The spectrum of kernel random matrices

Noureddine El Karoui [*]
Department of Statistics,
University of California, Berkeley

December 13, 2007

Abstract

We place ourselves in the setting of high-dimensional statistical inference, where the number of variables p in a dataset of interest is of the same order of magnitude as the number of observations n.

We consider the spectrum of certain kernel random matrices, in particular $n \times n$ matrices whose (i,j)-th entry is $f(X_i'X_j/p)$ or $f(\|X_i - X_j\|^2/p)$, where p is the dimension of the data, and X_i are independent data vectors. Here f is assumed to be a locally smooth function.

The study is motivated by questions arising in statistics and computer science, where these matrices are used to perform, among other things, non-linear versions of principal component analysis. Surprisingly, we show that in high-dimensions, and for the models we analyze, the problem becomes essentially linear - which is at odds with heuristics sometimes used to justify the usage of these methods. The analysis also highlights certain peculiarities of models widely studied in random matrix theory and raises some questions about their relevance as tools to model high-dimensional data encountered in practice.

1 Introduction

Recent years has seen newfound theoretical interest in the properties of large dimensional sample covariance matrices. With the increase in the size and dimensionality of datasets to be analyzed, questions have been raised about the practical relevance of information derived from classical asymptotic results concerning spectral properties of sample covariance matrices. To address these concerns, one line of analysis has been the consideration of asymptotics where both the sample size, n and the number of variables p in the dataset go to infinity, jointly, while assuming for instance that p/n had a limit.

This type of questions concerning the spectral properties of large dimensional matrices have been and are being addressed in variety of fields, from physics to various areas of mathematics. While the topic is classical, with the seminal contribution Wigner (1955) dating back from the 1950's, there has been renewed and vigorous interest in the study of large dimensional random matrices in the last decade or so. This has led to new insights and the appearance of new "canonical" distributions (Tracy and Widom (1994)), new tools (see Voiculescu (2000)) and, in Statistics, a sense that one needs to exert caution with familiar techniques of multivariate analysis when the dimension of the data gets large and the sample size is of the same order of magnitude as the dimension of the data.

So far in Statistics, this line of work has been concerned mostly with the properties of sample covariance matrices. In a seminal paper, Marčenko and Pastur (1967) showed a result that from a statistical standpoint may be interpreted as saying, roughly, that asymptotically, the histogram of the sample eigenvalues is a deterministic non-linear deformation of the histogram of population eigenvalues. Remarkably, they managed to characterize this deformation for fairly general population covariances. Their result was shown in great generality, and introduced new tools to the field, including one that has become ubiquitous, the

[*] I would like to thank Bin Yu for stimulating my interest in the questions considered in this paper and for interesting discussions on the topic. I would like to thank Elizabeth Purdom for discussions about kernel analysis and Peter Bickel for many stimulating discussions about random matrices and their relevance in statistics. Support from NSF grant DMS-0605169 is gratefully acknowledged. **AMS 2000 SC:** Primary: 62H10. Secondary: 60F99 **Key words and Phrases :** covariance matrices, kernel matrices, eigenvalues of covariance matrices, multivariate statistical analysis, high-dimensional inference, random matrix theory, machine learning, Hadamard matrix functions, concentration of measure. **Contact :** nkaroui@stat.berkeley.edu

Stieltjes transform of a distribution. In its best known form, their result says that when the population covariance is identity, and hence all the population eigenvalues are equal to 1, in the limit the sample eigenvalues are split and, if $p \leq n$, they are spread between $[(1-\sqrt{p/n})^2, (1+\sqrt{p/n})^2]$, according to a fully explicit density, known now as the density of the Marčenko-Pastur law. Their result was later rediscovered independently in Wachter (1978) (under slightly weaker conditions), and generalized to the case of non-diagonal covariance matrices in Silverstein (1995), under some particular distributional assumptions, which we discuss later in the paper.

