delta_{ij} \left(1 - \sum_l \beta_{li}^2\right) \quad \text{and} \quad \langle f_k \varepsilon_l \rangle = 0, \tag{7.67}$$

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

$$C_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ \left(\boldsymbol{\beta}\boldsymbol{\beta}^*\right)_{ij} & \text{otherwise.} \end{cases} \tag{7.68}$$

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 β using the sample data, which is due to the recent paper [155]. Note that the eigenvalue clipping (7.68) can recovered by setting $\beta \equiv \beta_{PCA}$ where

$$\beta_{PCA} := \mathbf{U}_{|M} \boldsymbol{\Lambda}_{|M}^{1/2}, \tag{7.69}$$

with $\mathbf{U}$ the sample eigenvectors, Λ 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 [155] suggests finding the $\boldsymbol{\beta}$s such that:

$$\widehat{\boldsymbol{\beta}} := \underset{\beta}{\text{argmin}}\ \mathcal{L}\left(\left\| \frac{1}{T}\mathbf{Y}\mathbf{Y}^* - \boldsymbol{\beta}\boldsymbol{\beta}^* \right\|_{\text{off-diag}}\right), \tag{7.70}$$

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 (7.69) and with a quadratic norm $\mathcal{L}$. We refer the reader to [155] for more details on the procedure and its implementation, as well as an extension of the model to non-linear (volatility) dependencies.

8. Numerical implementation and empirical results

This Section aims at putting all the above ideas into practice in a financial context, the final goal being 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 6.5 that the discrete version (6.26) of the optimal RIE (6.5) deviates systematically from its limiting value for small eigenvalues. But as we discussed in Section 7, the estimation of these small eigenvalues is particularly important since Markowitz optimal portfolios tend to overweight them 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 [146] and [156]) that attempt to correct this systematic underestimation of the small eigenvalues. We will then turn to numerical experiments on synthetic and real financial data and test the quality of the regularized RIE for real world applications.

8.1. Finite N regularization of the optimal RIE (6.26)

8.1.1. 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 (6.5) should yield $\widehat{\xi}(\lambda_i) = 1$ exactly as $N \to \infty$ (see Eq. (6.21)). We therefore compare the observable shrinkage function $\widehat{\xi}^N$ (6.26) for finite N with its limiting value $\widehat{\xi} = 1$. The results are reported in Fig. 26 where the observable estimator Eq. (6.26) appears as green points while 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 instead of remaining close to unity. This highlights, as stated in [39] or [113], that the behavior for small eigenvalues is more difficult to handle compared to the rest of the spectrum.

This underestimation can be investigated analytically. With $z = \lambda - \mathrm{i}\eta$, we actually see from Fig. 26 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 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 $\mathbb{R}^+$. Let us illustrate this: under the one-cut assumption, we can always decompose the Stieltjes transform as (see Eq. (2.31))

$$\mathfrak{g}_{\mathbf{E}}(z) = \mathfrak{h}(z) + Q(z)\sqrt{d_+(z)}\sqrt{d_-(z)}, \qquad d_\pm(z) := z - \lambda_\pm \tag{8.1}$$

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 $\mathbb{C}^+$. We place ourselves in the situation where $d_-(\lambda) = \varepsilon \ll \eta$, i.e. the eigenvalue λ is very close to zero. Then, we have

$$\begin{aligned} \mathfrak{g}_{\mathbf{E}}(z) &= \mathfrak{h}(z) + Q(z)\sqrt{-\mathrm{i}\eta}\sqrt{d_+(\lambda) - \mathrm{i}\eta} + \mathcal{O}(\varepsilon) \\ &= \mathfrak{h}(z) - (1+\mathrm{i})Q(z)\sqrt{\frac{\eta|d_+(\lambda)|}{2}} + \mathcal{O}(\varepsilon). \end{aligned} \tag{8.2}$$

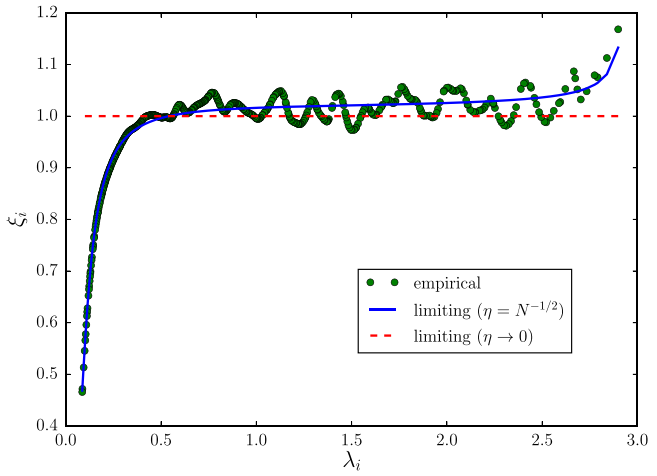


Fig. 26. Evaluation of the empirical RIE (6.26) (green points) for $\mathbf{C} = \mathbf{I}_N$ with $N = 500$. The matrix $\mathbf{E}$ is generated using Wishart matrix with parameter $q = 0.5$. We compare the result with its limiting value for $\eta = N^{-1/2}$ (blue line) and $\eta \to 0^+$ (red dotted line).

Specializing this last equation to the null hypothesis $\mathbf{C} = \mathbf{I}_N$, one infers from Eq. (2.41) that $1/Q(z) = 2qz$ and $\mathfrak{h}(z) = Q(z)(z + q - 1)$. Then plugging (8.2) into (6.5) yields, at the left edge:

$$\widehat{\xi}(\lambda_- - \mathrm{i}\eta) = 1 - \sqrt{\frac{2\eta\sqrt{q}}{(1-\sqrt{q})^2}} + \mathcal{O}(\eta), \tag{8.3}$$

that is to say, there is 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 η 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 η cannot be too small either [113]. We conclude that the underestimation effect we observe in Figs. 26 and 21 is purely due to a finite size effect and would furthermore occur for any model of $\rho_{\mathbf{C}}$ (see Fig. 21). We emphasize that this effect is different from the phase transition affecting left outliers, as displayed in Fig. 20.

8.1.2. *Regularizing the empirical RIE (6.26)*

There are two ways to address this problem. The first one is to use a simple ad-hoc de-noising 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 Fig. 26), 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. (2.53), for which we know that the associated optimal RIE is the linear shrinkage (6.24).[29] Within this specification, the parameter κ allows one to interpolate $\rho_{\mathbf{C}}$ between the infinitely wide measure on $\mathbb{R}^+$ ($\kappa \to 0^+$) and the null hypothesis ($\kappa \to \infty$).

Our procedure, *for the only purpose of regularization*, is to calibrate κ such that the lower edge λ_-^{iw} of the corresponding empirical spectrum (and given in Eq. (3.41)), coincides with the observed smallest eigenvalue λ_N. We then rescale the smallest eigenvalue using the exact factor that would be needed if $\mathbf{C}$ was indeed an Inverse-Wishart matrix, i.e.:

$$\widehat{\xi}_i^{\mathrm{reg}} = \widehat{\xi}_i^{N} \times \max(1, \Gamma_i^{\mathrm{iw}}), \qquad \Gamma_i^{\mathrm{iw}} = \frac{|1 - q + qz_i \mathfrak{g}_{\mathbf{E}}^{\mathrm{iw}}(z_i)|^2}{\lambda_i/(1 + \alpha_s(\lambda_i - 1))}, \quad z_i = \lambda_i - \mathrm{i}N^{-1/2}, \tag{8.4}$$

where $\alpha_s = 1/(1 + 2q\kappa)$ and $\mathfrak{g}_{\mathbf{E}}^{\mathrm{iw}}$ is given in Eq. (3.41). We give a more precise implementation of this "IW-regularization" in the Algorithm 1, and a numerical illustration for an Inverse Wishart matrix (2.58) with parameter $\kappa = 10$ and $q = 0.5$, for which $\alpha_s \approx 0.09$. The results are plotted in Fig. 27 where the empirical points come from a single simulation with $N = 500$.

We now reconsider the numerical examples given in Section 6.5, for which we apply the IW-regularization Algorithm 1. The results are plotted in Fig. 28 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 one observes a significant discrepancy for the empirical, bare estimator (green dots). Hence, the IW-regularization (Algorithm 1) provides a very simple way to correct

[29] A yet simpler solution, proposed in [156] is to consider a rescaled Marčenko–Pastur's spectrum in such a way to fit the smallest eigenvalue λ_N. This is indistinguishable from the IW procedure when κ is large enough, and provides very accurate predictions for US stocks return [156]. Nevertheless, in the presence of very small "true" eigenvalues, corresponding to e.g. very strongly pairs of correlated financial contracts, this simple recipe fails.

Algorithm 1 IW-regularization of the empirical RIE (6.26)

function G_IW(z, q, κ):
  $\lambda_\pm \leftarrow \left[(1+q)\kappa + 1 \pm \sqrt{(2\kappa+1)(2q\kappa+1)}\,\right]/\kappa$;
  return $\left[z(1+\kappa) - \kappa(1-q) - \sqrt{z-\lambda_+}\sqrt{z-\lambda_-}\right]/(z(z+2q\kappa))$;
end function

function RIE(z, q, g):
  return $\mathrm{Re}[z]/|1-q+qzg|^2$;
end function

function DENOISING_RIE($N, q, \{\lambda_i\}_{i=1}^N$): //$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$
  $\kappa \leftarrow 2\lambda_N/\left((1-q-\lambda_N)^2 - 4q\lambda_N\right)$;
  $\alpha \leftarrow 1/(1+2q\kappa)$;
  for $i = 1$ to N **do**
    $z \leftarrow \lambda_i - \mathrm{i}N^{-1/2}$;
    $g \leftarrow \left(\sum_{j\neq i}^N 1/(z-\lambda_j)\right)/(N-1)$;
    $\hat{\xi}_i \leftarrow$ RIE(z, q, g);
    $g \leftarrow$ G_IW(z, q, κ)
    $\Gamma_i \leftarrow (1+\alpha(\lambda_i - 1))/RIE(z, q, g)$;
    if $\Gamma_i > 1$ and $\lambda_i < 1$ **then**
      $\hat{\xi}_i \leftarrow \Gamma_i \hat{\xi}_i$;
    end if
  end for
  $s \leftarrow \sum_i \lambda_i / \sum_i \hat{\xi}_i$; //preserving the trace
  return $\{s \times \hat{\xi}_i\}_{i=1}^N$
end function

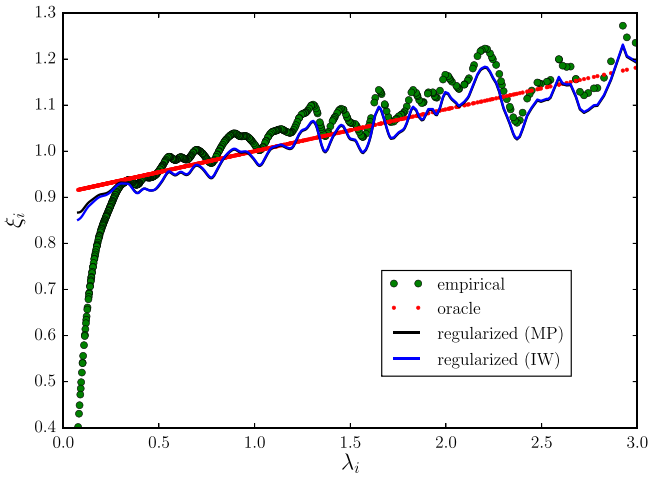
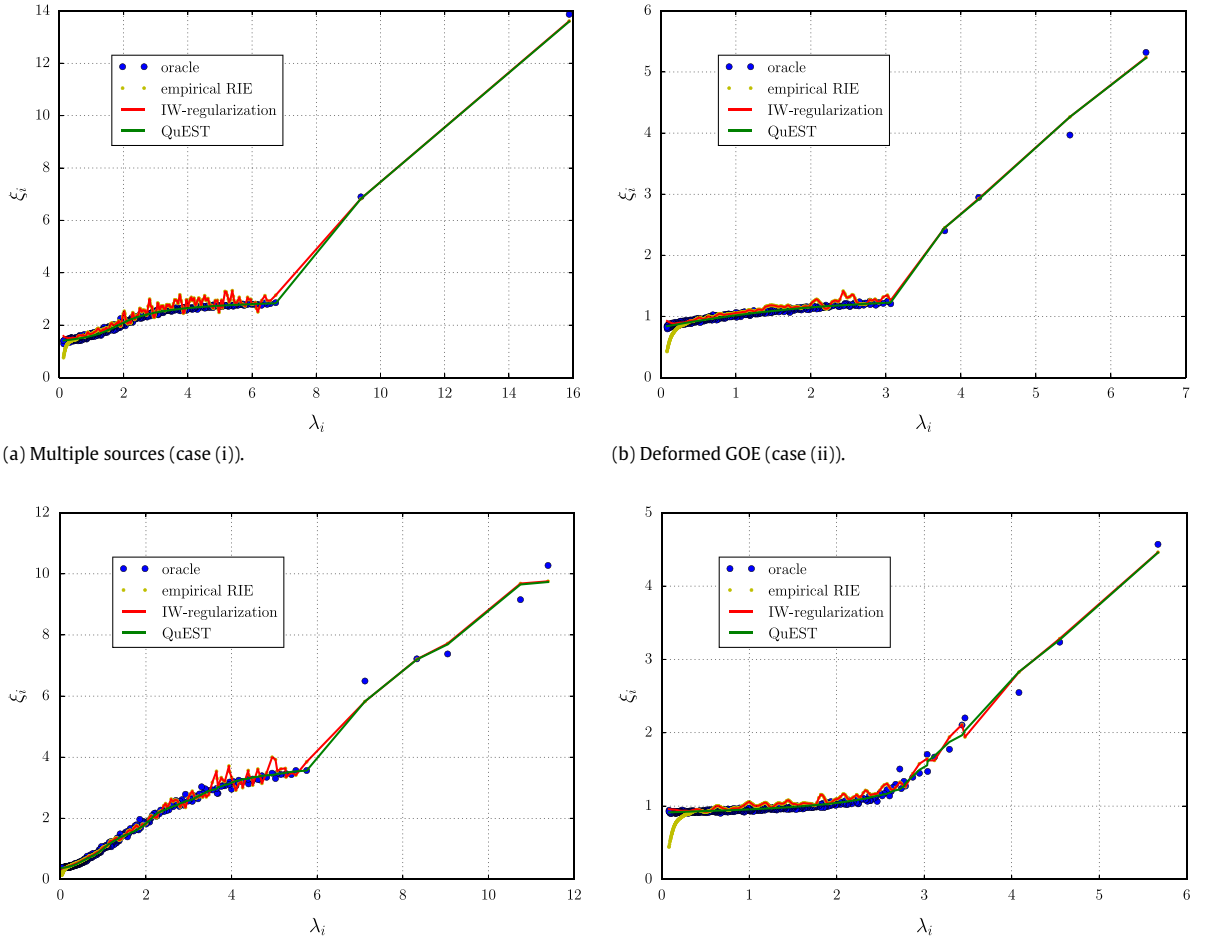


Fig. 27. We apply the IW-regularization $\widehat{\xi}_i^{\mathrm{reg}}$ with $z = \lambda - \mathrm{i}N^{-1/2}$ in the case where $\mathbf{C}$ is an Inverse-Wishart matrix with $\kappa = 10$ and $q = 0.5$. The finite size effect of the empirical RIE (6.26) (green points) is efficiently corrected. The 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 [156]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this systematic downside bias which is of crucial importance whenever 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 1).

8.1.3. Quantized Eigenvalues Sampling Transform (QuEST)

An alternative method, recently proposed by Ledoit and Wolf [146] to approximate numerically the optimal RIE (6.5), is to work with the Marčenko–Pastur equation (3.9). It is somewhat similar to the numerical scheme proposed by N. El Karoui (see Section 7.2.3) to solve the indirect problem of the Marčenko–Pastur equation.



(a) Multiple sources (case (i)).

(b) Deformed GOE (case (ii)).

(c) Toeplitz (case (iii)).

(d) Power law (case (iv)).

Fig. 28. Comparison of the IW-regularization (6.26) (red line) with the empirical RIE (6.26) (yellow dots) and the Oracle estimator (6.2) (blue points) for the four cases presented at the beginning of Section 6.5 with $N = 500$ and $T = 1000$. We also plot the estimation we get using QuEST estimator (8.10) (green line). The results generated with a single realization of $\mathbf{E}$ using a multivariate Gaussian measurement process, and the four specifications of Section 6.5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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.

$$\frac{i}{N} = \int_{-\infty}^{\lambda_i} \rho_{\mathbf{E}}(x)\mathrm{d}x, \tag{8.5}$$

and the aim is to find the quantile, as a function of the population eigenvalues $[\mu_i]_{i\in[\![1,N]\!]}$, such that (8.5) holds. Note that the representation (8.5) is the definition of the classical location of the *bulk* eigenvalues, encountered in Eq. (3.40). 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 these outliers. Nevertheless, for constructing 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 6.2.2). That being said, the "quantized" eigenvalues, expected to be close to the empirical eigenvalues, are defined as

$$\tilde{\gamma}_i(\boldsymbol{\mu}) := N \int_{(i-1)/N}^{i/N} F_{\mathbf{E}}^{-1}(p)\mathrm{d}p, \qquad i \in [\![1, N]\!],\ p \in [0, 1], \tag{8.6}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$, and

$$F_{\mathbf{E}}^{-1}(p) := \sup\left\{x \in \mathbb{R} \,:\, F_{\mathbf{E}}(x) \leqslant p\right\},$$

$$F_{\mathbf{E}}(x) := \begin{cases} \max\Big(1 - 1/q, N^{-1}\sum_{i=1}^{N}\delta_0(\mu_i)\Big) & \text{if } x = 0, \\ \int_0^x \rho_{\mathbf{E}}(u)\mathrm{d}u, & \text{otherwise}, \end{cases} \tag{8.7}$$

with $\rho_{\mathbf{E}}(u) = \lim_{\eta\downarrow 0} \operatorname{Im} \mathrm{g}_{\mathbf{E}}^N(u - \mathrm{i}\eta)$ and $\mathrm{g}_{\mathbf{E}}^N$ is the unique solution in $\mathbb{C}^+$ of the discretized Marčenko–Pastur equation (3.11)

$$\mathrm{g}_{\mathbf{E}}^N(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z - \mu_i(1 - q + qz\mathrm{g}_{\mathbf{E}}^N(z))}. \tag{8.8}$$

Even if the numerical scheme seems quite intricate, all these quantities are simply a discretized version of the Marčenko–Pastur equation. Indeed, Eq. (8.5) is equivalent to Eq. (3.11) for large N and (8.6) is nothing but a discrete estimator of Eq. (3.40).

Finally, the optimization program reads

$$\tilde{\boldsymbol{\mu}} := \begin{cases} \operatorname*{argmin}_{\boldsymbol{\mu}\in\mathbb{R}_+^N} \sum_{i=1}^{N}\Big[\tilde{\gamma}_i(\boldsymbol{\mu}) - \lambda_i\Big]^2, \\ \text{s.t. } \tilde{\gamma}_i(\boldsymbol{\mu}) \text{ satisfies Eqs. (8.6)–(8.8).} \end{cases} \tag{8.9}$$

From there, the regularization scheme of the empirical RIE (6.26) reads

$$\xi_i^{\text{QuEST}} = \frac{\lambda_i}{|1 - q + q\lambda_i \lim_{\eta\downarrow 0}\tilde{\mathrm{g}}_{\mathbf{E}}^N(\lambda_i - \mathrm{i}\eta)|^2}, \tag{8.10}$$

where $\tilde{\mathrm{g}}_{\mathbf{E}}^N(z) \in \mathbb{C}^+$ is the unique solution of

$$\tilde{\mathrm{g}}_{\mathbf{E}}^N(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z - \tilde{\mu}_i(1 - q + qz\tilde{\mathrm{g}}_{\mathbf{E}}^N(z))}. \tag{8.11}$$

We see that the above regularization scheme allows one to estimate – in principle – the limiting RIE (6.5) since we can now set η to be arbitrarily small. This means that, contrary to the empirical estimate (6.26), the QuEST procedure should not suffer from a systematic underestimation at the left edge. 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 (8.9) 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 of the implementation of QuEST is given in [146], where the authors advise to sort the cleaned eigenvalues $[\xi_i^{\text{QuEST}}]_{i\in[\![1,N]\!]}$ since, as said above, we expect the optimal cleaned eigenvalues to be monotonic with respect to the sample eigenvalues.

8.1.4. Empirical studies

We compare in Fig. 28 the above QuEST numerical scheme with the simple IW-regularization of Section 8.1.2. The eigenvalues coming from the QuEST regularization are shown as green lines 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 (8.5) 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 in the power law example (see Fig. 28(d)). On the other hand, and perhaps surprisingly, the much simpler, somewhat ad-hoc IW-regularization given in Algorithm 1 provides very similar results. However, the QuEST method requires solving a nonlinear and non-convex optimization problem (see Eq. (8.9)) which implies heavy numerical computations that may not even converge to the global minimum (when 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 Fig. 28(d) since this simple prior allows use to generate very complex spectrum with possibly "clustered" outliers, similar to financial data. We emphasize again 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 (6.2). Using the mean squared error (MSE), we may also analyze the relative performance (RP) in percentage

Table 1
We reconsider the setting of Fig. 28(d) and check the consistency over 100 samples. The population density $\rho_{\mathbf{C}}$ is drawn from (6.28) 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 (6.2), stdev stands for the standard deviation of the squared error and the RP defined in Eq. (8.12). 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 (s)
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

Table 2
Check of the consistency of the three regularizations with respect to the dimension N. The population density $\rho_{\mathbf{C}}$ is drawn from (6.28) 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 (6.2) 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)	0.44 (0.15)	0.5 (0.13)

Table 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 (6.28) 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 (6.2) 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)

compared to the sample covariance. This is given by

$$\mathrm{RP}(\Xi) := 100 \times \left(1 - \frac{\mathbb{E}\|\Xi - \Xi^{\mathrm{ora.}}\|_2}{\mathbb{E}\|\mathbf{E} - \Xi^{\mathrm{ora.}}\|_2}\right), \tag{8.12}$$

where $\Xi \equiv \Xi(\mathbf{E})$ is a RIE of $\mathbf{C}$ and $\Xi^{\mathrm{ora.}}$ is the Oracle estimator. We also report in each case the average computational time needed to perform the estimation.[30]

First, let us assess the usefulness of sorting the cleaned eigenvalues. We report in Table 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 1 that it is indeed better to sort the eigenvalues when using the IW-regularization (8.4) as the difference is 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 statistically 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 matrix is very substantial.

We now investigate how these conclusions change when N varies with $q = 0.5$ fixed. The results are given in Table 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 $N \geqslant 100$, 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 s to get the regularized RIE while the QuEST algorithm requires 80 seconds on 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 (8.9) becomes very restrictive, while improvement over the simple IWs method is no longer significant.

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 as in Table 2 and the results are reported in Table 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 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 Section 9.

To conclude, we observed using synthetic data that we are able to estimate accurately the Oracle estimator for finite 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 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

[30] Simulations were implemented in Python and based on an Intel® Core™ i7-4700HQ and CPU of 8 × 2.40 GHz processor.

and non-convex optimization problem (8.9) becomes insignificant as N increases (see Tables 2 and 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.

8.2. Optimal RIE and out-of-sample risk for optimized portfolios

As alluded above (see Section 7.1), the concept of correlations between different assets is the cornerstone of Markowitz' optimal portfolio theory, and more generally for risk management purposes [157]. 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 7.1.3 that the best estimator inside the space of estimators restricted to possess the sample eigenvectors is precisely the Oracle estimator (6.2) 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 [25] 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 of historical returns, or simply implied volatilities from option markets, and the reader can choose his/her favorite 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

$$\widehat{\sigma}_{it} := \sqrt{\sum_j r_{jt}^2}, \tag{8.13}$$

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 σ_i is the sample estimator of the $\widetilde{r}_i$ which is now, to a first approximation, stationary.

We may now compute the sample covariance matrix $\mathbf{E}$ as in Eq. (3.3). 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 normalization by the cross-sectional volatility can be seen as a proxy for a robust estimator of the covariance matrix (3.8) with $U(x) = x^{-1}$ which can be studied using the tools of Sections 3 and 4 (see Section 3.1.3 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 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. We plot in Fig. 29 the average nonlinear shrinkage curve for the IWs-regularization (blue line) and for the QuEST regularization (red dashed line) – where we sorted the eigenvalues in both cases – and compare it with the estimated population eigenvalues obtained from (8.9). We see that IWs-regularization and QuEST still yield very similar results. Furthermore, we notice that the spectrum of the cleaned eigenvalues is, as expected, narrower than the spectrum of the (estimated) population matrix.

Interestingly, the Oracle estimator (6.2) can be estimated empirically and used to directly test the accuracy of the IWs-regularized RIE (8.4). The trick is to remark that the Oracle eigenvalues (6.2) can be interpreted as the "true" (out-of-sample) risk associated to a portfolio whose weights are given by the ith eigenvector. Hence, assuming that the data generating process is stationary, we estimate the Oracle estimator through the realized risk associated to such eigen-portfolios [136]. More precisely, we split the total length of our time series T_{tot} into n consecutive, non-overlapping samples of length T_{out}. The "training" period has length T, so n is given by:

$$n := \left\lfloor \frac{T_{\text{tot}} - T - 1}{T_{\text{out}}} \right\rfloor. \tag{8.14}$$

The Oracle estimator (6.2) is then computed as:

$$\hat{\xi}_i^{\text{ora.}} \approx \frac{1}{n}\sum_{j=0}^{n-1} \mathcal{R}^2_{\text{out}}(t_j, \mathbf{u}_i) \quad i = 1, \ldots, N, \tag{8.15}$$

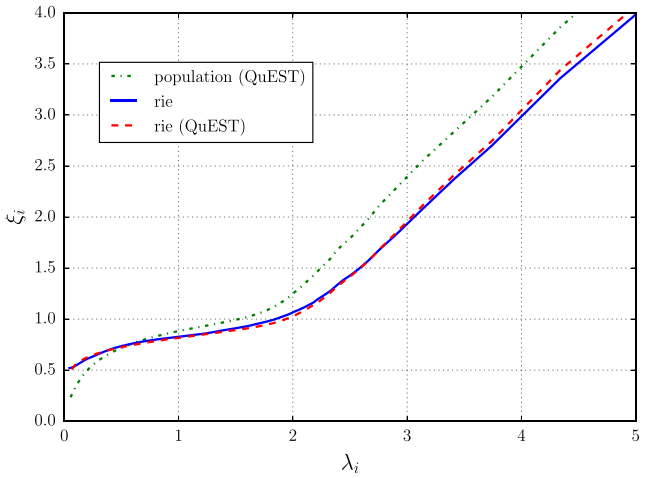


Fig. 29. Comparison of the IWs-regularization (8.4) (blue) with the QuEST procedure (8.10) (red dashed line) using 500 US stocks from 1970 to 2012. The agreement between those two regularizations is quite remarkable. We also provide the estimation of the population eigenvalues obtained from (8.9) (green dashed–dotted line).

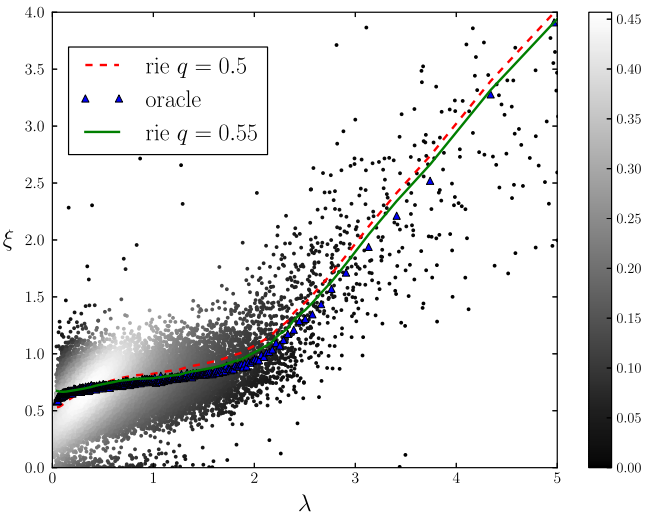


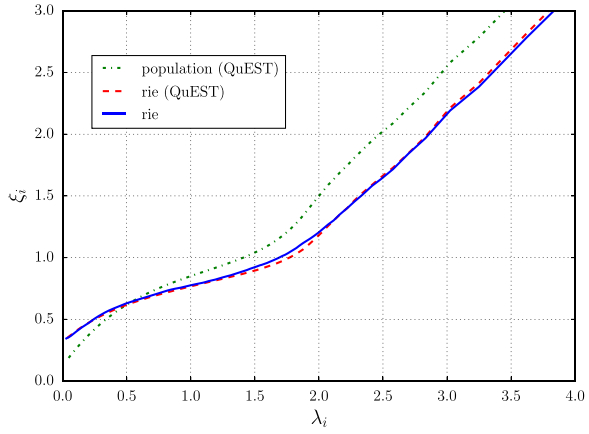
Fig. 30. Comparison of the IWs-regularized RIE (8.4) with the proxy (8.15) using 500 US stocks from 1970 to 2012. The points represent the density map of each realization of (8.15) and the color code indicates the density of data points. The average IWs-regularized RIE is plotted with the red dashed line and the average realized risk in blue. We also provide the prediction of the IWs-regularized RIE with an effective observation ratio q_{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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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

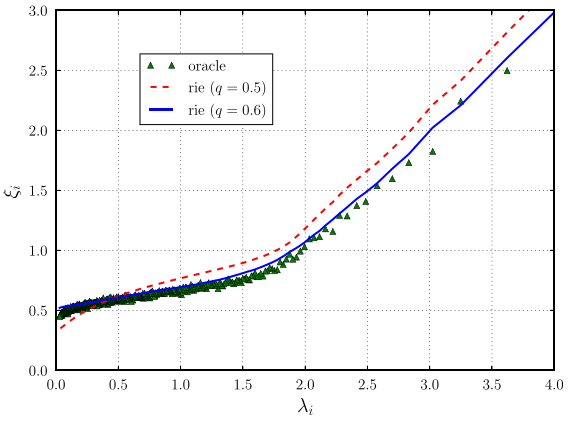
$$\mathcal{R}^2_{\text{out}}(t, \mathbf{w}) := \frac{1}{T_{\text{out}}} \sum_{\tau=t+1}^{t+T_{\text{out}}} \left(\sum_{i=1}^{N} \mathbf{w}_i Y_{i\tau} \right)^2, \tag{8.16}$$

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 (8.15) using US data in Fig. 30, which we compare with the IWs-regularized RIE. The results are, we believe, quite remarkable: the RIE formula (8.4) (red dashed line) tracks very closely the average realized risk (blue triangles), especially 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 Fig. 31(a) the estimation of the population eigenvalues (using Eq. (8.9)) and the regularized RIE (using Algorithm 1 or Eq. (8.10)). Again, the results we get from the simple IWs-regularization and QuEST procedure are nearly indistinguishable.

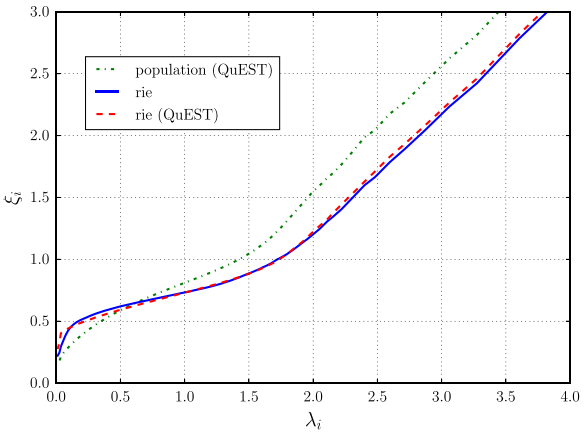


(a) Population and optimal RIE bulk eigenvalues.

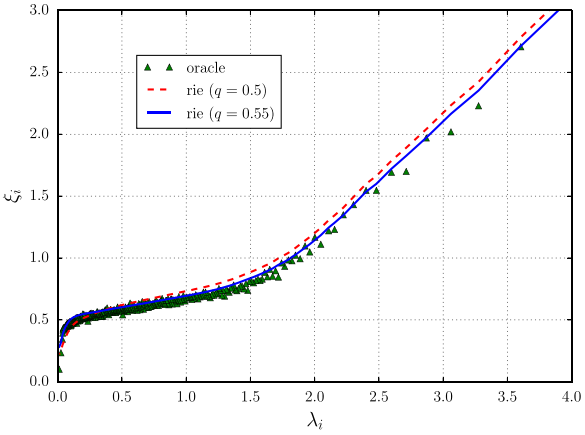


(b) Comparison with Oracle estimator (8.15).

Fig. 31. Left figure: analysis of the population (green dashed line) and optimal RIE bulk eigenvalues (red dashed line for Eq. (8.10) 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 (8.15) (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 (8.15).

Fig. 32. Left figure: analysis of the population (green dashed line) and optimal RIE bulk eigenvalues (red dashed line for Eq. (8.10) 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 (8.15) (green triangle). We also provide the plot of the IWs-regularized RIE with an effective observation ratio (blue line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This is another manifestation of the robustness of both algorithms at a finite N. We then plot in Fig. 31(b) the comparison between the IWs-regularized RIE (red dashed line) and the Oracle estimator, approximated by (8.15) (green triangles). We observe that the overall estimation is not as convincing as for US stocks (Fig. 30) but as above, the deviation can be explained by the presence of weak autocorrelations in the return time series (more on this below). Indeed, there exists an effective ratio $q_{\text{eff}} = 1.2q$ such that the estimation is extremely good (see blue line in Fig. 31(b)).

Finally we look at European stocks where the conclusions are similar than for the US stocks. In particular, we notice in Fig. 32(b) 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 quantities. This study highlights that the optimal RIE is robust with respect to the data generating process, as financial stock markets are certainly not Gaussian. The cross sectional volatility estimator (8.13) does not remove entirely heteroskedastic effects, nor 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 model of $\mathbf{E}$. The presence of autocorrelations has been shown to widen the spectrum of the sample matrix $\mathbf{E}$ [86]. We shall come back to the open problem of calibrating q_{eff} on empirical data in Section 9. It would be

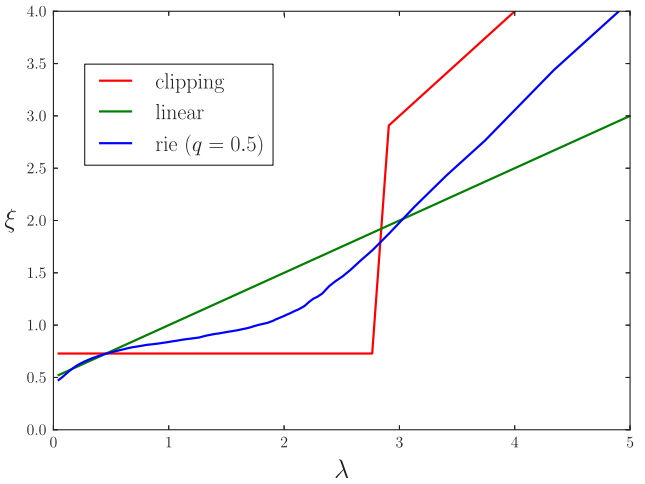


Fig. 33. Comparison of the de-biased RIE (8.4) (blue line) with clipping at the edge of the Marčenko–Pastur (red dashed line) and the linear shrinkage with $\alpha = 0.5$ (green dotted line). We use here the same dataset as in Fig. 30. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

interesting to quantify the information kept by the optimal RIE compared to other estimators using e.g. the Kullback–Leibler distance as in [158,102].

8.3. Out-of-sample risk minimization

It is interesting to compare the different shrinkage functions that map the empirical eigenvalues λ_i onto their "cleaned" counterparts $\hat{\xi}_i$. We show these functions in Fig. 33 for the three schemes we retained here, i.e. linear shrinkage, clipping and RIE, using the same dataset as in Fig. 30. 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 λ_i's is not captured. Furthermore, the larger eigenvalues are systematically overestimated. For the linear shrinkage (green dotted line), it is immediate from Fig. 33 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:

$$\mathbf{w} := \frac{\widehat{\mathbf{\Sigma}}^{-1}\mathbf{g}}{\mathbf{g}^* \widehat{\mathbf{\Sigma}}^{-1}\mathbf{g}}, \tag{8.17}$$

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 $i, j \in [\![1, N]\!]$. 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 volatility in the out-of-sample 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 $g_i = \mathcal{N}\, \widetilde{r}_{i,t}(T_{\text{out}})$ where $r_{i,t}(\tau) = (P_{i,t+\tau} - P_{i,t})/P_{i,t}$ with $P_{i,t}$ the price of the ith 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 normalization 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. (8.16) 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 4, for the three above cleaning schemes and the three geographical zones, keeping the same value of T (the learning period) and T_{out} (the out-of-sample period) as above. The linear shrinkage estimator uses a shrinkage intensity α estimated from the data following [159] (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 ($\alpha_s = 1$).

Table 4
Annualized average volatility (in %) of the different strategies. Standard deviations are given in bracket.