On the other hand, recent developments have been concerned with fine properties of the largest eigenvalue of random matrices, which became amenable to analysis after mathematical breakthroughs which happened in the 1990's (see Tracy and Widom (1994), Tracy and Widom (1996) and Tracy and Widom (1998)). Classical statistical work on joint distribution of eigenvalues of sample covariance matrices (see Anderson (2003) for a good reference) then became usable for analysis in high-dimensions. In particular, in the case of gaussian distributions, with Id covariance, it was shown in Johnstone (2001) that the largest eigenvalue of the sample covariance matrix is Tracy-Widom distributed. More recent progress (El Karoui (2007c)) managed to carry out the analysis for essentially general population covariance. On the other hand, models for which the population covariance has a few separated eigenvalues have also been of interest: see for instance Paul (2007) and Baik and Silverstein (2006). Beside the particulars of the different type of fluctuations that can be encountered (Tracy-Widom, Gaussian or other), researchers have been able to precisely localize these largest eigenvalues. One striking aspect of those results is the fact that in the high-dimensional setting of interest, the largest eigenvalues are always positively biased, with the bias being sometime really large. (We also note that in the case of i.i.d data - which naturally less interesting in statistics - results on the localization of the largest eigenvalue have been available for quite some time now, after the works Geman (1980) and Yin et al. (1988) to cite a few.) This is naturally in sharp contrast to classical results of multivariate analysis, which show $\sqrt{n}$-consistency of all sample eigenvalues - though the possibility of bias is a simple consequence of Jensen's inequality.

On the other hand, there has been much less theoretical work on kernel random matrices. In particular, we are not aware of any work in the setting of high-dimensional data analysis. However, given the practical success and flexibility of these methods (we refer to Schölkopf and Smola (2002) for an introduction), it is natural to try to investigate theoretically their properties. Further, as illustrated in the data analytic part of Williams and Seeger (2000), the n/p boundedness assumption in not unrealistic, as far as applications of kernel methods are concerned. The aim of the present paper is to shed some theoretical light on the properties of these kernel random matrices, and to do so in relatively wide generality. We note that the choice of renormalization that we make is motivated in part by the arguments of Williams and Seeger (2000) and their practical choices of kernels for data of varying dimensions.

Existing theory on kernel random matrices (see for instance the interesting Koltchinskii and Giné (2000)), for fixed dimensional input data, predicts that the eigenvalues of kernel random matrices behave - at least for the largest ones - like the eigenvalues of the corresponding operator on $L^2(dP)$, if the data is i.i.d with probability distribution P. These insights have also been derived through more heuristic but nonetheless enlightening arguments in, for instance, Williams and Seeger (2000). By contrast, we show that for the models we analyze, kernel random matrices essentially behave like sample covariance matrices and hence their eigenvalues suffer from the same bias problems that affect sample covariance matrices in high-dimensions. In particular, if one were to try to apply the heuristics of Williams and Seeger (2000), which were developed for low-dimensional problems, to the high-dimensional case, the predictions would be quite wildly wrong. (A simple example is provided by the Gaussian kernel with i.i.d Gaussian data, where the computations can be done completely explicitly, as explained in Williams and Seeger (2000).) We also note that the scaling we use is different from the one used in low dimensions, where the matrices are scaled by $1/n$. This is because the high-dimensional problem would be completely degenerate if we used this normalization in our setting. However, our results still give information about the problem when it is scaled by $1/n$.

We note that from a random matrix point of view, our study is connected to the study of Euclidian random matrices and distance matrices, which is of some interest in, for instance, Physics. We refer to Bogomolny et al. (2003) and Bordenave (2006) for work in this direction in the low (or fixed) dimensional setting. We also note that at the level of generality we place ourselves in, the random matrices we study do not seem to be amenable to study through the classical tools of random matrix theory. Hence, beside their obvious statistical interest, they are also very interesting on purely mathematical grounds.

We now turn to the gist of our paper, which will show that high-dimensional kernel random matrices behave spectrally essentially like sample covariance matrices. We will get two types of results: in Theorems 1 and 2, we get a strong approximation result (in operator norm) for standard models studied in random matrix theory. In Theorems 3 and 4, we characterize the limiting spectral distribution of our kernel random matrices, for a wide class of data distributions. In Section 2, we also state clearly the consequences of our theorems and review the relevant theory of high-dimensional sample covariance matrices. From a technical standpoint, we adopt a point of view centered on the concentration of measure phenomenon, as exposed for instance in Ledoux (2001), as it provides a unified way to treat the two types of results we are interested in. Finally, we discuss in our conclusion (Section 3) the consequences of our results, and in particular some possible limitations of standard random matrix models as tools to model data encountered in practice.