$\langle\mathcal{R}\rangle_e$	US	Japan	Europe
Minimum variance portfolio			
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			
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			
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			
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)

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. This is in agreement with previous work of Pantaleo et al. [160]; (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 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 valid in all cases as soon as $N \geqslant 300$. We see that apart from some fluctuations for $N = 100$, the result for out-of-sample test with the RIE is robust to the dimension N as indicated in Table 5.

8.4. 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 4.2.

Let us reconsider the two-sample self-overlap formula (4.41) for which the key object is the *limiting* Stieltjes transform (4.36). As we saw in Section 8.1.2, using the "raw" empirical Stieltjes transform yields a systematic bias for small eigenvalues which can be problematic when applying Eq. (4.41). Hence, we shall split the numerical computation of the overlap formula (4.40) or (4.41) into two steps. The first step is to estimate the population eigenvalues using the QuEST method of Ledoit and Wolf (see Section 8.1.3). Since these eigenvalues are designed to solve the Marčenko–Pastur equation, the second step consists in extracting from Eq. (8.8) an estimation of the Stieltjes transform of $\mathbf{E}$ for an arbitrarily small imaginary part η, 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. (4.36) allows us to obtain the overlaps.

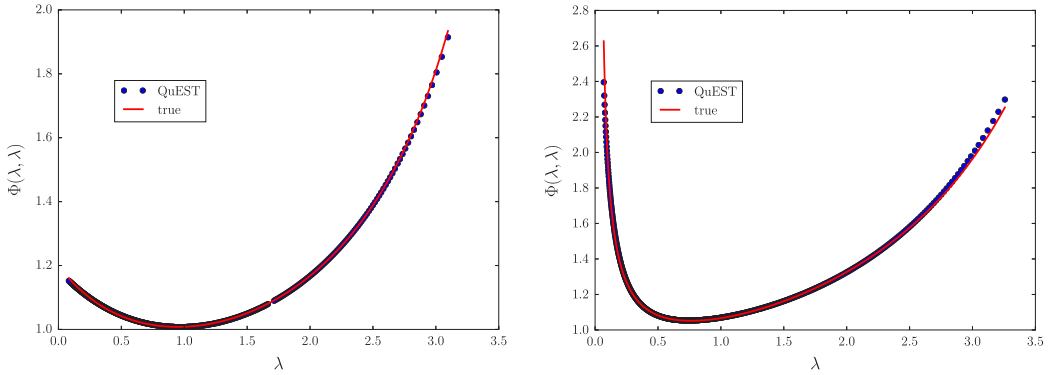
8.4.1. Synthetic data

We test this procedure on synthetic data first. Our numerical procedure is as follows. As in Section 4.2, we consider 100 independent realizations of the Wishart noise $\mathcal{W}$ with parameter T and covariance $\mathbf{C}$. Then, for each pair of samples, we compute the smoothed overlaps as:

$$\langle \mathbf{u}_i, \tilde{\mathbf{u}}_i \rangle^2 = \frac{1}{Z_i} \sum_{j=1}^{N} \frac{\langle \mathbf{u}_i, \tilde{\mathbf{u}}_j \rangle^2}{(\lambda_i - \tilde{\lambda}_j)^2 + \eta^2}, \tag{8.18}$$

Table 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.

N	US			Japan			Europe		
	100	200	300	100	200	300	100	200	300
Minimum variance portfolio									
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									
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									
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									
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)



(a) Inverse Wishart ($\kappa = 10$). (b) $\mathbf{I}_N$+ GOE ($\sigma = 0.35$).

Fig. 34. Evaluation of the self-overlap $\Phi(\lambda,\lambda)$ as a function of the sample eigenvalues λ 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 (4.41) using analytical solution (red line) and the estimated from the sample eigenvalues using QuEST algorithm (blue points).

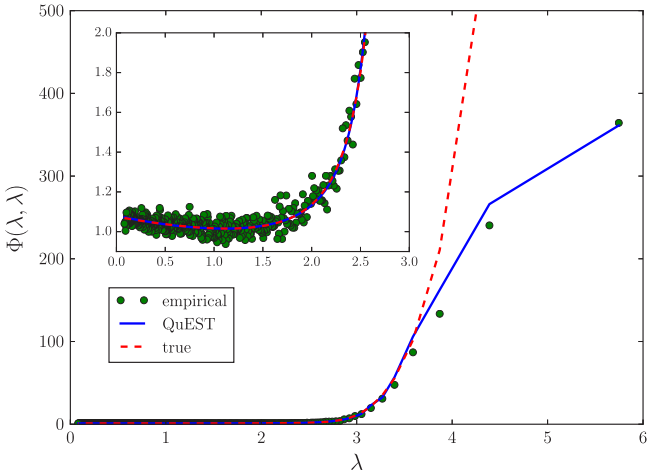


Fig. 35. Main figure: Evaluation of the self-overlap $\Phi(\lambda,\lambda)$ as a function of the sample eigenvalues λ when $\rho_{\mathbf{C}}$ is obtained from the power law proxy (6.28) with $\lambda_0 = 0.8$. We compare the analytical true solution using Eq. (3.50) (red dashed line) with the QuEST estimation (blue plain line) and also an empirical estimate over 100 realizations of $\mathbf{E}$ using Eq. (2.38) (green points). Inset: zoom in the bulk region of the main figure.

with $Z_i = \sum_{k=1}^{N}((\lambda_i - \tilde{\lambda}_k)^2 + \eta^2)^{-1}$ the normalization constant and η 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. (4.4) 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 (4.41) using the sample eigenvalues. We compare in Fig. 34 the estimation that we get using QuEST algorithm (blue points) with the limiting "true" analytical solution (4.46) (red line) and we see that the fit is indeed excellent. The same conclusion is reached when $\mathbf{C}$ is a GOE centered around the identity matrix.

Next, we proceed to the same test using the power law distribution proxy (6.28) for $\rho_{\mathbf{C}}$ with $\lambda_0 = -0.6$ (see Eq. (3.49) for the precise definition of λ_0). We emphasize again that this model is quite complex since it naturally generates a finite number of outliers. The result is reported in Fig. 35 where we plotted the self-overlap obtained by the limiting exact spectral density using Eq. (3.50) (red dashed line), the QuEST algorithm (blue plain line) and the empirical estimate (8.18) 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 (8.5) 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.

8.4.2. 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

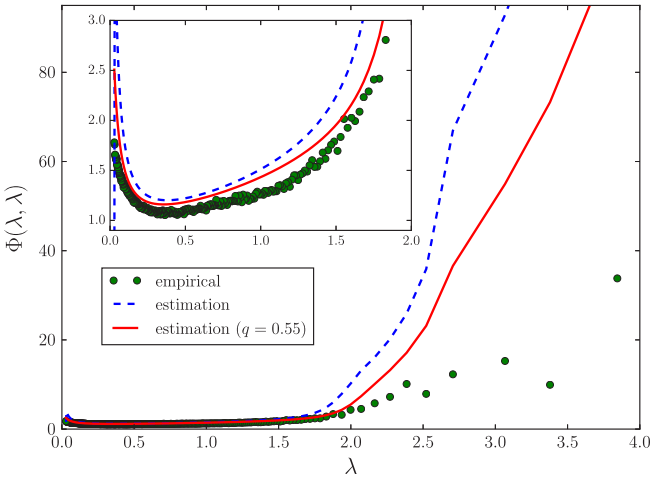


Fig. 36. Evaluation of the self-overlap $\Phi(\lambda, \lambda)$ as a function of the sample eigenvalues λ using the $N = 300$ most liquid US equities from 2004 to 2013. We split the data into two non-overlapping periods with same sample size 1200 business days. For each period, we randomly select $T = 600$ days and we repeat $B = 100$ bootstraps of the original data. The empirical self-overlap is computed using Eq. (8.18) over these 100 bootstraps (green points) and the limiting formula (4.41) 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_{eff} than in Fig. 30. Inset: focus in the bulk of eigenvalues.

correlation matrices $\mathbf{E}$ and $\mathbf{E}'$, as in Eq. (8.18), because we do not have enough data points to evaluate accurately an average over the noise as required in Eq. (4.4). To bypass this problem, we use a Bootstrap procedure to increase the size of the data.[31] 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 Bootstrap sample $b \in \{1, \ldots, B\}$, 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 (4.4) and also the theoretical limit (4.40) – using QuEST algorithm – from these B bootstrap data-sets.

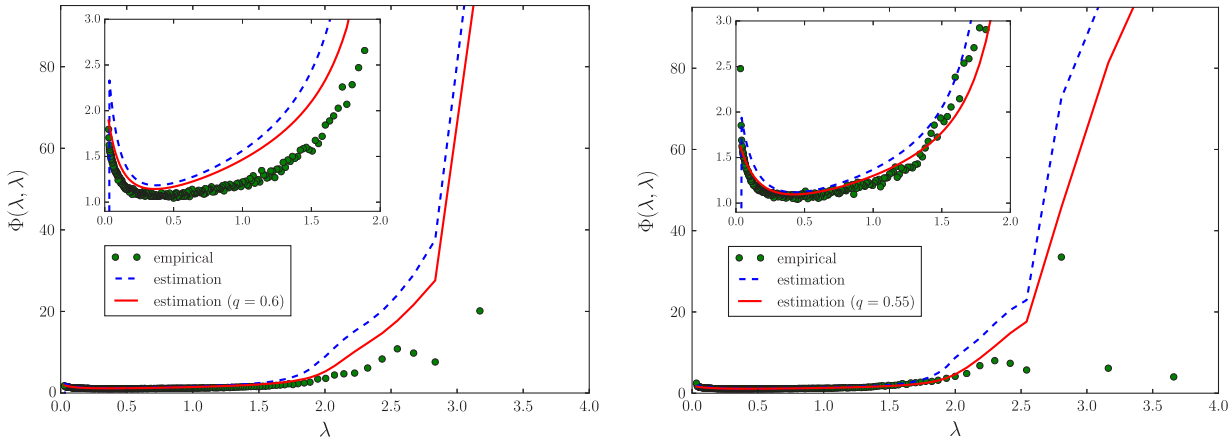
For our simulations, we set $B = 100$ and plot in Fig. 36 the resulting estimation of Eq. (4.4) we get from the QuEST algorithm (blue dashed line) and the empirical bootstrap estimate (8.18) (green points) using US stocks. We also perform the estimation with an effective observation ratio q_{eff} (red plain line) where we use for each market the values of q_{eff} obtained above (see Figs. 30-31(b)-32(b)). Note that the behavior in bulk is quite well estimated by the asymptotic prediction Eq. (4.41) for both periods. This is consistent with the conclusion of Fig. 30.

It is however clear from Fig. 36 that the eigenvectors associated to large eigenvalues are 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_{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 [161]. Note also the discrepancy at the left edge between the theoretical and empirical data in Fig. 36, which can be partly corrected using the effective ratio q_{eff}. This suggests that one can still improve the Marčenko–Pastur framework by adding e.g. autocorrelation or heavy tailed entries which allows one to widen the LSD of $\mathbf{E}$ (see e.g. [86,143] for autocorrelation and [102,142,103] for heavy tailed entries).

All the above results can be extended and confirmed in the case of Japanese and European stocks, for which the results are plotted respectively in Figs. 37(a) and 37(b).

To conclude, these observations suggest further improvements upon the time independent framework of Marčenko and Pastur, that would allow one to account for some "true" dynamics of the underlying correlation matrix. Such dynamics exist for eigenvectors corresponding to the largest eigenvalues is intuitively reasonable, and empirically confirmed by the analysis of Ref. [161]. The full correlation matrix might in fact evolve and jump between different "market states", as suggested in various recent papers of the Guhr group (see e.g. [162,163] 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_{eff} which was systematically found to be larger than $q = N/T$. This could be an indication of non-stationarity effects. This is particularly apparent for the Japanese stocks (see e.g. Fig. 37(a)) where the theoretical prediction deviates significantly from the empirical one even if we calibrate the effective quality ratio q_{eff}. The case of eigenvectors associated to the small eigenvalues is particularly striking and probably need further scrutiny, in particular in the case of future 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 Section, for further investigations.

31 This technique is especially useful in machine learning and we refer the reader to e.g. [4, Section 7.11] for a more detailed explanation.



(a) TOPIX (Japan). (b) Bloomberg 500 (Europe).

Fig. 37. Evaluation of the self-overlap $\Phi(\lambda, \lambda)$ as a function of the sample eigenvalues λ 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. (8.18) over these 100 bootstraps (green points) and the limiting formula (4.41) 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 Fig. 30. Inset: focus in the bulk of eigenvalues.

9. Conclusion and perspectives

In this review, we have discussed some of the most advanced techniques in RMT and their usefulness for estimating 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 review with an (incomplete) list of potentially interesting open problems that represent natural extensions of the results obtained above.

9.1. Extension to more general models of covariance matrices

One important assumption of the sample covariance matrix model (3.3) is the absence of temporal correlations and/or temporal structure in the data. However, this assumption does not hold in most real life applications (see e.g. Section 8.4). It is thus natural to extend the present work to estimators that account 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 [86,142,143]:

$$\mathbb{E}[Y_{it} Y_{jt'}] = C_{ij} \exp\big[-|t - t'|/\tau\big], \tag{9.1}$$

where τ controls the range of the time correlations.

Another frequent situation is when covariances are measured through an *Exponential Weighted Moving Average* (EWMA)[142,164][32]:

$$M_{ij}(\tau, T) = (1 - \alpha) \sum_{t=0}^{T} \alpha^t Y_{i,\tau-t} Y_{j,\tau-t}, \tag{9.2}$$

where τ 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 (9.2) can be rewritten as

$$M_{ij}(\tau) = \sum_{t=0}^{T} H_{it} H_{jt}, \quad \text{with} \quad \mathbb{E}\big[H_{it} H_{it'}\big] = \delta_{tt'}(1 - \alpha)\alpha^t, \tag{9.3}$$

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.

[32] We denote in the following the different estimators of $\mathbf{C}$ by $\mathbf{M}$ to avoid confusion with Pearson's sample estimator $\mathbf{E} = \mathbf{X}\mathbf{X}^*/T$.

Again, this assumption may not be satisfied in real dataset, especially in finance [104]. As alluded in Section 3.1.3, a more robust estimate of the covariance matrix in then needed [101]. 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 σ_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 [101]:

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

where U is a non-increasing function. As mentioned in Section 3.1.3, it is possible to show that for the $U(x) = x^{-1}$, one has $\mathbf{M} \to \mathbf{E}$ in the large N limit [102,103,107,108], where $\mathbf{E} = \mathbf{C}^{1/2} \mathcal{W} \mathbf{C}^{1/2}$ and $\mathcal{W}$ is a Wishart matrix. However, the asymptotic limit is more complex for general U's and reads:

$$\mathbf{M} \to \mathbf{C}^{1/2} \mathbf{X} \mathbf{B} \mathbf{X}^* \mathbf{C}^{1/2}, \tag{9.4}$$

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. [108] for the exact expression of the matrix $\mathbf{B}$).

Interestingly, all the above models, (9.1), (9.3) and (9.4), can be wrapped into a general multiplicative framework that reads:

$$\mathbf{M} := \mathbf{C}^{1/2} \mathbf{X} \mathbf{B} \mathbf{X}^* \mathbf{C}^{1/2}, \tag{9.5}$$

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 (9.1), we have $B_{tt'} = \exp[-|t - t'|/\tau]$ while we set $B_{tt'} = \delta_{tt'}(1 - \alpha)\alpha^t$ for (9.3).

The optimal RIE for this model has been briefly mentioned in Section 6.6 and can be found in exquisite details in [38]. We saw that the oracle estimator associated to the model (9.5) converges – at least for bulk eigenvalues – to a limiting function that does not depend explicitly on the spectral density of $\mathbf{C}$ (see Eq. (6.31)). It is thus interesting to see whether one of the aforementioned models can be solved in full generality using e.g. the results of [86] for the model (9.1) and whether one can explain the appearance of an effective ratio $q_{\text{eff}} > q$, as encountered in Section 8. Furthermore, another important result would be to see whether the estimator (6.31) is also valid for outliers, as is the case for the time-independent sample covariance matrices.

9.2. Singular value decomposition

A natural extension of the work presented in this review 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 $\mathcal{C}$ defined as:

$$\mathcal{C}_{ia} := \mathbb{E}[x_i y_a], \tag{9.6}$$

where we assumed that both quantities have zero mean and unit variance.

What can be said about the structure of this rectangular and non symmetric correlation matrix (9.6)? 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 [58] or [165,166] 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 $\mathcal{C}\mathcal{C}^*$, which is now symmetric and has N non negative eigenvalues. Indeed, the trick behind this change of variable is that the eigenvalues of $\mathcal{C}\mathcal{C}^*$ are equal to the square of a singular value of $\mathcal{C}$ 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 $\mathcal{C}$, one forms the $M \times M$ matrix $\mathcal{C}^*\mathcal{C}$ that has exactly the same non zero eigenvalues as $\mathcal{C}\mathcal{C}^*$; 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 $\mathcal{C}^*\mathcal{C}$ has $M - N$ additional zero eigenvalues; whereas in the other case, it is $\mathcal{C}\mathcal{C}^*$ that has an excess of $N - M$ zero eigenvalues.

However, as for standard correlation matrices, the knowledge of the true population matrix Eq. (9.6) is unavailable. Hence, one resorts to an empirical determination of $\mathcal{C}$ that is strewn with measurement noise, as above. We expect to be able to use tools from RMT to understand 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 $\mathcal{C}$ is given by

$$\mathcal{E}_{ia} := \frac{1}{T}\sum_{t=1}^{T} X_{it} Y_{at}, \tag{9.7}$$

and the aim is to study the singular values of this matrix. Indeed, as in Section 3, 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 section, a convenient way to perform this analysis is to consider the eigenvalues of $\mathcal{E}\mathcal{E}^*$ (or $\mathcal{E}^*\mathcal{E}$). Using tools from Appendix B, especially Eq. (B.10), we see that

$$\det(\mathcal{E}\mathcal{E}^* - z\mathbf{I}_N) = \det\left(\mathbf{S}_\mathbf{X}\mathbf{S}_\mathbf{Y} - z\mathbf{I}_T\right), \qquad \mathbf{S}_\mathbf{X} := \frac{\mathbf{X}^*\mathbf{X}}{T}, \quad \mathbf{S}_\mathbf{Y} := \frac{\mathbf{Y}^*\mathbf{Y}}{T}$$

so that $\mathcal{E}\mathcal{E}^*$ shares the same non-zero eigenvalues than the product of the dual $T \times T$ samples covariance matrix $\mathbf{S}_\mathbf{X}$ and $\mathbf{S}_\mathbf{Y}$.

It is easy to see that when $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated, i.e. $\mathcal{C} = \mathbf{0}$, one can compute the spectral density of $\mathbf{S}_\mathbf{X}\mathbf{S}_\mathbf{Y}$ using the free multiplication formula (2.81). 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 [60] and [59] for an early derivation without using free probability methods), and is given by:

$$\rho(c) = \max(m + n - 1, 0)\delta(c - 1) + \mathrm{Re}\frac{\sqrt{(c^2 - \gamma_-)(\gamma_+ - c^2)}}{\pi c(1 - c^2)}, \tag{9.8}$$

where $\gamma_\pm$ are given by:

$$\gamma_\pm = n + m - 2mn \pm 2\sqrt{mn(1 - n)(1 - m)}, \quad 0 \leqslant \gamma_\pm \leqslant 1 \tag{9.9}$$

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 [165,60,166].

The general case where when $\mathbf{X}$ and $\mathbf{Y}$ are correlated, i.e. $\mathcal{C} \neq \mathbf{0}$, 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 (3.9). Moreover, an analysis of the left and right eigenvectors like in Section 4 would certainly be of interest in many real life problems (see e.g. [4,167–169] 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 [170].

9.3. 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 [153,155], ultra-metric tree models (*eigenvector clustering*) [171,172], or constraints on the participation ratios [41].

Very recently, an attempt to "clean" empirical outlier eigenvectors was formulated in [41]. Let us focus for example on the top eigenvector; the prior is then defined as a weighted sum of the sample eigenvectors:

$$\widehat{\mathbf{v}}_1 = \sqrt{\Phi(\mu_1, \lambda_1)}\,\mathbf{u}_1 + \sum_{j=2}^{N} \varepsilon_j \sqrt{\Phi(\mu_1, \lambda_j)}\mathbf{u}_j, \tag{9.10}$$

where the bivariate mean squared overlap Φ is defined in Eq. (4.3) 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. (9.10) can be computed using Eq. (4.14) and the second one can be inferred from Eq. (4.16). 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 [41] 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.

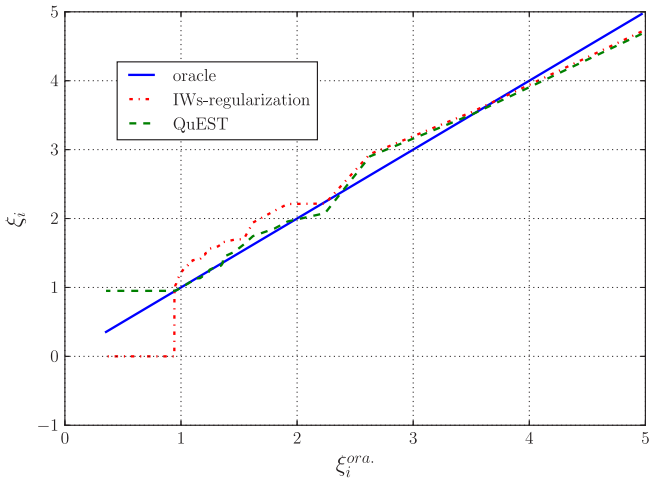


Fig. 38. We apply the IWs (red dash-dotted line) and QuEST (green dashed line) regularization of Section 8 as a function of the oracle estimator (6.2) with $\rho_{\mathbf{C}}$ given by Eq. (6.28) with $\lambda_0 = 0.8$ and $N = 1000$. The sample covariance matrix $\mathbf{E}$ is a Wishart matrix with $q = 2$. We see that both regularizations provide results that are far from the optimal solution (blue plain line).

9.4. Cleaning recipe for $q > 1$

As observed in Section 8, the optimal RIE (8.4) returns very satisfactory results in terms of estimating the oracle estimator either with synthetic 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 Section 8 fail to estimate correctly the small eigenvalues in this case (see Fig. 38). 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 $N - T$ zero 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 7.2. 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.

9.5. A Brownian motion model for correlated Wishart matrices

We present in Appendix D 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. (D.20) 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. (3.12) suggest that such a process should exist. In the case where $\mathbf{C} = \mathbf{I}_N$, Bru's Wishart process [173] allows one to obtain many interesting properties about both the eigenvalues and eigenvectors—see [72,174], 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. [126,161] for a related question in the additive model). As this review was being completed, we managed to characterize this process, and the reader is referred to [112] for details.

Acknowledgments

We want to warmly thank all our collaborators on the different topics considered in this review, in particular Romain Allez, Antti Knowles and Satya N. Majumdar. We also acknowledge insightful discussions with Marc Abeille, Jean-Yves Audibert, Yanis Bahroun, Florent Benaych-Georges, Raphaël Benichou, Alexios Beveratos, Giulio Biroli, Rémy Chicheportiche, Benoit Collins, Romain Couillet, Noureddine El Karoui, Sandrine Péché, Adam Rej, Emmanuel Sérié, Guillaume Simon and Denis Ullmo.

Appendix A. Harish-Chandra–Itzykson–Zuber integrals

A.1. Definitions and results

The (generalized) Harish-Chandra–Itzykson–Zuber (HCIZ) integral [88,89] $\mathcal{I}_\beta(\mathbf{A},\mathbf{B})$ is defined as:

$$\mathcal{I}_\beta(\mathbf{A},\mathbf{B}) = \int_{G(N)} \mathcal{D}\boldsymbol{\Omega}\; e^{\frac{\beta N}{2}\mathrm{Tr}\mathbf{A}\boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^\dagger}, \tag{A.1}$$

where the integral is over the (flat) Haar measure of the compact group $\boldsymbol{\Omega} \in \mathbf{G}(N) = \mathbf{O}(N)$, $\mathbf{U}(N)$ or $Sp(N)$ in N dimensions and $\mathbf{A}$, $\mathbf{B}$ are arbitrary $N \times N$ symmetric (hermitian or symplectic) matrices. The parameter β 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 [175]). 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 the unitary case $\mathbf{G}(N) = \mathbf{U}(N)$ 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:

$$\mathcal{I}_{\beta=2}(\mathbf{A},\mathbf{B}) = \frac{c_N}{N^{(N^2-N)/2}} \frac{\det\left((e^{Na_ib_j})_{1\le i,j\le N}\right)}{\Delta(\mathbf{A})\Delta(\mathbf{B})} \tag{A.2}$$

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 asymptotic 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 $\mathbf{A}$, $\mathbf{B}$ [176]. Using Dyson's Brownian motion, one can find [177,178]: $F_{\beta=2}(\mathbf{A},\mathbf{B}) = \lim_{N\to\infty} N^{-2}\ln \mathcal{I}_2(\mathbf{A},\mathbf{B})$:

$$F_2(\mathbf{A},\mathbf{B}) = -\frac{3}{4} - S_2(\mathbf{A},\mathbf{B}) + \frac{1}{2}\int \mathrm{d}x\, x^2(\rho_\mathbf{A}(x)+\rho_\mathbf{B}(x)) - \frac{1}{2}\int \mathrm{d}x\mathrm{d}y\,[\rho_\mathbf{A}(x)\rho_\mathbf{A}(y)+\rho_\mathbf{B}(x)\rho_\mathbf{B}(y)]\ln|x-y|,$$

where

$$S_2(\mathbf{A},\mathbf{B}) = \frac{1}{2}\int \mathrm{d}t \int d\lambda\, \rho(\lambda,t)\left\{v^2(\lambda,t) + \frac{\pi^2}{3}\rho^2(\lambda,t)\right\} \tag{A.3}$$

with $\rho(\lambda,t)$ and $v(\lambda,t)$ solution of the following Euler equation

$$\begin{cases} \partial_t\rho(\lambda,t) + \partial_\lambda[\rho(\lambda,t)v(\lambda,t)] = 0, \\ \partial_t v(\lambda,t) + v(\lambda,t)\partial_\lambda v(\lambda,t) = \dfrac{\pi^2}{2}\partial_\lambda\rho^2(\lambda,t), \\ \text{with } \rho(\lambda,0) = \rho_\mathbf{A}(\lambda), \text{ and } \rho(\lambda,1) = \rho_\mathbf{B}(\lambda). \end{cases} \tag{A.4}$$

In fact, this result can be extended to arbitrary value of β 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 $\beta = 1$ [179] (see also [180–182] for arbitrary β).

Nonetheless, explicit results concerning the asymptotic of this integral are scarce. When $\mathbf{A}$ and $\mathbf{B}$ are both Wigner matrices, the Euler–Matytsin system of equation can be solved explicitly [177]. Another soluble case is when one of the two matrix has a Flat distribution [180]. 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 $a_1, a_2, \ldots, a_n$ and $N-n$ zero eigenvalues. Then we have [183,178,182]:

$$\mathcal{I}_\beta(\mathbf{A},\mathbf{B}) = \exp\left[\frac{N\beta}{2}\sum_{i=1}^{n}\mathcal{W}_\mathbf{B}(a_i)\right], \tag{A.5}$$

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 2.4). 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.5) in the Rank-1 case

This section is devoted to the derivation of the result (A.5) in the sample case where $\mathbf{A} = \mathrm{diag}(a_1, 0, \ldots, 0)$ and $\mathbf{B} = \mathrm{diag}(b_1, \ldots, b_N)$. Firstly, we rewrite (A.1) (we set $\beta = 1$ for simplicity):

$$\mathcal{I}_1(\mathbf{A}, \mathbf{B}) = \frac{1}{\mathcal{Z}} \int \left(\prod_{k=1}^{N} \mathrm{d}\Omega_{1k} \right) \exp\left[\frac{N}{2} a_1 \sum_{k=1}^{N} \Omega_{1k}^2 b_k \right] \delta\left(\sum_{k=1}^{N} \Omega_{1k}^2 - 1 \right), \tag{A.6}$$

where the Dirac delta function enforces the orthogonality and $\mathcal{Z}$ is normalization constant defined as:

$$\mathcal{Z} := \int \left(\prod_{k=1}^{N} \mathrm{d}\Omega_{1k} \right) \delta\left(\sum_{k=1}^{N} \Omega_{1k}^2 - 1 \right), \tag{A.7}$$

which allows us to omit constant variables in the following. We then use the following integral representation of the delta function:

$$\delta\left(\sum_{k=1}^{N} \Omega_{1k}^2 - 1 \right) = \frac{1}{2\pi} \int \exp\left[\mathrm{i}\zeta \left(\sum_{k=1}^{N} \Omega_{1k}^2 - 1 \right) \right] \mathrm{d}\zeta, \tag{A.8}$$

so that we have (after renaming $\zeta \to -2\mathrm{i}\zeta/N$)

$$\begin{aligned} \mathcal{I}_1(\mathbf{A}, \mathbf{B}) &\propto \frac{N}{4\pi} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}\zeta \int \left(\prod_{k=1}^{N} \mathrm{d}\Omega_{1k} \right) \exp\left[\frac{N}{2} \left(a_1 \sum_{k=1}^{N} \Omega_{1k}^2 b_k + \zeta \left(\sum_{k=1}^{N} \Omega_{1k}^2 - 1 \right) \right) \right] \\ &= \frac{N}{4\pi} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}\zeta \exp\left[\frac{N\zeta}{2} \right] \int \left(\prod_{k=1}^{N} \mathrm{d}\Omega_{1k} \right) \exp\left[-\frac{N}{2} \sum_{k=1}^{N} \Omega_{1k}^2 \left(\zeta - a_1 b_k \right) \right] \\ &= \frac{N}{4\pi} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \exp\left[-\frac{N}{2} \left(\frac{1}{N} \sum_{k=1}^{N} \log(\zeta - a_1 b_k) - \zeta \right) \right] \mathrm{d}\zeta. \end{aligned} \tag{A.9}$$

Since we consider $N \to \infty$, the integral over ζ is performed by a saddle-point method, leading to the following equation:

$$\frac{1}{N} \sum_{k=1}^{N} \frac{1}{\zeta - a_1 b_k} = 1, \tag{A.10}$$

which is equivalent to

$$\mathfrak{g}_{\mathbf{B}}(\zeta/a_1) = a_1. \tag{A.11}$$

We therefore find that

$$\zeta = a_1 \mathcal{B}_{\mathbf{B}}(a_1) = a_1 \mathcal{R}_{\mathbf{B}}(a_1) + 1. \tag{A.12}$$

By plugging this solution into (A.9), we obtain

$$\frac{2}{N} \log \mathcal{I}_1(\mathbf{A}, \mathbf{B}) \sim a_1 \mathcal{R}_{\mathbf{B}}(a_1) - \frac{1}{N} \sum_{k=1}^{N} \log\Big(1 + a_1 (\mathcal{R}_{\mathbf{B}}(a_1 - b_k)) \Big). \tag{A.13}$$

One can then check, by taking the derivative of both sides, that

$$a_1 \mathcal{R}_{\mathbf{B}}(a_1) - \frac{1}{N} \sum_{k=1}^{N} \log\Big(1 + a_1 (\mathcal{R}_{\mathbf{B}}(a_1 - b_k)) \Big) = \mathcal{W}_{\mathbf{B}}(a_1), \tag{A.14}$$

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

$$\frac{2}{N} \log \mathcal{I}_1(\mathbf{A}, \mathbf{B}) \sim \mathcal{W}_{\mathbf{B}}(a_1), \tag{A.15}$$

which is the claim.

Let us now explain briefly how to extend this derivation to the rank-n case. Formally, the integral reads

$$\mathcal{I}_1(\mathbf{A}, \mathbf{B}) = \frac{1}{\mathcal{Z}} \int \left(\prod_{i=1}^{n} \prod_{k=1}^{N} \mathrm{d}\Omega_{ik} \right) \exp\left[\frac{N}{2} \sum_{i=1}^{n} a_i \sum_{k=1}^{N} \Omega_{ik}^2 b_k \right] \prod_{i,j=1}^{n} \delta\left(\sum_{k=1}^{N} \Omega_{ik} \Omega_{jk} - \delta_{ij} \right), \tag{A.16}$$

where the normalization $\mathcal{Z}$ is easily deduced from (A.7), 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 normalization 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.5). For a more rigorous proof that this result holds as long as $n \ll \sqrt{N}$, see [178].

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

$$\mathbf{M} := \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}, \tag{B.1}$$

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

$$\mathbf{M}/\mathbf{D} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}. \tag{B.2}$$

Using it, one obtains after using block Gaussian elimination (or LU decomposition) that the determinant of $\mathbf{M}$ can be expressed as

$$\det(\mathbf{M}) = \det(\mathbf{D})\det(\mathbf{M}/\mathbf{D}). \tag{B.3}$$

Moreover, one can write the inverse matrix $\mathbf{M}^{-1}$ in terms of $\mathbf{D}^{-1}$ and the inverse of the Schur complement (B.2)

$$\mathbf{M}^{-1} = \begin{pmatrix} (\mathbf{M}/\mathbf{D})^{-1} & -(\mathbf{M}/\mathbf{D})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{M}/\mathbf{D})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{M}/\mathbf{D})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}. \tag{B.4}$$

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

$$\mathbf{M}/\mathbf{A} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}. \tag{B.5}$$

One easily obtains $\det(\mathbf{M})$ in terms of $\mathbf{A}$ and $\mathbf{M}/\mathbf{A}$ from (B.3) by replacing $\mathbf{D}$ by $\mathbf{A}$

$$\det(\mathbf{M}) = \det(\mathbf{A})\det(\mathbf{M}/\mathbf{A}). \tag{B.6}$$

The inverse matrix $\mathbf{M}^{-1}$ can also be written in terms of $\mathbf{A}^{-1}$ and the inverse of the Schur complement (B.5)

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{M}/\mathbf{A})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{M}/\mathbf{A})^{-1} \\ -(\mathbf{M}/\mathbf{A})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{M}/\mathbf{A})^{-1} \end{pmatrix}. \tag{B.7}$$

B.2. Matrix identities

There are several useful identities that can be inferred from Schur complement formula. Firstly, using (B.4) and (B.7), we may immediately deduce the so-called *Woodbury* matrix identity

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}. \tag{B.8}$$

Moreover, if $\mathbf{D} = I_M$, we get the *matrix determinant lemma* from (B.3) and (B.6)

$$\det(\mathbf{A} - \mathbf{B}\mathbf{C}) = \det(\mathbf{A})\det(\mathbf{I}_M - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}), \tag{B.9}$$

and if $\mathbf{A} = \mathbf{I}_N$ in addition, one gets *Sylvester's determinant identity*

$$\det(\mathbf{I}_N - \mathbf{B}\mathbf{C}) = \det(\mathbf{I}_M - \mathbf{C}\mathbf{B}). \tag{B.10}$$

Now, assuming that both $\mathbf{B}$ and $\mathbf{C}$ are column vectors, one readily find from (B.8) the *Sherman–Morrison* formula.

B.3. Resolvent identities

Another useful application of Schur complement formula concerns the resolvent. We keep the notations of Section 2.1.2 and thus

$$\mathbf{G}(z) = \mathbf{H}^{-1}(z), \qquad \mathbf{H}(z) \coloneqq z\mathbf{I}_N - \mathbf{M}, \tag{B.11}$$

with $\mathbf{G}$ a $N \times N$ symmetric matrix. We now rewrite $\mathbf{H}(z)$ as a block matrix:

$$\mathbf{H}(z) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{pmatrix}, \tag{B.12}$$

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.2) the Schur complement $\mathbf{D} \coloneqq \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^*$. In the following, we consider $K = 2$ for simplicity. We have for any $i, j \in \{1, 2\}$, we have from (B.4):

$$G_{ij} = (\mathbf{D}^{-1})_{ij}. \tag{B.13}$$

As a warm-up 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.11), one obtains $\mathbf{A} = z - M_{11}$, $\mathbf{B} = [M_{12}, \ldots, M_{1N}]$ and $\mathbf{C} = \mathbf{H}^{(1)}(z)$ where $\mathbf{H}^{(i)}$ denotes the "minor" of $\mathbf{H}$, i.e. $\mathbf{H}^{(i)} \coloneqq \left(H_{st} : s, t \in [\![1, N]\!] \backslash \{i\}\right)$. Hence, it is easy to see from the very definition of $\mathbf{D}$ that

$$\mathbf{D} \equiv D_{11} = z - M_{11} - \sum_{\alpha,\beta}^{(1)} M_{1\alpha} G^{(1)}_{\alpha,\beta} M_{\beta 1}, \tag{B.14}$$

where and we used the abbreviation

$$\sum_{\alpha,\beta}^{(i)} \equiv \sum_{\alpha,\beta \in [\![1,N]\!] \backslash \{i\}}. \tag{B.15}$$

Therefore, we deduce from (B.13) that

$$G_{11}(z) = \frac{1}{z - M_{11} - \sum_{\alpha\beta}^{(1)} M_{1\alpha} G^{(1)}_{\alpha,\beta} M_{\beta 1}}. \tag{B.16}$$

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 2×2 matrix. Again, using the block representation (B.12) and Eq. (B.11), one deduces that:

$$D_{kl} = z\delta_{kl} - M_{kl} - \sum_{\alpha,\beta}^{(kl)} M_{k\alpha} G^{(kl)}_{\alpha,\beta} M_{\beta l}, \quad k, l \in [\![i, j]\!]. \tag{B.17}$$

It is not hard to see that D_{kk} yields Eq. (B.14) as it should. Using that (B.17) is a 2×2 matrix, one can readily invert the matrix $\mathbf{D}$ to obtain the relation

$$G_{ij} - G^{(m)}_{ij} = \frac{G_{im} G_{mj}}{G_{mm}}, \tag{B.18}$$

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).