2 Spectrum of kernel random matrices

Kernel random matrices do not seem to be amenable to analysis through the usual tools of random matrix theory. In particular, for general f, it seems difficult to carry out either a method of moments proof, or a Stieltjes transform proof, or a proof that relies on knowing the density of the eigenvalues of the matrix.

Hence, we take an indirect approach. Our strategy is to find approximations of the kernel random matrix that have two properties. First, the approximation matrix is analyzable or has already been analyzed in random matrix theory. Second, the quality of the approximation is good enough that spectral properties of the approximating matrix can be shown to carry over to the kernel matrix.

The strategy in the first two theorems is to derive an operator norm "consistent" approximation of our kernel matrix. In other words, if we call M our kernel matrix, we will find K such that $|||M-K|||_2 \to 0$, as n and p tend to ∞. Note that both M and K are real symmetric (and hence Hermitian) here. We explain after the statement of Theorem 1 why operator norm consistency is a desirable property. But let us say that in a nutshell, it implies consistency for each individual eigenvalue as well as eigenspaces corresponding to separated eigenvalues.

For the second set of theorems (Theorems 3 and 4), we will relax the distributional assumptions made on the data, but at the expense of the precision of the results we will obtain: we will characterize the limiting spectral distribution of our kernel random matrices.

Our theorems below show that kernel random matrices can be well approximated by matrices that are closely connected to large-dimensional covariance matrices. The spectral properties of those matrices have been the subject of a significant amount of work in recent and less recent years, and hence this knowledge, or at least part of it, can be transferred to kernel random matrices. In particular, we refer the reader to Marčenko and Pastur (1967), Wachter (1978), Geman (1980), Yin et al. (1988), Silverstein (1995), Bai and Silverstein (1998), Johnstone (2001), Baik and Silverstein (2006), Paul (2007), El Karoui (2007c), Bai et al. (2007) and El Karoui (2007a) for some of the most statistically relevant results in this area. We review some of them now.

2.1 Some results on large dimensional sample covariance matrices

Since our main theorems are approximating theorems, we first wish to state some of the properties of the objects we will use to approximate kernel random matrices. In what follows, we consider an $n \times p$ data matrix, with, say p/n having a finite non-zero limit. Most of the results that have been obtained are of two types: either they are so-called "bulk" results and concern essentially the spectral distribution (or loosely speaking the histogram of eigenvalues) of the random matrices of interest. Or they concern the localization and fluctuation behavior of extreme eigenvalues of these random matrices.

2.1.1 Spectral distribution results

An object of great interest in random matrix theory is the spectral distribution of random matrices. Let us call l_i the decreasingly ordered eigenvalues of our random matrix, and let us assume we are working with an $n \times n$ matrix, M_n. The empirical spectral distribution of a $n \times n$ matrix is the probability measure which puts mass $1/n$ at each of its eigenvalues. In other words, if we call F_n this probability measure, we have

$$dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{l_i}(x) \; .$$

Note that the histogram of eigenvalues represent an integrated version of this measure.

For random matrices, this measure F_n is naturally a random measure. A very striking result in the area of covariance matrices is that if we observe i.i.d data vectors X_i, with $X_i = \Sigma_p^{1/2} Y_i$, where Y_i is a vector with i.i.d entries, under weak moment conditions on Y_i, F_n converges to a *non-random* measure, which we call F.

We call the models $X_i = \Sigma_p^{1/2} Y_i$ the standard models of random matrix theory because most results have been derived under these assumptions. In particular, striking results (Geman (1980), Bai and Silverstein (1998), Bai and Silverstein (1999)) show, among many other things, that when the entries of the vector Y have 4 moments, the largest eigenvalues of the sample covariance matrix $X'X/n$, where X_i now occupies the first row of the $n \times p$ matrix X, stay close to the endpoint of the support of F.