Appendix C. Self-consistent relation for Green's function 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. (4.5) are actually a consequence of the CLT.

C.1. Wigner matrices

As a warm-up exercise, we consider the simplest ensemble of random matrices where all elements of the matrix $\mathbf{M}$ are i.i.d random variables, with the only constraint that the matrix be symmetrical. This is the well-known Wigner ensemble

where we assume that

$$\mathbb{E}[M_{ij}] = 0, \quad \mathbb{E}[M_{ij}^2] = \frac{\sigma^2}{N}, \tag{C.1}$$

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 (C.1) and Wick's theorem into (B.17) and one can check that

$$\mathbb{E}\left[\sum_{\alpha,\beta}^{(kl)} M_{k\alpha} G^{(kl)}_{\alpha\beta} M_{\beta l}\right] = \delta_{kl} \frac{\sigma^2}{N} \sum_{\alpha}^{(k)} G^{(k)}_{\alpha\alpha}$$

$$\mathbb{V}\left[\sum_{\alpha,\beta}^{(kl)} M_{k\alpha} G^{(kl)}_{\alpha\beta} M_{\beta l}\right] \sim \frac{\sigma^4}{N}. \tag{C.2}$$

Consequently, using the Central Limit Theorem, we conclude that for Wigner matrices, (B.17) converges for large N towards

$$D_{kl} = \delta_{kl}\left(z - \frac{\sigma^2}{N} \sum_{\alpha}^{(k)} G^{(k)}_{\alpha\alpha}\right) + O(N^{-1/2}) \quad k, l \in \{i, j\}, \tag{C.3}$$

from which one deduces that $G_{ij} \sim N^{-1/2}$ using (B.13). 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.18) and we therefore obtain for any $i \in [\![1, N]\!]$:

$$G_{ii} \sim \frac{1}{z - \sigma^2 \mathfrak{g}(z)} + O(N^{-1/2}). \tag{C.4}$$

By taking the normalized trace in this last equation, we obtain at leading order the equation of the semi-circle law's Stieltjes transform

$$\mathfrak{g}(z) = \frac{1}{z - \sigma^2 \mathfrak{g}(z)}, \tag{C.5}$$

so that we conclude

$$G_{ij}(z) \sim \delta_{ij} \mathfrak{g}(z) + O(N^{-1/2}). \tag{C.6}$$

This result has been extended in a much more general framework—see e.g. the recent reviews [91,184]. In particular, it is possible to show that the error term we obtain in Eq. (C.6) is quite similar to (4.7) and reads for $\eta = \widehat{\eta} N$ with $\widehat{\eta} \gg 1$:

$$\Psi_{\mathrm{GOE}}(z) := \sqrt{\frac{\operatorname{Im} \mathfrak{g}_{\mathbf{S}}(z)}{\widehat{\eta}}} + \frac{1}{\widehat{\eta}}, \tag{C.7}$$

provided that N is large enough. We illustrate this ergodic behavior for the GOE in Fig. 11, and we see the agreement is excellent and each diagonal entry indeed converges to the semicircle law (see Fig. C.1).

C.2. Sample covariance matrices

We now want to derive (4.5) using the same type of arguments than in the previous section. Suppose that $\mathbf{E}$ is defined as in (3.3) and we denote by $\mathbf{G}(z)$ its resolvent. Let us assume for simplicity that $\mathbf{C} = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_N)$. 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:

$$\mathbf{R}(z) := \mathbf{H}^{-1}(z), \quad \mathbf{H}(z) := \begin{pmatrix} \mathbf{C}^{-1} & \mathbf{X} \\ \mathbf{X}^* & z\,\mathbf{I}_T \end{pmatrix}. \tag{C.8}$$

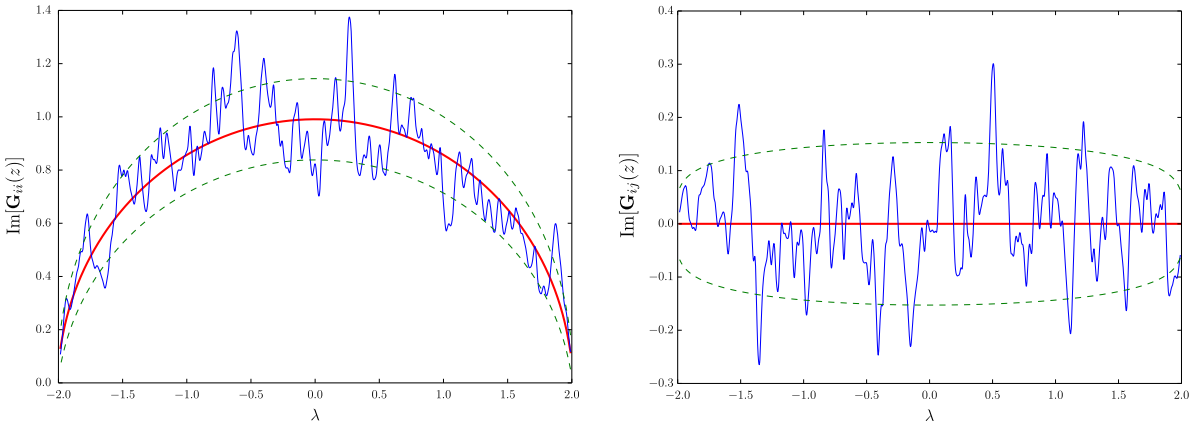
To simplify the notations, we introduce the set of indexes $\mathcal{I}_N := [\![1, N]\!]$ and $\mathcal{I}_T := [\![1, T]\!]$. Then using (B.4) and (B.7), we see that

$$R_{ij}(z) = z(\mathbf{C}^{1/2} \mathbf{G}_{\mathbf{E}}(z) \mathbf{C}^{1/2})_{ij}, \quad i, j \in \mathcal{I}_N, \tag{C.9}$$

where $\mathbf{E}$ is the sample covariance matrix defined in Eqs. (3.3) and (3.4), but also

$$R_{\alpha\beta}(z) = (\mathbf{G}_{\mathbf{S}}(z))_{\alpha\beta}, \qquad \alpha, \beta \in \mathcal{I}_T, \tag{C.10}$$

where the $T \times T$ matrix $\mathbf{S}$ is defined in Eq. (3.32).



(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 $j = 2$.

Fig. C.1. Illustration of the imaginary part of Eq. (C.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. (C.6). The green dotted corresponds to the confidence interval whose formula is given by Eq. (C.7). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We are interested in the computations of R_{ij} for $i, j \in \mathcal{I}_N$ and this can be done using (B.13) and (B.17). Note that one can find $R_{\alpha\beta}$ by proceeding in the same way. We obtain from (B.13) and (B.17) that

$$R_{ij}(z) = (\mathbf{D}^{-1})_{ij}, \qquad D_{kl} := \frac{\delta_{kl}}{\mu_k} - \sum_{\alpha,\beta \in \mathcal{I}_T} X_{k\alpha} R^{(kl)}_{\alpha\beta} X_{l\alpha} \tag{C.11}$$

for any $k, l \in \{i, j\}$. Using that $\mathbb{E}[X_{it}] = 0$ and $\mathbb{E}[X_{it}^2] = T^{-1}$ from (3.5), we remark thanks to Wick's theorem that the sum in the term D_{kl} obeys

$$\begin{aligned} \mathbb{E}\Big[\sum_{\alpha,\beta \in \mathcal{I}_T} X_{k\alpha} R^{(kl)}_{\alpha\beta} X_{l\alpha} \Big] &= \frac{\delta_{kl}}{T} \sum_{\alpha}^{(k)} R^{(k)}_{\alpha\alpha} \\ \mathbb{V}\Big[\sum_{\alpha,\beta \in \mathcal{I}_T} X_{k\alpha} R^{(kl)}_{\alpha\beta} X_{l\alpha} \Big] &\sim \frac{1}{T}, \end{aligned} \tag{C.12}$$

where we used the notation (B.15) for the sum. Invoking once again the CLT, we find that the entry D_{kl} converges for large N towards

$$D_{kl} \sim \delta_{kl} \Big(\frac{1}{\mu_k} - \frac{1}{T} \sum_{\alpha \in \mathcal{I}_T} R^{(k)}_{\alpha\alpha} \Big) + O(T^{-1/2}), \tag{C.13}$$

so that we may conclude from (C.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

$$D_{\alpha\beta} \sim \delta_{\alpha\beta} \Big(z - \frac{1}{T} \sum_{k \in \mathcal{I}_N} R^{(\alpha)}_{kk} \Big) + \mathcal{O}(T^{-1/2}). \tag{C.14}$$

Let us now investigate $R^{(k)}_{\alpha\alpha}$ which can be rewritten thanks to (B.18) as:

$$R^{(k)}_{\alpha\alpha} = R_{\alpha\alpha} - \frac{R_{k\alpha} R_{\alpha k}}{R_{kk}}. \tag{C.15}$$

We deduce from (C.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.7) to (C.8) to find

$$R_{k\alpha} = -\big(\mathbf{C}\mathbf{X}\mathbf{G}_{\mathbf{S}}\big)_{k\alpha} = -\mu_k \sum_{\beta \in \mathcal{I}_T} X_{k\beta} (\mathbf{G}_{\mathbf{S}})_{\beta\alpha}. \tag{C.16}$$

Using Eqs. (C.10), (C.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. (C.15), we see that for $N \to \infty$:

$$\frac{1}{T} \sum_{\alpha}^{(k)} R^{(k)}_{\alpha\alpha} = \frac{1}{T} \sum_{\alpha}^{(k)} R_{\alpha\alpha} + O(T^{-1}) = \mathfrak{g}_{\mathbf{S}}(z) + O(T^{-1}), \tag{C.17}$$

and we therefore have from Eqs. (C.13) and (C.11):

$$R_{ij}(z) = \delta_{ij}\left(\frac{\mu_k}{1-\mu_k \mathfrak{g}_S(z)}\right) + O(T^{-1/2}). \tag{C.18}$$

Finally, recalling that $\mathfrak{g}_{\mathbf{S}}(z) = q\mathfrak{g}_{\mathbf{E}}(z) + (1-q)/z$ from Eq. (3.33) and $R_{ii} = z\mu_i G_{ii}$ from Eq. (C.9), we conclude that

$$(\mathbf{G}_{\mathbf{E}}(z))_{ij} = \delta_{ij}\left(\frac{1}{z-\mu_k(1-q+qz\mathfrak{g}_{\mathbf{E}}(z))}\right) + O(T^{-1/2}), \qquad i,j \in [\![1,N]\!], \tag{C.19}$$

which is the prediction obtained in (4.6) with the Replica method. Similarly, we obtain for the $T \times T$ block that:

$$(\mathbf{G}_{\mathbf{S}}(z))_{\alpha\beta} = \frac{\delta_{\alpha\beta}}{z - \frac{1}{T}\sum\limits_{k\in\mathcal{I}_N}(\mathbf{G}_{\mathbf{E}}(z))_{ij}} + O(T^{-1/2}). \tag{C.20}$$

Moreover, by using (C.18) and (3.34), we see that for $N \to \infty$

$$z - \frac{1}{T}\sum_{k\in\mathcal{I}_N}(\mathbf{G}_{\mathbf{E}}(z))_{kk} = \frac{1}{\mathfrak{g}_{\mathbf{S}}(z)}, \tag{C.21}$$

so that we may conclude

$$(\mathbf{G}_{\mathbf{S}}(z))_{\alpha\beta} = \delta_{\alpha\beta}\mathfrak{g}_{\mathbf{S}}(z) + O(T^{-1/2}). \tag{C.22}$$

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 [113].

Appendix D. Additive noise model

In this review, we mainly focus on sample covariance matrices which is a particular case of models of random matrices with multiplicative noise. In this appendix, we consider the case of an additive external noise which can also be important in many situations, in particular in quantum chaos and quantum transport [185], with renewed interest coming from problems of quantum ergodicity ("eigenstate thermalization") [130,186], entanglement and dissipation (for recent reviews see [187,188]). We will show briefly here how we can extend the results of Section 4 to this specific case with a special focus to the overlaps (4.3).

As above, we shall denote the $N \times N$ real symmetric population matrix, i.e. the one we wish to infer, by $\mathbf{C}$ and to avoid confusion, we denote by $\mathbf{M}$ the sample matrix that is the matrix we measure with the data. Throughout this section, we deal with models of the form

$$\mathbf{M} = \mathbf{C} + \boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^*, \tag{D.1}$$

where $\mathbf{B}$ is a fixed matrix with eigenvalues $b_1 > b_2 > \cdots > b_N$, spectral $\rho_{\mathbf{B}}$, and $\boldsymbol{\Omega}$ 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 therefore posit without loss of generality that $\mathbf{C}$ is diagonal. The most common example of such models in the literature [83] is the case where $\mathbf{B}$ belongs to the GOE but for now, we do not specify any distribution or structure assumption on the fixed matrix $\mathbf{B}$. We first present this simple model and then show that we can generalize it to the general case (D.1). We shall also provide an elementary derivation of the free addition in the limit $N \to \infty$.

D.1. Gaussian external noise

In order to give some insights about the general model (D.1), we focus first on the case where the external noise $\mathbf{B}$ belongs to the GOE with a variance of σ^2. More formally, we consider $\mathbf{B}$ to be a $N \times N$ real symmetric matrix with Gaussian entries that satisfies

$$\mathbb{E}[B_{ij}] = 0 \qquad \mathbb{E}[B_{ij}^2] = \begin{cases} 2\sigma^2/N & \text{if } i = j, \\ \sigma^2/N & \text{otherwise.} \end{cases} \tag{D.2}$$

In the case where $\mathbf{B}$ satisfies (D.2), we say that $\mathbf{M}$ defined as (D.1) is a *deformed GOE* matrix. As usual, all the information about the eigenvalues and eigenvectors of $\mathbf{M}$ can be analyzed through the resolvent. In fact, as for sample covariance matrices, it is possible to show that each entry of the resolvent $\mathbf{G}_{\mathbf{M}}$ converges to a deterministic limit for $N \to \infty$. There are a lot of different mathematical methods to prove this last assertion and we shall cover only two of them: the first method is to use a straightforward generalization of the arguments of Appendix C.1 above. The second method is based on the representation of a GOE matrix as a (dynamical) stochastic process, known as *Dyson's Brownian motion*. As we shall see below, this second approach provides insightful physical interpretation about the behavior of $\mathbf{M}$.

D.1.1. Schur complement arguments

Let us start with the first method. We expect the resolvent of $\mathbf{M}$ to be in the same basis than $\mathbf{C}$, at least in the limit $N \to \infty$, meaning that we can work in the basis where $\mathbf{C}$ is diagonal. Moreover, since matrix $\mathbf{C}$ is deterministic, one may easily repeat the arguments of Appendix C.1 to generalize Eq. (C.3) to:

$$D_{kl} = \delta_{kl}\left(z - \mu_k - \frac{\sigma^2}{N}\sum_{\alpha}^{(k)} G^{(k)}_{\alpha\alpha}\right) + \mathcal{O}(N^{-1/2}), \qquad k, l \in \{i, j\}. \tag{D.3}$$

As above, we can consistently check that $G_{ij} \sim N^{-1/2}$ using (B.13). Moreover, we also obtain that $G^{(k)}_{\ell\ell} \sim G_{\ell\ell} + \mathcal{O}(N^{-1})$ for any $\ell \in [\![1, N]\!]$ thanks to (B.18). Therefore, we obtain for any $i \in [\![1, N]\!]$:

$$G_{ii} \sim \frac{1}{z - \sigma^2 \mathfrak{g}(z) - \mu_i} + \mathcal{O}(N^{-1/2}), \tag{D.4}$$

which is the result obtained in e.g. [125,113] using more rigorous arguments.

D.1.2. Dyson Brownian motion

Since the seminal paper of Dyson in 1962 [189], it is well known that the spectrum induced by the addition of free random matrices in the Gaussian orthogonal ensemble[33] can be investigated through the evolution of a time-dependent real symmetric $N \times N$ Brownian motion. More precisely, let us introduce a fictitious time t and rewrite the model (D.1) as :

$$\mathbf{M}(t) = \mathbf{C} + \mathbf{B}(t) \tag{D.5}$$

with

$$B_{ii}(t) = \sqrt{\frac{2\sigma^2}{N}} W_{ii}(t), \qquad B_{ij}(t) = \sqrt{\frac{\sigma^2}{N}} W_{ij}(t) \quad (i \neq j), \tag{D.6}$$

where the $W_{ij}(t)$, $i \leqslant j$ are independent and identically distributed real Brownian motions. We see that $\mathbf{B}(t)$ is an external noise whose variance increases as the time t grows. We suppose that the eigenvalues of $\mathbf{C}$ are all distinct and satisfy $\mu_1 \geqslant \mu_2 \geqslant \dots \mu_N$. Then, the dynamics of the eigenvalues of $\mathbf{M}(t)$ may also be characterized by a stochastic differential equation (SDE), known as *Dyson's Brownian motion*:

$$\begin{aligned} \mathrm{d}\lambda_i(t) &= \sqrt{\frac{2\sigma^2}{N}}\mathrm{d}b_i(t) + \frac{1}{N}\sum_{j\neq i}^{N}\frac{\mathrm{d}t}{\lambda_i(t) - \lambda_j(t)}, \\ \lambda_i(0) &= \mu_i, \end{aligned} \tag{D.7}$$

for any $i = 1, \dots, N$, and where the $b_i(t)$ are independent real Brownian motions. We observe that the eigenvalues of $\mathbf{M}(t)$ defines Dyson's Coulomb gas model that describes positively charged particles on a line interacting via a logarithmic potential and subject to a thermal noise $\mathrm{d}b_i(t)$.

Conditionally to the eigenvalues paths, the trajectories of the associated eigenvectors $\mathbf{u}_i(t)$ can also be characterized by a SDE:

$$\begin{aligned} \mathrm{d}\mathbf{u}_i(t) &= \frac{1}{\sqrt{N}}\sum_{k\neq i}\frac{\mathrm{d}w_{ik}(t)}{\lambda_i(t) - \lambda_k(t)}\mathbf{u}_k(t) - \frac{1}{2N}\sum_{k\neq i}\frac{\mathrm{d}t}{(\lambda_i(t) - \lambda_k(t))^2}\mathbf{u}_i(t), \\ \mathbf{u}_i(0) &= \mathbf{v}_i, \end{aligned} \tag{D.8}$$

where the family of independent (up to symmetry) Brownian motions $\{w_{ij}\}$ is independent from the Brownian motions $\{b_i\}$ that drive the eigenvalues trajectories. As a result, in order to study the dynamics of the eigenvectors, we may always freeze the eigenvalues paths and work conditionally to the realized trajectories. This is the approach used in [125,126,174] in order to study the mean squared overlap (4.3) in this additive model.