A very natural question is therefore to try to characterize F. Except in particular situations, it is difficult to do so explicitly. However, it is possible to characterize a certain transformation of F. The tool of choice in this context is the *Stieltjes transform* of a distribution. It is a function defined on $\mathbb{C}^+$ by the formula, if we call St_F the Stieltjes transform of F,

$$\mathrm{St}_F(z) = \int \frac{dF(\lambda)}{\lambda - z} \, , \mathrm{Im}\,[z] > 0.$$

In particular for empirical spectral distributions, we see that, if F_n is the spectral distribution of the matrix M_n,

$$\mathrm{St}_{F_n}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{l_i - z} = \frac{1}{n}\mathrm{trace}\left((M_n - z\mathrm{Id})^{-1}\right) \; .$$

The importance of the Stieltjes transform in the context of random matrix theory stems from two facts: on the one hand, it is connected fairly explicitly to the matrices that are being analyzed. On the other hand, pointwise convergence of Stieltjes transform implies weak convergence of distributions, if a certain mass preservation condition is satisfied. This is how a number of bulk results are therefore proved. For a clear and self-contained introduction to the connection between Stieltjes transforms and weak convergence of probability measures, we refer the reader to Geronimo and Hill (2003).

The result of Marčenko and Pastur (1967), later generalized by Silverstein (1995) for standard random matrix models with non-diagonal covariance, and more recently by El Karoui (2007a) away from those standard models, is a functional characterization of the limit F. If one calls $w_n(z)$ the Stieltjes transform of the empirical spectral distribution of XX'/n, $w_n(z)$ converges pointwise (and almost surely after Silverstein (1995)) to a *non-random* $w(z)$, which, as a function, is a Stieltjes transform. Moreover, w, the Stieltjes transform of F, satisfies the equation, if $p/n \to \rho$:

$$-\frac{1}{w(z)} = z - \rho \int \frac{\lambda dH(\lambda)}{1 + \lambda w} \; ,$$

where H is the limiting spectral distribution of Σ_p, assuming that such a distribution exists. We note that Silverstein (1995) proved the result under a second moment condition on the entries of Y_i.

From this result, Marčenko and Pastur (1967) derived that in the case where $\Sigma_p = \mathrm{Id}$, and hence $dH = \delta_1$, the limiting spectral distribution has a limit whose density is, if $\rho \leq 1$,

$$f_\rho(x) = \frac{1}{2\pi\rho}\frac{\sqrt{(b-x)(x-a)}}{x}$$

where $a = (1 - \rho^{1/2})^2$ and $b = (1 + \rho^{1/2})^2$. The difference between the population spectral distribution (a point mass at 1) and the limit of the empirical spectral distribution is very striking.

2.1.2 Largest eigenvalues results

Another line of work has been focused on the behavior of extreme eigenvalues of sample covariance matrices. In particular, Geman (1980) showed, under some moment conditions, that when $\Sigma_p = \text{Id}_p$, $l_1(X'X/n) \to (1+\sqrt{p/n})^2$. In other words, the largest eigenvalue stays close to the endpoint of the limiting spectral distribution of $X'X/n$. This result was later generalized in Yin et al. (1988), and shown to be true under the assumption of finite 4th moment only, for data with mean 0. In recent years, fluctuation results have been obtained for this largest eigenvalue, which is of practical interest in PCA. Under Gaussian assumptions, Johnstone (2001) and El Karoui (2003) (see also Forrester (1993) and Johansson (2000)) showed that the fluctuations of the largest eigenvalue are Tracy-Widom. For the general covariance case, similar results, as well as localization information were recently obtained in El Karoui (2007c). We note that the localization information (i.e a formula) that was discovered in this latter paper, through appeal to Bai and Silverstein (1998), was shown to hold for a wide variety of standard random matrix models. We refer the interested reader to Fact 2 in El Karoui (2007c) for more information. Interesting work has also been done on so-called "spiked" models, where a few population eigenvalues are separated from the bulk of them. In particular, in the case where all population eigenvalues are equal, except for one that is significantly larger (see Baik et al. (2005) for the discovery of a very interesting phase transition), Paul (2007) showed, in the Gaussian case, inconsistency of the largest sample eigenvalue, as well as the fact that the angle between the population and sample principal eigenvectors is bounded away from 0. Paul (2007) also obtained fluctuation information about the largest empirical eigenvalue. Finally, we note that the same inconsistency of eigenvalue result was also obtained in Baik and Silverstein (2006), beyond the Gaussian case.

2.1.3 Notations

Let us now define some notations and add some clarifications.

First, let us remind the reader that if A and B are two rectangular matrices, AB and BA have the same eigenvalues, except for possibly, a certain number of zeros. We will make repeated use of this fact, both for matrices like $X'X$ and XX' and in the case where A and B are both square, in which case, AB and BA have the same eigenvalues.

We will also need various norms on matrices. We will use the so-called operator norm, which we denote by $|||A|||_2$, which corresponds to the largest singular value of A, i.e $\max_i \sqrt{l_i(A^*A)}$. We occasionally denote the largest singular value of A by $\sigma_1(A)$. Clearly, for positive semi-definite matrices, the largest singular value is equal to the largest eigenvalue. Finally, we will sometime need to use the Frobenius (or Hilbert-Schmidt) norm of a matrix A. We denote it by $\|A\|_F$. By definition, it is simply

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2 \; .$$

Further, we use $\circ$ to denote the Hadamard (i.e entrywise) product of two matrices. We denote by μ_m the m-th moment of a random variable. Note that by a slight abuse of notation, we might also use the same notation to refer to the m-th absolute moment (i.e $E|X|^m$) of a random variable, but if there is any ambiguity, we will naturally precise which definition we are using.

Finally, in the discussion of standard random matrix models that follows, there will be arrays of random variables and a.s convergence. We work with random variables defined on a common probability space. To each ω corresponds an infinite dimensional array of numbers. The $n \times p$ matrices we will use in what follows are the "upper-left" corner of this array.

We now turn to the study of kernel random matrices. We will show that we can approximate them by matrices that are closely connected to sample covariance matrices in high-dimensions and, therefore, that a number of the results we just reviewed also apply to them.

2.2 Inner-product kernel matrices: $\mathbf{f(X_i'X_j/p)}$

Theorem 1 (Spectrum of inner product kernel random matrices)**.** *Let us assume that we observe n i.i.d random vectors, X_i in $\mathbb{R}^p$. Let us consider the kernel matrix M with entries*

$$M_{i,j} = f\left(\frac{X_i'X_j}{p}\right) .$$

We assume that

a) $n \asymp p$, i.e n/p and p/n remain bounded as $p \to \infty$

b) Σ is a positive semi-definite $p \times p$ matrix , and $|||\Sigma|||_2 = \sigma_1(\Sigma)$ remains bounded in p.

c) $\mathrm{trace}\,(\Sigma)/p$ has a finite limit.

d) $X_i = \Sigma^{1/2}Y_i$.

e) The entries of Y_i are i.i.d, and have $4+\epsilon$ moments, for some $\epsilon > 0$.

f) f is a C^1 function in a neighborhood of $\lim \mathrm{trace}\,(\Sigma)/p$ and a C^3 function in a neighborhood of 0.

Under these assumptions, the kernel matrix M can (in probability) be approximated consistently in operator norm, when p and n tend to ∞, by the matrix K, where

$$K = \left(f(0) + f''(0)\frac{\mathrm{trace}\,(\Sigma^2)}{2p^2}\right)11' + f'(0)\frac{XX'}{p} + \upsilon_p \mathrm{Id}_n ,$$
$$\upsilon_p = f\left(\frac{\mathrm{trace}\,(\Sigma)}{p}\right) - f(0) - f'(0)\frac{\mathrm{trace}\,(\Sigma)}{p} .$$

In other words,

$$|||M - K|||_2 \to 0 \text{ , in probability, when } p \to \infty .$$

The advantages of obtaining an operator norm consistent estimator are many. We list some here:

- Asymptotically, M and K have the same j-largest eigenvalue: this is simply because for symmetric matrices, if l_j is the j-th largest eigenvalue of a matrix, Weyl's inequality implies that

$$|l_j(M) - l_j(K)| \leq |||M - K|||_2 .$$

- The limiting spectral distributions of M and K (if they exist) are the same. This is a consequence of Lemma 1 below.
- We have subspace consistency for eigenspaces corresponding to separated eigenvalues. (For a proof, we refer to El Karoui (2007b), Corollary 3.)