In this appendix, we present an alternative approach that considers directly the time evolution of the full resolvent, which we have not seen in the literature before. To that end, we define

$$\mathbf{G}(z, t) := \mathbf{H}^{-1}(z, t), \qquad \mathbf{H}(z, t) := z\mathbf{I}_N - \mathbf{M}(t). \tag{D.9}$$

Using Itô formula and the fact that $\mathrm{d}M_{kl} = \mathrm{d}B_{kl}$, one has

$$\mathrm{d}G_{ij}(z, t) = \sum_{k,l=1}^{N}\frac{\partial G_{ij}}{\partial M_{kl}}\mathrm{d}B_{kl} + \frac{1}{2}\sum_{k,l,m,n=1}^{N}\sum_{m,n=1}^{N}\frac{\partial^2 G_{ij}}{\partial M_{kl}\partial M_{mn}}\mathrm{d}\big[B_{kl}B_{mn}\big]. \tag{D.10}$$

[33] All these results may be easily extended to the Hermitian case.

Next, we compute the derivatives:

$$\frac{\partial G_{ij}}{\partial M_{kl}} = \frac{1}{2}\left[G_{ik}G_{jl} + G_{jk}G_{il}\right], \tag{D.11}$$

from which we deduce the second derivatives

$$\frac{\partial^2 G_{ij}}{\partial M_{kl}\partial M_{mn}} = \frac{1}{4}\left[\left(G_{im}G_{kn} + G_{im}G_{kn}\right)G_{jl} + \cdots\right], \tag{D.12}$$

where we have not written the other 6 GGG products. Now, using (D.6), the quadratic co-variation reads

$$\mathrm{d}\left[B_{kl}B_{mn}\right] = \frac{\sigma^2 \mathrm{d}t}{N}\left(2\delta_{k=l=m=n} + \delta_{k=m}\delta_{l=n} + \delta_{k=n}\delta_{l=m}\right) \tag{D.13}$$

so that we get from (D.10) and taking into account symmetries:

$$\mathrm{d}G_{ij}(z,t) = \sum_{k,l=1}^{N} G_{ik}G_{jl}\mathrm{d}B_{kl} + \frac{\sigma^2}{N}\sum_{k,l=1}^{N}\left(G_{ik}G_{lk}G_{lj} + G_{ik}G_{kj}G_{ll}\right)\mathrm{d}t. \tag{D.14}$$

As above, we expect the entries of $\mathbf{G}$ to be self-averaging. Hence, we consider the average with respect to the Brownian motion W_{kl} defined in Eq. (D.6), we find the following evolution for the average resolvent:

$$\partial_t \mathbb{E}[\mathbf{G}(z,t)] = \sigma^2 \mathfrak{g}(z,t)\,\mathbb{E}[\mathbf{G}^2(z,t)] + \frac{1}{N}\mathbb{E}[\mathbf{G}^3(z,t)]. \tag{D.15}$$

Now, one can notice that:

$$\mathbf{G}^2(z,t) = -\partial_z \mathbf{G}(z,t); \qquad \mathbf{G}^3(z,t) = \partial_{zz}^2 \mathbf{G}(z,t), \tag{D.16}$$

which hold even before averaging. By sending $N \to \infty$, we obtain the following matrix PDE for the resolvent:

$$\partial_t \mathbb{E}[\mathbf{G}(z,t)] = -\sigma^2 \mathfrak{g}(z,t)\,\partial_z \mathbb{E}[\mathbf{G}(z,t)], \quad \text{with} \quad \mathbb{E}[\mathbf{G}(z,0)] = \mathbf{G}_{\mathbf{C}}(z). \tag{D.17}$$

Taking the trace of this equation immediately leads to a Burgers equation for the Stieltjes transform [125,126]:

$$\partial_t \mathfrak{g}(z,t) = -\sigma^2 \mathfrak{g}(z,t)\,\partial_z \mathfrak{g}(z,t), \quad \text{with} \quad \mathfrak{g}(z,0) = \mathfrak{g}_{\mathbf{C}}(z). \tag{D.18}$$

Its solution can be found using the method of characteristics and reads:

$$\mathfrak{g}(z,t) = \mathfrak{g}_{\mathbf{C}}(Z(z,t)), \qquad Z(z,t) := z - \sigma^2 t\,\mathfrak{g}(z,t). \tag{D.19}$$

The solution of Eq. (D.17) then reads [125,190]:

$$\mathbb{E}[\mathbf{G}(z,t)] = \mathbf{G}_{\mathbf{C}}(Z(z,t)), \tag{D.20}$$

and is exactly equivalent to (D.4) except that the variance here is given by $\sigma^2 t$.

Note that we can then easily study from (D.20) the mean squared overlap between the *perturbed* eigenvectors $\mathbf{u}_i(t)$ and the *pure* ones $\mathbf{u}_j(0) = \mathbf{v}_j$ for any $i, j \in [\![1, N]\!]$. Indeed, it suffices to consider in the basis where $\mathbf{C}$ is diagonal the following projection $\langle \mathbf{v}_j, G_{ii}(z,t)\mathbf{v}_j\rangle$ with $z = \lambda_i - \mathrm{i}\eta$ as in Section 4 and we finally obtain

$$N\mathbb{E}\left[\langle \mathbf{u}_i(t), \mathbf{v}_j\rangle^2\right] = \frac{\sigma^2 t}{\left|\lambda_i(t) - \sigma^2 t\,\mathfrak{g}_{\mathbf{M}}(z,t) - \mu_j\right|^2}. \tag{D.21}$$

D.2. Extension to an arbitrary rotational invariant noise

D.2.1. An elementary derivation of the free addition formula

We now turn on the general case where the noise term $\mathbf{B}$ is a (asymptotically) rotational invariant random matrix. We saw in Section 2.3 that the limiting spectrum of such models can be investigated using the free probability formalism. The first part of this section is dedicated to a formal but elementary derivation of Voiculescu's free addition (2.66) [45] by following the arguments of [38]. From this result, we will be able to derive the asymptotic behavior of the resolvent of the model (D.1) using the Replica formalism of Section 2.4.

As in Section 2.3.3, 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 (2.92) for the resolvent of (D.1) which

yields[34]

$$\mathbf{G}_{\mathbf{M}}(z)_{i,j} = \int \left(\prod_{\alpha=1}^{n}\prod_{k=1}^{N} d\eta_k^\alpha\right) \eta_i^1 \eta_j^1 \prod_{\alpha=1}^{n} e^{-\frac{1}{2}\sum_{k=1}^{N}(\eta_k^\alpha)^2(z-c_k)} \left\langle e^{-\frac{1}{2}\sum_{k,l=1}^{N}\eta_k^\alpha(\boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^*)_{k,l}\eta_l^\alpha}\right\rangle_{\boldsymbol{\Omega}}. \tag{D.22}$$

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 Appendix A.2. Hence, one deduces from (A.5) that

$$\mathcal{I}_1\left(\sum_{\alpha=1}^{n}\eta^\alpha(\eta^\alpha)^*, \mathbf{B}\right) = \exp\left[\frac{N}{2}\sum_{\alpha=1}^{n}\mathcal{W}_{\mathbf{B}}\left(\frac{1}{N}(\eta^\alpha)^\dagger\eta^\alpha\right)\right], \tag{D.23}$$

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 (D.22) becomes

$$\mathbf{G}_{\mathbf{M}}(z)_{i,j} = \int \left(\prod_{k=1}^{N} d\eta_k\right) \eta_i^1\eta_j^1 \exp\left\{\frac{N}{2}\sum_{\alpha=1}^{n}\left[\mathcal{W}_{\mathbf{B}}\left(\frac{1}{N}(\eta^\alpha)^\dagger\eta^\alpha\right) - \frac{1}{2}\sum_{k=1}^{N}(\eta_k^\alpha)^2(z-\mu_k)\right]\right\}, \tag{D.24}$$

and by introducing a Lagrange multiplier $p^\alpha := \frac{1}{N}(\eta^\alpha)^\dagger\eta^\alpha$, we obtain using Fourier transform (and renaming $\zeta^\alpha \to -2i\zeta^\alpha/N$)

$$\begin{aligned}\mathbf{G}_{\mathbf{M}}(z)_{i,j} \propto{}& \int\int \left(\prod_{\alpha=1}^{n} dp^\alpha d\zeta^\alpha\right) \exp\left\{\frac{N}{2}\sum_{\alpha=1}^{n}[\mathcal{W}_{\mathbf{B}}(p^\alpha) - p^\alpha\zeta^\alpha]\right\} \\ &\times \int \left(\prod_{\alpha=1}^{n}\prod_{k=1}^{N} d\eta_k^\alpha\right) \eta_i^1\eta_j^1 \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{k=1}^{N}(\eta_k^\alpha)^2(z-\zeta^\alpha-\mu_k)\right\}.\end{aligned}$$

One can readily find

$$\mathbf{G}_{\mathbf{M}}(z)_{i,j} \propto \int\int \left(\prod_{\alpha=1}^{n} dp^\alpha d\zeta^\alpha\right) \frac{\delta_{ij}}{z+\zeta^1-\mu_i} \exp\left\{-\frac{Nn}{2}F_0(p^\alpha,\zeta^\alpha)\right\}, \tag{D.25}$$

where the 'free energy' F_0 is given by

$$F_0(p^\alpha,\zeta^\alpha) = \frac{1}{Nn}\sum_{\alpha=1}^{n}\left[\sum_{k=1}^{N}\log(z-\zeta^\alpha-\mu_k) - \mathcal{W}_{\mathbf{B}}(p^\alpha) + p^\alpha\zeta^\alpha\right]. \tag{D.26}$$

As in Section 2.3.3, the integral (D.25) 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

$$p^\alpha = p \quad \text{and} \quad \zeta^\alpha = \zeta, \qquad \forall\,\alpha \in \{1,\dots,n\}, \tag{D.27}$$

from which, we obtain the following set of equations:

$$\zeta^* = \mathcal{R}_{\mathbf{B}}(p^*) \qquad \text{and} \qquad p^* = \mathfrak{g}_{\mathbf{C}}(z-\zeta^*). \tag{D.28}$$

If we apply the Blue transform of $\mathbf{C}$ on the second equation of (D.28), we obtain

$$z = \mathcal{B}_{\mathbf{C}}(p^*) + \mathcal{R}_{\mathbf{B}}(p^*) \equiv \mathcal{R}_{\mathbf{C}}(p^*) + \mathcal{R}_{\mathbf{B}}(p^*) - \frac{1}{p^*}. \tag{D.29}$$

On the other hand, we see that the resolvent (D.25) is given in the large N limit and the limit $n \to 0$ by

$$\mathbf{G}_{ij}(z) \sim \frac{\delta_{ij}}{z - \mathcal{R}_{\mathbf{B}}(p^*) - \mu_i}. \tag{D.30}$$

The trick is to see that we can get rid of one variable by taking the normalized trace in this later equation as it yields

$$\mathfrak{g}_{\mathbf{M}}(z) = \mathfrak{g}_{\mathbf{C}}(z - \mathcal{R}_{\mathbf{B}}(p^*)) = p^* \tag{D.31}$$

where the last equation follows from (D.28). Therefore, we conclude by plugging this last equation into (D.29) that

$$z - \frac{1}{\mathfrak{g}_{\mathbf{M}}(z)} = \mathcal{R}_{\mathbf{C}}(\mathfrak{g}_{\mathbf{M}}(z)) + \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z)),$$

[34] One may also use the Replica formalism for the Stieltjes transform as well.

from which one can check by renaming $z = \mathcal{B}_{\mathbf{M}}(\omega)$ that

$$\mathcal{R}_{\mathbf{M}}(\omega) = \mathcal{R}_{\mathbf{C}}(\omega) + \mathcal{R}_{\mathbf{B}}(\omega), \tag{D.32}$$

which is exactly the free addition formula (2.66).

D.2.2. Asymptotic resolvent of (D.1)

A trivial application of the result above is the evaluation of the resolvent entry-wise for the general model (D.1). Indeed, we see by plugging Eq. (D.31) into Eq. (D.30) that

$$\mathbf{G}_{\mathbf{M}}(z)_{ij} \sim \frac{\delta_{ij}}{z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z)) - \mu_i}, \tag{D.33}$$

which is equivalent to

$$\mathbf{G}_{\mathbf{M}}(z)_{ij} = \mathbf{G}_{\mathbf{C}}(Z(z))_{ij}, \qquad Z(z) \coloneqq z - \mathcal{R}_{\mathbf{B}}(\mathfrak{g}_{\mathbf{M}}(z)). \tag{D.34}$$

One notices that this formula is indeed the generalization of the formula (2.67) as a matrix. Moreover, we see that in the large N limit, each entry of the random resolvent of $\mathbf{M}$ converges to a deterministic quantity that lies in the basis of $\mathbf{C}$. We moreover 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 4 can be performed nearly verbatim for the additive model (D.1) and the exact results can be found in [38].

D.3. Overlap and optimal RIE formulas in the additive case

D.3.1. Mean squared overlaps

We were able to show that each entries of the resolvent of $\mathbf{M}$ in the general additive model (D.1) converges to a deterministic limit that is given in Eq. (D.34). 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 4.1.1. We omit details that may be found in [38] and one finds that the overlap for the free additive noise is given by:

$$\Phi(\lambda, \mu) = \frac{\beta_1(\lambda)}{(\lambda - c - \alpha_a(\lambda))^2 + \pi^2 \beta_a(\lambda)^2 \rho_{\mathbf{M}}(\lambda)^2}, \tag{D.35}$$

where μ is the corresponding eigenvalue of the true matrix $\mathbf{C}$, and where we defined:

$$\begin{cases} \alpha_a(\lambda) \coloneqq \mathrm{Re}[\mathcal{R}_{\mathbf{B}}(\mathfrak{h}_{\mathbf{M}}(\lambda) + \mathrm{i}\pi\rho_{\mathbf{M}}(\lambda))], \\ \beta_a(\lambda) \coloneqq \dfrac{\mathrm{Im}[\mathcal{R}_{\mathbf{B}}(\mathfrak{h}_{\mathbf{M}}(\lambda) + \mathrm{i}\pi\rho_{\mathbf{M}}(\lambda))]}{\pi\rho_{\mathbf{M}}(\lambda)}. \end{cases} \tag{D.36}$$

As a simple consistency check, we specialize our result to the case where $\boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^*$ is a GOE matrix such that the entries have a variance equal to σ^2/N. Then, one has $\mathcal{R}_{\mathbf{B}}(z) = \sigma^2 z$ meaning that $Z(z)$ of Eq. (D.34) simply becomes $Z(z) = z - \sigma^2 \mathfrak{g}_{\mathbf{M}}(z)$. This allows us to get a simpler expression for the overlap:

$$\Phi(\lambda, \mu) = \frac{\sigma^2}{(c - \lambda + \sigma^2 \mathfrak{h}_{\mathbf{M}}(\lambda))^2 + \sigma^4 \pi^2 \rho_{\mathbf{M}}(\lambda)^2}, \tag{D.37}$$

which is exactly the result obtained in Eq. (D.21). In Fig. D.1, we illustrate this formula in the case where $\mathbf{C} = \mathcal{W}$ with parameter q. We set $N = 500$, $T = 1000$, and take $\boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^*$ as a GOE matrix with variance $1/N$. For a fixed $\mathbf{C}$, we generate 200 samples of $\mathbf{M}$ given by Eq. (D.1) for which we can measure numerically the overlap (4.3). We see that the theoretical prediction (D.37) agrees remarkably with the numerical simulations.

D.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 (6.2) for the bulk eigenvalues in the model (D.1). Indeed, it is easy to see from Eqs. (2.6) and (6.2) that:

$$\xi_i^{\mathrm{ora.}} \sim \frac{1}{\pi\rho_{\mathbf{M}}(\lambda_i)} \lim_{z \to \lambda_i - \mathrm{i}0^+} \mathrm{Im}\left[\int \frac{\mu\,\rho_{\mathbf{C}}(\mu)}{Z(z) - \mu}\,\mathrm{d}\mu\right] = \frac{1}{N\pi\rho_{\mathbf{M}}(\lambda_i)} \lim_{z \to \lambda_i - \mathrm{i}0^+} \mathrm{Im}\,\mathrm{Tr}\left[\mathbf{G}_{\mathbf{M}}(z)\mathbf{C}\right], \tag{D.38}$$

where $Z(z)$ is given by Eq. (D.34). From Eq. (D.34) one also has $\mathrm{Tr}[\mathbf{G}_{\mathbf{M}}(z)\mathbf{C}] = N(Z(z)\mathfrak{g}_{\mathbf{M}}(z) - 1)$, and using Eqs. (D.34) and (D.36), we end up with:

$$\lim_{z \to \lambda - \mathrm{i}0^+} \mathrm{Im}\,\mathrm{Tr}\left[\mathbf{G}_{\mathbf{M}}(z)\mathbf{C}\right] = N\pi\rho_M(\lambda)\left[\lambda - \alpha(\lambda) - \beta(\lambda)\mathfrak{h}_{\mathbf{M}}(\lambda)\right].$$

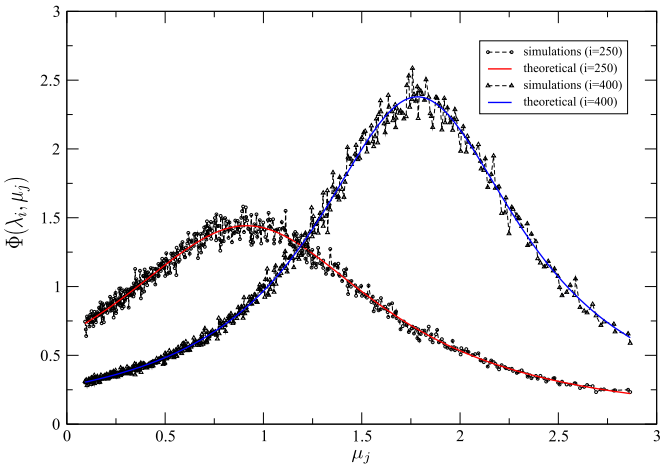


Fig. D.1. Computations of the rescaled overlap $\Phi(\lambda, \mu)$ as a function of μ in the free addition perturbation. We chose $i = 250$, $\mathbf{C}$ a Wishart matrix with parameter $q = 0.5$ and $\mathbf{B}$ a Wigner matrix with $\sigma^2 = 1$. The black circle points are computed using numerical simulations and the plain red curve is the theoretical predictions Eq. (D.35). 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. (D.1) with $\mathbf{C}$ fixed.

We therefore find the following optimal RIE nonlinear "shrinkage" function F_a:

$$\xi_i^{\text{ora.}} \sim F_a(\lambda_i); \qquad F_a(\lambda) = \lambda - \alpha_a(\lambda) - \beta_a(\lambda)\mathfrak{h}_{\mathbf{M}}(\lambda), \tag{D.39}$$

where α_a, β_a are defined in Eq. (D.36). This result states that if we consider a model where the signal $\mathbf{C}$ is perturbed by 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 (D.39). 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 (D.39). Using the definition of α_a and β_a given in Eq. (D.36), the nonlinear shrinkage function is given by

$$F_a(\lambda) = \lambda - 2\sigma^2\mathfrak{h}_{\mathbf{M}}(\lambda). \tag{D.40}$$

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:

$$F_a(\lambda) = \lambda\left(\frac{\sigma_{\mathbf{C}}^2}{\sigma_{\mathbf{C}}^2 + \sigma^2}\right), \tag{D.41}$$

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 [135], and this is exactly the result that we have obtained with (D.41). 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 (D.41) that $\widehat{\Xi}(\mathbf{M})$ is a Wigner matrix with variance $\sigma_{\Xi}^2 \times \text{SNR}$ which leads to

$$\sigma_{\mathbf{M}}^2 \geq \sigma_{\mathbf{C}}^2 \geq \sigma_{\mathbf{C}}^2 \times \text{SNR}, \tag{D.42}$$

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 non-negative eigenvalues. However, we expect that the noisy effect coming from the GOE matrix pushes some true eigenvalues towards the negative side of the real axis. In Fig. D.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. (D.40), we invoke once again the free addition formula to find the following equation for the Stieltjes transform of $\mathbf{M}$:

$$-q_0\sigma^2\mathfrak{g}_{\mathbf{M}}(z)^3 + (\sigma^2 + q_0 z)\mathfrak{g}_{\mathbf{M}}(z)^2 + (1 - q_0 - z)\mathfrak{g}_{\mathbf{M}}(z) + 1 = 0,$$

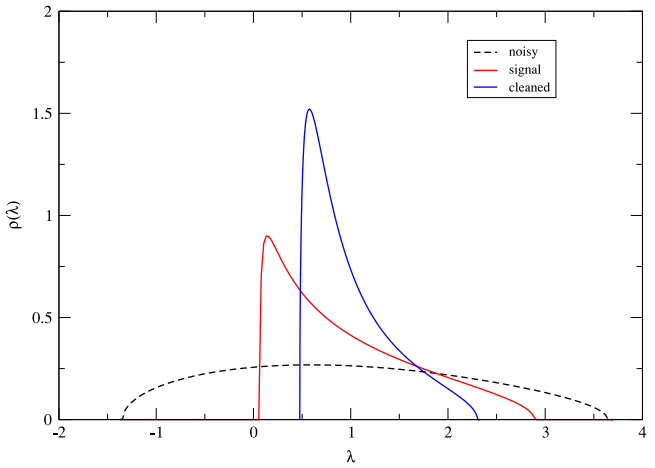


Fig. D.2. Eigenvalues of the noisy measurement **M** (black dotted line) compared to the true signal **C** drawn from a 500×500 Wishart matrix of parameter $q_0 = 0.5$ (red line). We have corrupted the signal by adding a GOE matrix with radius 1. The eigenvalues density of **M** allows negative values 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

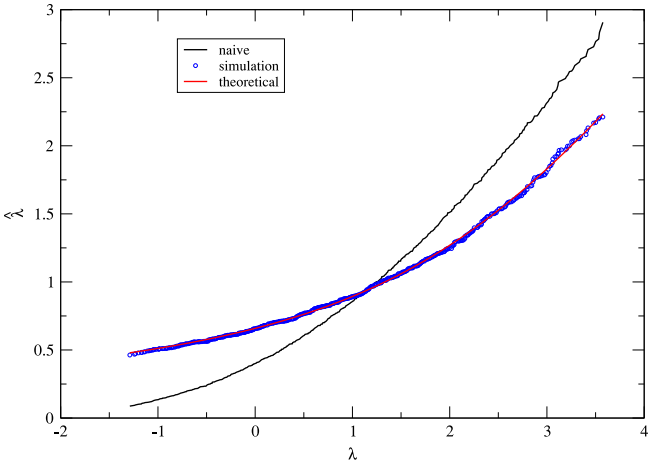


Fig. D.3. Eigenvalues according to the optimal cleaning formula (D.41) (red line) as a function of the observed noisy eigenvalues λ. The parameter are the same as in Fig. D.2. We also provide a comparison against the naive eigenvalues substitution method (black line) and we see that the optimal cleaning scheme indeed narrows the spacing between eigenvalues. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for any $z = \lambda - i\eta$ with $\eta \to 0$. It then suffices to take the real part of the Stieltjes transform $g_{\mathbf{M}}(z)$ that solves this equation[35] to get the Hilbert transform. In order to check formula Eq. (D.39) using numerical simulations, we have generated a matrix of **M** given by Eq. (D.1) with **C** a fixed white Wishart matrix with parameter q_0 and $\boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}^*$ a GOE matrix with radius 1. As we know exactly **C**, we can compute numerically the oracle estimator as given in (6.2) for each sample. In Fig. D.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. D.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 Section 6.

Appendix E. 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 $a, b \in \mathbb{N}$.

35 We take the solution which has a strictly non-negative imaginary part.

Mathematical symbols

We list here some of the most important notations of the review.

Symbol	Description
$\mathcal{B}_{\mathbf{M}}$	Blue transform of $\mathbf{M}$ (2.15)
$\mathbf{C}$	Population/True covariance matrix (3.1)
$\underline{\mathbf{C}}$	Spikeless version of $\mathbf{C}$ (3.56)
$\mathbb{C}_{\pm}$	Complex upper/lower half plane
$\mathbf{E}$	Sample/Empirical covariance matrix (3.3)
$\mathbb{E}$	Expectation value over the noise
$\mathbf{G}_{\mathbf{M}}$	Resolvent of $\mathbf{M}$, (2.5)
$\mathfrak{g}^N_{\mathbf{M}}$	Empirical Stieltjes transform of $\rho_{\mathbf{M}}$ (2.7)
$\mathfrak{g}_{\mathbf{M}}$	Stieltjes transform of $\rho_{\mathbf{M}}$ (2.8)
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 $\mathbf{M}$ (2.16)
$\mathcal{R}^2_{\text{in}}$	In-sample/predicted risk (7.7)
$\mathcal{R}^2_{\text{out}}$	Out-of-sample/realized risk (7.9)
$\mathcal{R}^2_{\text{true}}$	True risk (7.5)
$\mathbf{S}$	"Dual" sample covariance matrix (3.32)
$\mathcal{S}_{\mathbf{M}}$	S-transform of $\mathbf{M}$ (2.23)
T	Sample size
$\mathcal{T}_{\mathbf{M}}$	T-transform of $\mathbf{M}$ (2.21)
$\mathbf{u}_i$	Sample eigenvector associated to λ_i
$\mathbf{v}_i$	Population eigenvector associated to μ_i
$\mathbb{V}$	Variance of a random variable
$\mathcal{W}_{\mathbf{M}}$	Primitive of the $\mathcal{R}$-Transform of $\mathbf{M}$ (2.96)
$\mathbf{Y}$	$N \times T$ normalized data matrix
α_s	Linear shrinkage intensity (5.19)
λ_i	ith sample eigenvalue
μ_i	ith population (true) eigenvalue
$\Xi^{\text{lin.}}$	Linear Shrinkage estimator (5.19)
$\hat{\Xi}(\mathbf{E})$	Optimal RIE of $\mathbf{C}$ depending on $\mathbf{E}$
$\Xi^{\text{ora.}}$	Oracle estimator (6.2)
$\Xi(\mathbf{E})$	RIE of $\mathbf{C}$ depending on $\mathbf{E}$
$\rho^N_{\mathbf{M}}$	Empirical spectral density of $\mathbf{M}$ (2.3)
$\rho_{\mathbf{M}}$	Limiting spectral density of $\mathbf{M}$ (2.4)
Φ	Rescaled mean squared overlap (4.3) and (4.4)
$\varphi(\mathbf{M})$	Normalized trace of $\mathbf{M}$ (2.61)
$\boldsymbol{\Omega}$	Rotation matrix
$\langle\cdot\rangle_{\mathbf{M}}$	Expectation value with respect to $\mathcal{P}(\mathbf{M})$
$\langle\ ,\ \rangle$	inner product

Abbreviations

Symbol	Description
CCA	Canonical Correlation Analysis
ESD	Empirical Spectral Density
GOE	Gaussian Orthogonal Ensemble
HCIZ	Harish-Chandra–Itzykson–Zuber

(*continued on next page*)

Symbol	Description
IW	Inverse Wishart
IWs	Inverse Wishart + sorting
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
RI	Rotational Invariance
RIE	Rotational Invariant Estimator
RP	Relative Performance
RMT	Random Matrix Theory
SVD	Singular Value Decomposition

References

[1] A.W. Van der Vaart, Asymptotic Statistics, vol. 3, Cambridge university press, 2000.
[2] T. Amemiya, Advanced Econometrics, Harvard University Press, 1985.
[3] L.P. Hansen, Large sample properties of generalized method of moments estimators, Econometrica (1982) 1029–1054.
[4] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, in: Springer Series in Statistics, vol. 1, Springer, Berlin, 2001.
[5] H. Markowitz, Portfolio selection, J. Finance 7 (1) (1952) 77–91.
[6] F. Caccioli, I. Kondor, G. Papp, Portfolio optimization under expected shortfall: contour maps of estimation error, arXiv preprint arXiv:1510.04943.
[7] S. Ciliberti, I. Kondor, M. Mézard, On the feasibility of portfolio optimization under expected shortfall, Quant. Finance 7 (4) (2007) 389–396.
[8] M. Pourahmadi, High-Dimensional Covariance Estimation: With High-Dimensional Data, John Wiley & Sons, 2013.
[9] J. Wishart, The generalised product moment distribution in samples from a normal multivariate population, Biometrika (1928) 32–52.
[10] T.W. Anderson, An Introduction to Multivariate Statistics, Wiley, New York, 1984, p. 675.
[11] C. Stein, Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1956, pp. 197–206.
[12] B. Efron, C.N. Morris, Stein's Paradox in Statistics, WH Freeman, 1977.
[13] W. James, C. Stein, Estimation with quadratic loss, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1961, pp. 361–379.
[14] B. Efron, C. Morris, Multivariate empirical Bayes and estimation of covariance matrices, Ann. Statist. (1976) 22–32.
[15] L. Haff, Minimax estimators for a multinormal precision matrix, J. Multivariate Anal. 7 (3) (1977) 374–385.
[16] L. Haff, Empirical Bayes estimation of the multivariate normal covariance matrix, Ann. Statist. (1980) 586–597.
[17] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2) (2004) 365–411.
[18] V.A. Marchenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Mat. Sb. 114 (4) (1967) 507–536.
[19] T.W. Anderson, Asymptotic theory for principal component analysis, Ann. Math. Stat. (1963) 122–148.
[20] Y.Q. Yin, Limiting spectral distribution for a class of random matrices, J. Multivariate Anal. 20 (1) (1986) 50–68.
[21] J.W. Silverstein, Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices, J. Multivariate Anal. 55 (2) (1995) 331–339.
[22] A. Sengupta, P.P. Mitra, Distributions of singular values for some random matrices, Phys. Rev. E 60 (3) (1999) 3389.
[23] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, Noise dressing of financial correlation matrices, Phys. Rev. Lett. 83 (7) (1999) 1467.
[24] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, T. Guhr, H.E. Stanley, Random matrix approach to cross correlations in financial data, Phys. Rev. E 65 (6) (2002) 066126.
[25] J.-P. Bouchaud, M. Potters, Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management, Cambridge University Press, 2003.
[26] I.M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Statist. (2001) 295–327.
[27] C.A. Tracy, H. Widom, Level-spacing distributions and the Airy kernel, Comm. Math. Phys. 159 (1) (1994) 151–174.
[28] L. Laloux, P. Cizeau, M. Potters, J.-P. Bouchaud, Random matrix theory and financial correlations, Int. J. Theor. Appl. Finance 3 (03) (2000) 391–397.
[29] J.-P. Bouchaud, M. Potters, Financial applications of random matrix theory: a short review, in: The Oxford Handbook of Random Matrix Theory, Oxford University Press, 2011, pp. 824–850.
[30] J.W. Silverstein, S.-I. Choi, Analysis of the limiting spectral distribution of large dimensional random matrices, J. Multivariate Anal. 54 (2) (1995) 295–309.
[31] X. Mestre, Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates, IEEE Trans. Inform. Theory 54 (11) (2008) 5113–5129.
[32] J. Yao, A. Kammoun, J. Najim, Eigenvalue estimation of parameterized covariance matrices of large dimensional data, IEEE Trans. Signal Process. 60 (11) (2012) 5893–5905.
[33] N. El Karoui, Spectrum estimation for large dimensional covariance matrices using random matrix theory, Ann. Statist. (2008) 2757–2790.
[34] J.W. Silverstein, Eigenvalues and eigenvectors of large dimensional sample covariance matrices, Contemp. Math. 50 (1986) 153–159.
[35] J.W. Silverstein, On the eigenvectors of large dimensional sample covariance matrices, J. Multivariate Anal. 30 (1) (1989) 1–16.
[36] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statist. Sinica (2007) 1617–1642.
[37] O. Ledoit, S. Péché, Eigenvectors of some large sample covariance matrix ensembles, Probab. Theory Related Fields 151 (1–2) (2011) 233–264.
[38] J. Bun, R. Allez, J.P. Bouchaud, M. Potters, Rotational invariant estimator for general noisy matrices, IEEE Trans. Inform. Theory (ISSN: 0018-9448) 62 (12) (2016) 7475–7490. http://dx.doi.org/10.1109/TIT.2016.2616132.
[39] J. Bun, A. Knowles, An optimal rotational invariant estimator for general covariance matrices, in preparation (2017).

[40] F. Benaych-Georges, R.R. Nadakuditi, The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices, Adv. Math. 227 (1) (2011) 494–521.
[41] R. Monasson, D. Villamaina, Estimating the principal components of correlation matrices from all their empirical eigenvectors, Europhys. Lett. 112 (5) (2015) 50001.
[42] E. Brézin, C. Itzykson, G. Parisi, J.-B. Zuber, Planar diagrams, Comm. Math. Phys. 59 (1) (1978) 35–51.
[43] E.P. Wigner, On the statistical distribution of the widths and spacings of nuclear resonance levels, in: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 47, Cambridge Univ. Press, 1951, pp. 790–798.
[44] D. Voiculescu, Symmetries of Some Reduced Free Product C*-Algebras, Springer, 1985.
[45] D. Voiculescu, Limit laws for random matrices and free products, Invent. Math. 104 (1) (1991) 201–220.
[46] S. Edwards, R.C. Jones, The eigenvalue spectrum of a large symmetric random matrix, J. Phys. A: Math. Gen. 9 (10) (1976) 1595.
[47] M. Mézard, M.A. Virasoro, G. Parisi, Spin Glass Theory and Beyond, World Scientific, 1987.
[48] G. Parisi, Replica theory and spin glasses, in: Statistical Physics, Optimization, Inference, and Message-Passing Algorithms: Lecture Notes of the Les Houches School of Physics: Special Issue, October 2013, Florent Krzakala, Federico Ricci-Tersenghi, Lenka Zdeborova, Riccardo Zecchina, Eric W. Tramel, and Leticia F. Cugliandolo, Oxford Scholarship Online, 2015.
[49] H. Weidenmüller, G. Mitchell, Random matrices and chaos in nuclear physics: Nuclear structure, Rev. Modern Phys. 81 (2) (2009) 539.
[50] G. Akemann, J. Baik, P. Di Francesco, The Oxford Gandbook of Random Matrix Theory, Oxford University Press, 2011.
[51] D.S. Dean, S.N. Majumdar, Large deviations of extreme eigenvalues of random matrices, Phys. Rev. Lett. 97 (16) (2006) 160201.
[52] S. Majumdar, Random matrices, the Ulam problem, Directed polymers and Growth models, and Sequence matching, in: Les Houches-Session LXXXV, Elsevier, 2006, pp. 179–216. (Chapter 4).
[53] G.W. Anderson, A. Guionnet, O. Zeitouni, An Introduction to Random Matrices, Cambridge University Press, 2010.
[54] T. Tao, Topics in Random Matrix Theory, vol. 132, American Mathematical Soc., 2012.
[55] A.M. Tulino, S. Verdú, Random matrix theory and wireless communications, Commun. Inform. Theory 1 (1) (2004) 1–182.
[56] Z. Bai, J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, Springer, 2009.
[57] R. Couillet, M. Debbah, Random Matrix Methods for Wireless Communications, Cambridge University Press, Cambridge, MA, 2011.
[58] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.
[59] K.W. Wachter, The limiting empirical measure of multiple discriminant ratios, Ann. Statist. (1980) 937–957.
[60] J.-P. Bouchaud, L. Laloux, M.A. Miceli, M. Potters, Large dimension forecasting models and random singular value spectra, Eur. Phys. J. B 55 (2) (2007) 201–207.
[61] T. Tao, V. Vu, Random matrices: universality of local eigenvalue statistics, Acta Math. 206 (1) (2011) 127–204.
[62] B. Charles, D. Chafai, Around the circular law, Probab. Surv. 9 (2012) 1–89.
[63] D. Voiculescu, K. Dykema, A. Nica, Free Random Variables, American Mathematical Soc., 1992.
[64] R. Speicher, Free probability theory, in: The Oxford Handbook of Random Matrix Theory, Oxford University Press, 2011, pp. 452–470.
[65] A. Zee, Law of addition in random matrix theory, Nuclear Phys. B 474 (3) (1996) 726–744.
[66] Z. Burda, Free products of large random matrices–a short review of recent developments, J. Phys.: Conf. Ser. 473 (2013) 012002.
[67] Z. Burda, A. Görlich, A. Jarosz, J. Jurkiewicz, Signal and noise in correlation matrix, Physica A 343 (2004) 295–310.
[68] T. Guhr, Supersymmetry, in: The Oxford Handbook of Random Matrix Theory, Oxford University Press, 2011, pp. 135–154.
[69] M.L. Mehta, Random Matrices, vol. 142, Academic Press, 2004.
[70] D.S. Dean, S.N. Majumdar, Extreme value statistics of eigenvalues of Gaussian random matrices, Phys. Rev. E 77 (4) (2008) 041108.
[71] J.-B. Zuber, Introduction to random matrices, 2012.
[72] R. Allez, J.-P. Bouchaud, S.N. Majumdar, P. Vivo, Invariant β-Wishart ensembles, crossover densities and asymptotic corrections to the Marčenko–Pastur law, J. Phys. A 46 (1) (2012) 015001.
[73] P. Cizeau, J.-P. Bouchaud, Theory of Lévy matrices, Phys. Rev. E 50 (3) (1994) 1810.
[74] G. Ben Arous, A. Guionnet, The spectrum of heavy tailed random matrices, Comm. Math. Phys. 278 (3) (2008) 715–751.
[75] E. Tarquini, G. Biroli, M. Tarzia, Level statistics and localization transitions of Lévy matrices, Phys. Rev. Lett. 116 (1) (2016) 010601.
[76] Z.D. Bai, Convergence rate of expected spectral distributions of large random matrices. Part I. Wigner matrices, Ann. Probab. (1993) 625–648.
[77] Z. Bai, B. Miao, J.-F. Yao, Convergence rates of spectral distributions of large sample covariance matrices, SIAM J. Matrix Anal. Appl. 25 (1) (2003) 105–127.
[78] P.S. Dwyer, Some applications of matrix derivatives in multivariate analysis, J. Amer. Statist. Assoc. 62 (318) (1967) 607–625.
[79] L. Haff, An identity for the Wishart distribution with applications, J. Multivariate Anal. 9 (4) (1979) 531–544.
[80] R. Speicher, Multiplicative functions on the lattice of non-crossing partitions and free convolution, Math. Ann. 298 (1) (1994) 611–628.
[81] G. Hooft, A planar diagram theory for strong interactions, Nuclear Phys. B 72 (3) (1974) 461–473.
[82] A.M. Khorunzhy, L. Pastur, On the eigenvalue distribution of the deformed Wigner ensemble of random matrices, Adv. Sov. Math. 19 (1994) 97–127.
[83] E. Brézin, S. Hikami, A. Zee, Universal correlations for deterministic plus random Hamiltonians, Phys. Rev. E 51 (6) (1995) 5442.
[84] P. Zinn-Justin, Adding and multiplying random matrices: a generalization of voiculescus formulas, Phys. Rev. E 59 (5) (1999) 4884.
[85] Z. Burda, R. Janik, M. Nowak, Multiplication law and S transform for non-hermitian random matrices, Phys. Rev. E 84 (6) (2011) 061125.
[86] Z. Burda, J. Jurkiewicz, B. Wacław, Spectral moments of correlated Wishart matrices, Phys. Rev. E 71 (2) (2005) 026111.
[87] G. Parisi, A sequence of approximated solutions to the SK model for spin glasses, J. Phys. A: Math. Gen. 13 (4) (1980) L115.
[88] Harish-Chandra, Differential operators on a semisimple Lie algebra, Amer. J. Math. (1957) 87–120.
[89] C. Itzykson, J.-B. Zuber, The planar approximation. II, J. Math. Phys. 21 (1980) 411–421.
[90] M. Talagrand, The parisi formula, Ann. of Math. 163 (2006) 221–263.
[91] F. Benaych-Georges, A. Knowles, Lectures on the local semicircle law for Wigner matrices, arXiv preprint arXiv:1601.04055.
[92] V. Kargin, Subordination for the sum of two random matrices, Ann. Probab. 43 (4) (2015) 2119–2150.
[93] D. Paul, A. Aue, Random matrix theory in statistics: a review, J. Statist. Plann. Inference 150 (2014) 1–29.
[94] J. Yao, Z. Bai, S. Zheng, Large Sample Covariance Matrices and High-Dimensional Data Analysis, 39, Cambridge University Press, 2015.
[95] P. Vivo, S.N. Majumdar, O. Bohigas, Large deviations of the maximum eigenvalue in Wishart random matrices, J. Phys. A 40 (16) (2007) 4317.
[96] S.N. Majumdar, P. Vivo, Number of relevant directions in principal component analysis and wishart random matrices, Phys. Rev. Lett. 108 (20) (2012) 200601.
[97] A. Perret, G. Schehr, Finite N corrections to the limiting distribution of the smallest eigenvalue of Wishart complex matrices, Random Matrices (2015) 1650001.
[98] T. Wirtz, T. Guhr, Distribution of the smallest eigenvalue in the correlated Wishart model, Phys. Rev. Lett. 111 (9) (2013) 094101.
[99] A. Bloemendal, A. Knowles, H.T. Yau, et al., Probab. Theory Relat. Fields 164 (2016) 459. http://dx.doi.org/10.1007/s00440-015-0616-x.
[100] P.J. Huber, Robust Statistics, Springer, 2011.
[101] R.A. Maronna, D.R. Martin, V.J. Yohai, Robust Statistics: Theory and Methods, John Wiley and Sons, 2006.
[102] G. Biroli, J.P. Bouchaud, M. Potters, The student ensemble of correlation matrices: eigenvalue spectrum and Kullback–Leibler entropy, Acta Phys. Polon. Ser. B 39 (1) (2008) 4009–4026.
[103] N. El Karoui, Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond, Ann. Appl. Probab. 19 (6) (2009) 2362–2405.
[104] R. Chicheportiche, Non-linear dependences in finance, arXiv preprint arXiv:1309.5073, and references therein.
[105] R. Couillet, F. Pascal, J.W. Silverstein, The random matrix regime of Maronnas M-estimator with elliptically distributed samples, J. Multivariate Anal. 139 (2015) 56–78.
[106] D.E. Tyler, A distribution-free M-estimator of multivariate scatter, Ann. Statist. 15 (1) (1987) 234–251.

[107] T. Zhang, X. Cheng, A. Singer, Marchenko–Pastur Law for Tyler's and Maronna's M-estimators, arXiv preprint arXiv:1401.3424.
[108] R. Couillet, A. Kammoun, F. Pascal, Second order statistics of robust estimators of scatter. application to glrt detection for elliptical signals, J. Multivariate Anal. 143 (2016) 249–274.
[109] J.W. Silverstein, Z. Bai, On the empirical distribution of eigenvalues of a class of large dimensional random matrices, J. Multivariate Anal. 54 (2) (1995) 175–192.
[110] J.A. Mingo, A. Nica, Annular noncrossing permutations and partitions, and second-order asymptotics for random matrices, Int. Math. Res. Not. IMRN 2004 (28) (2004) 1413–1460.
[111] O. Ledoit, M. Wolf, Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions, J. Multivariate Anal. 139 (2015) 360–384.
[112] J. Bun, J.-P. Bouchaud, M. Potters, A Dyson Brownian motion for the resolvant of Wigner and Wishart random matrices, in preparation (2017).
[113] A. Knowles, J. Yin, Anisotropic local laws for random matrices, Probab. Theory Relat. Fields (ISSN: 1432-2064) (2016) 1–96. http://dx.doi.org/10.1007/s00440-016-0730-4.
[114] E. Dobriban, Efficient computation of limit spectra of sample covariance matrices, Random Matrices: Theory Appl. 4 (04) (2015) 1550019.
[115] G. Biroli, J.-P. Bouchaud, M. Potters, On the top eigenvalue of heavy-tailed random matrices, Europhys. Lett. 78 (1) (2007) 10001.
[116] M.J. Bowick, É. Brézin, Universal scaling of the tail of the density of eigenvalues in random matrix models, Phys. Lett. B 268 (1) (1991) 21–28.
[117] S.N. Majumdar, G. Schehr, Top eigenvalue of a random matrix: large deviations and third order phase transition, J. Stat. Mech. Theory Exp. 2014 (1) (2014) P01012.
[118] C. Nadal, S.N. Majumdar, A simple derivation of the Tracy–Widom distribution of the maximal eigenvalue of a Gaussian unitary random matrix, J. Stat. Mech. Theory Exp. 2011 (04) (2011) P04001.
[119] J. Ramirez, B. Rider, B. Virág, Beta ensembles, stochastic airy spectrum, and a diffusion, J. Amer. Math. Soc. 24 (4) (2011) 919–944.
[120] K. Johansson, Shape fluctuations and random matrices, Comm. Math. Phys. 209 (2) (2000) 437–476.
[121] J. Baik, G. Ben Arous, S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, Ann. Probab. (2005) 1643–1697.
[122] S. Péché, Universality results for the largest eigenvalues of some sample covariance matrix ensembles, Probab. Theory Related Fields 143 (3–4) (2009) 481–516.
[123] S. Péché, Universality of Local Eigenvalue Statistics for Random Sample Covariance Matrices (Ph.D. thesis), EPFL, 2003.
[124] W. Hachem, A. Hardy, J. Najim, A survey on the eigenvalues local behavior of large complex correlated wishart matrices, ESAIM: Proceedings and Surveys 51 (2015) 150–174.
[125] R. Allez, J. Bun, J.-P. Bouchaud, The eigenvectors of Gaussian matrices with an external source, arXiv preprint arXiv:1412.7108.
[126] R. Allez, J.-P. Bouchaud, Eigenvector dynamics under free addition, Random Matrices 03 (2014) 1450010.
[127] J. Bun, J.-P. Bouchaud, M. Potters, On the overlaps between eigenvectors of correlated random matrices, arXiv preprint arXiv:1603.04364.
[128] D. Hoyle, M. Rattray, Limiting Form of the Sample Covariance Eigenspectrum in PCA and Kernel PCA, in: Advances in Neural Information Processing Systems, 2003.
[129] H. Weyl, Inequalities between the two kinds of eigenvalues of a linear transformation, Proc. Nat. Acad. Sci. 35 (7) (1949) 408–411.
[130] J.M. Deutsch, Quantum statistical mechanics in a closed system, Phys. Rev. A 43 (4) (1991) 2046.
[131] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, vol. 2, Taylor & Francis, 2014.
[132] A. Takemura, An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population, Tech. Rep., DTIC Document, 1983.
[133] N. El Karoui, H. Kösters, Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods, arXiv preprint arXiv:1105.1404.
[134] L.H. Dicker, et al., Ridge regression and asymptotic minimax estimation over spheres of growing dimension, Bernoulli 22 (1) (2016) 1–37.
[135] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series, vol. 2, MIT press, Cambridge, MA, 1949.
[136] S. Pafka, I. Kondor, Noisy covariance matrices and portfolio optimization II, Physica A 319 (2003) 487–494.
[137] N. El Karoui, et al., High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation, Ann. Stat. 38 (6) (2010) 3487–3566.
[138] B. Collins, D. McDonald, N. Saad, Compound wishart matrices and noisy covariance matrices: risk underestimation, arXiv preprint arXiv:1306.5510.
[139] N.E. Karoui, On the realized risk of high-dimensional Markowitz portfolios, SIAM J. Financ. Math. 4 (1) (2013) 737–783.
[140] O. Ledoit, M. Wolf, Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks, Available at SSRN 2383361 (2014).
[141] D. Bartz, Advances in High-dimensional Covariance Matrix Estimation (Doctoral thesis), Technische Universität, Berlin, 2015.
[142] Z. Burda, J. Jurkiewicz, M.A. Nowak, G. Papp, I. Zahed, Free Lévy matrices and financial correlations, Phys. A 343 (2004) 694–700.
[143] D. Bartz, K.-R. Müller, Covariance Shrinkage for Autocorrelated Data, in: Advances in Neural Information Processing Systems, 2014, pp. 1592–1600.
[144] D. Bartz, K. Hatrick, C.W. Hesse, K.-R. Müller, S. Lemm, Directional variance adjustment: Bias reduction in covariance matrices based on factor analysis with an application to portfolio optimization, PLoS One 8 (7) (2013) e67503.
[145] M. Marsili, Dissecting financial markets: sectors and states, Quant. Finance 2 (4) (2002) 297–302.
[146] O. Ledoit, M. Wolf, Numerical implementation of the quest function, Tech. Rep., Department of Economics-University of Zurich, 2016.
[147] P. McCullagh, J.A. Nelder, Generalized Linear Models, vol. 37, CRC press, 1989.
[148] G. Chamberlain, M. Rothschild, Arbitrage, factor structure, and mean-variance analysis on large asset markets, Econometrica 51 (5) (1983) 1281–1304. www.jstor.org/stable/1912275.
[149] G. Kapetanios, A new method for determining the number of factors in factor models with large datasets, Tech. Rep., Working Paper, Department of Economics, Queen Mary, University of London, 2004.
[150] A. Onatski, Determining the number of factors from empirical distribution of eigenvalues, Rev. Econ. Stat. 92 (4) (2010) 1004–1016.
[151] D. Paul, J.W. Silverstein, No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix, J. Multivariate Anal. 100 (1) (2009) 37–57.
[152] R.C. Merton, An intertemporal capital asset pricing model, Econometrica (1973) 867–887.
[153] E.F. Fama, K.R. French, Common risk factors in the returns on stocks and bonds, J. Financ. Econ. 33 (1) (1993) 3–56.
[154] A. Tanskanen, J. Lukkarinen, K. Vatanen, Random factor approach for large sets of equity time-series, arXiv preprint arXiv:1604.05896.
[155] R. Chicheportiche, J.-P. Bouchaud, A nested factor model for non-linear dependencies in stock returns, Quant. Finance 15 (11) (2015) 1–16.
[156] J. Bun, J.-P. Bouchaud, M. Potters, Cleaning correlation matrices, Risk magazine.
[157] H.M. Markowitz, Portfolio Selection: Efficient Diversification of Investments, vol. 16, Yale University Press, 1968.
[158] M. Tumminello, F. Lillo, R.N. Mantegna, Kullback-leibler distance as a measure of the information filtered from multivariate data, Phys. Rev. E 76 (3) (2007) 031123.
[159] O. Ledoit, M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, J. Empir. Finance 10 (5) (2003) 603–621.
[160] E. Pantaleo, M. Tumminello, F. Lillo, R.N. Mantegna, When do improved covariance matrix estimators enhance portfolio optimization? an empirical comparative study of nine estimators, Quant. Finance 11 (7) (2011) 1067–1080.
[161] R. Allez, J.-P. Bouchaud, Eigenvector dynamics: general theory and some applications, Phys. Rev. E 86 (4) (2012) 046202.
[162] T.A. Schmitt, D. Chetalova, R. Schäfer, T. Guhr, Non-stationarity in financial time series: Generic features and tail behavior, Europhys. Lett. 103 (5) (2013) 58003.
[163] S. Wang, R. Schäfer, T. Guhr, Average cross-responses in correlated financial market, arXiv preprint arXiv:1603.01586.

[164] S. Pafka, M. Potters, I. Kondor, Exponential weighting and random-matrix-theory-based filtering of financial covariance matrices for portfolio optimization, arXiv preprint cond-mat/0402573.
[165] I.M. Johnstone, Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence, Ann. Statist. 36 (6) (2008) 2638.
[166] Y. Yang, G. Pan, et al., Independence test for high dimensional data based on regularized canonical correlation coefficients, Ann. Statist. 43 (2) (2015) 467–500.
[167] V.C. Klema, A.J. Laub, The singular value decomposition: its computation and some applications, IEEE Trans. Automat. Control 25 (2) (1980) 164–176.
[168] G.W. Furnas, S. Deerwester, S.T. Dumais, T.K. Landauer, R.A. Harshman, L.A. Streeter, K.E. Lochbaum, Information retrieval using a singular value decomposition model of latent semantic structure, in: Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1988, pp. 465–480.
[169] O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Nat. Acad. Sci. 97 (18) (2000) 10101–10106.
[170] F. Benaych-Georges, R.R. Nadakuditi, The singular values and vectors of low rank perturbations of large rectangular random matrices, J. Multivariate Anal. 111 (2012) 120–135.
[171] M. Tumminello, F. Lillo, R.N. Mantegna, Correlation, hierarchies, and networks in financial markets, J. Econ. Behav. Organ. 75 (1) (2010) 40–58.
[172] I.I. Dimov, P.N. Kolm, L. Maclin, D.Y. Shiber, Hidden noise structure and random matrix models of stock correlations, Quant. Finance 12 (4) (2012) 567–572.
[173] M.-F. Bru, Wishart processes, J. Theoret. Probab. 4 (4) (1991) 725–751.
[174] P. Bourgade, H.-T. Yau, The eigenvector moment flow and local quantum unique ergodicity, Commun. Math. Phys. (2013) 1–48.
[175] T. Tao, http://terrytao.wordpress.com/2013/02/08/the-harish-chandra-itzykson-zuber-integral-formula/, 2013.
[176] A. Matytsin, On the large-N limit of the Itzykson-Zuber integral, Nuclear Phys. B 411 (1994) 805–820.
[177] J. Bun, J.P. Bouchaud, S.N. Majumdar, M. Potters, Instanton approach to large N Harish-Chandra-Itzykson-Zuber integrals, Phys. Rev. Lett. 113 (2014) 070201.
[178] A. Guionnet, M. Maïda, A Fourier view on the R-transform and related asymptotics of spherical integrals, J. Funct. Anal. 222 (2) (2005) 435–490.
[179] J.-B. Zuber, The large-N limit of matrix integrals over the orthogonal group, J. Phys. A 41 (38) (2008) 382001.
[180] A. Guionnet, O. Zeitouni, Large deviations asymptotics for spherical integrals, J. Funct. Anal. 188 (2) (2002) 461–515.
[181] B. Collins, A. Guionnet, E. Maurel-Segala, Asymptotics of unitary and orthogonal matrix integrals, Adv. Math. 222 (1) (2009) 172–215.
[182] T. Tanaka, Asymptotics of Harish-Chandra-Itzykson-Zuber integrals and free probability theory, in: Journal of Physics: Conference Series, vol. 95, IOP Publishing, 2008, 012002.
[183] E. Marinari, G. Parisi, F. Ritort, Replica field theory for deterministic models. ii. a non-random spin glass with glassy behaviour, J. Phys. A 27 (23) (1994) 7647.
[184] L. Erdős, Universality of Wigner random matrices: a survey of recent results, Russian Math. Surveys 66 (3) (2011) 507.
[185] C.W. Beenakker, Random-matrix theory of quantum transport, Rev. Modern Phys. 69 (3) (1997) 731.
[186] G. Ithier, F. Benaych-Georges, Thermalisation of a quantum system from first principles, arXiv preprint arXiv:1510.04352.
[187] R. Nandkishore, D.A. Huse, Many-body localization and thermalization in quantum statistical mechanics, Annual Review of Condensed Matter Physics 6 (1) (2015) 15–38.
[188] J. Eisert, M. Friesdorf, C. Gogolin, Quantum many-body systems out of equilibrium, Nat. Phys. 11 (2) (2015) 124–130.
[189] F.J. Dyson, A Brownian-motion model for the eigenvalues of a random matrix, J. Math. Phys. 3 (1962) 1191–1198.
[190] D. Shlyakhtenko, Random Gaussian band matrices and freeness with amalgamation, Int. Math. Res. Not. 1996 (20) (1996) 1013–1025.