(Note that the statements we just made assume that both M and K are symmetric, which is the case here.)

The strategy for the proof is the following. According to the results of Lemma A-3, the matrix $X_i'X_j/p$ has "small" entries off the diagonal, whereas on the diagonal, the entries are essentially constant and equal to $\mathrm{trace}\,(\Sigma)/p$. Hence, it is natural to try to use the δ-method (i.e do a Taylor expansion) entry by entry. By contrast to standard problems in Statistics, the fact that we have to perform n^2 of those Taylor expansions means that the second order term is not negligible, a priori. The proof shows that this approach can be carried out rigorously, and that, surprisingly, the second order term is not too complicated to approximate in operator norm. Also, it is shown that the third order term plays essentially no role.

Proof. First, let us call

$$\tau \triangleq \frac{\text{trace}\,(\Sigma)}{p} \; .$$

We can rewrite our kernel matrix as:

$$f(X_i'X_j/p) = f(0) + f'(0)X_i'X_j/p + \frac{f''(0)}{2}(X_i'X_j/p)^2 + \frac{f^{(3)}(\xi_{i,j})}{6}(X_i'X_j/p)^3 \; , \text{ if } i \neq j \; ,$$
$$f(\|X_i\|_2^2/p) = f(\tau) + f'(\xi_{i,i})\left(\frac{\|X_i\|_2^2}{p} - \tau\right) \text{ on the diagonal.}$$

The proof can be separated in different steps. We will break the kernel matrix into a diagonal term and an off diagonal term. The results of Lemma A-3, after they are shown, will allow us to take care of the diagonal matrix at relatively lost cost. So we postpone that part of the analysis to the end of the proof and we first focus on the off-diagonal matrix.

A) Study of the off-diagonal matrix

• Truncation and centralization

This step is classical. Following the arguments of Lemma 2.2 in Yin et al. (1988), we see that because we have assumed that we have $4+\epsilon$ moments, and $n \asymp p$, the array $Y = Y_{1\leq i\leq n, 1\leq j\leq p}$ is almost surely equal to the array $\widetilde{Y}$ of same dimensions, with

$$\widetilde{Y}_{i,j} = Y_{i,j}1_{|Y_{i,j}|\leq B_p} \; , \text{ where } B_p = p^{1/2-\delta} \; , \text{ and } \delta > 0.$$

We will therefore carry out the analysis on this $\widetilde{Y}$ array. Note that most of the results we will rely on require vectors of i.i.d entries with mean 0. Of course, $\widetilde{Y}_{i,j}$ has in general a mean different from 0. In other words, if we call $\mu = \mathbf{E}\left(\widetilde{Y}_{i,j}\right)$, we need to show that we do not lose anything in operator norm by replacing $\widetilde{Y}_i$'s by U_i's with $U_i = \widetilde{Y}_i - \mu 1$. Note that, as seen in Lemma A-3, by plugging in $t = 1/2 - \delta$ in the notation of this lemma, which corresponds to the $4+\epsilon$ moment assumption here, we have

$$|\mu| \leq p^{-3/2-\delta} \; .$$

Now let us call S the matrix XX'/p, except that its diagonal is replaced by zeros. From Yin et al. (1988), and the fact that n/p stays bounded, we know that $|||XX'/p|||_2 \leq \sigma_1(\Sigma)|||YY'|||_2/p$ stays bounded. Using Lemma A-3, we see that the diagonal of XX'/p stays bounded a.s in operator norm. Therefore, $|||S|||_2$ is bounded a.s.

Now, as in the proof of Lemma A-3, we have

$$S_{i,j} = \frac{U_i'\Sigma U_j}{p} + \mu\left(\frac{1'\Sigma U_j}{p} + \frac{1'\Sigma U_i}{p}\right) + \mu^2\frac{1'\Sigma 1}{p} \triangleq \frac{U_i'\Sigma U_j}{p} + R_{i,j} \text{ a.s} \; .$$

Note that this equality is true a.s only because it involves replacing Y by $\widetilde{Y}$. The proof of Lemma A-3 shows that

$$|R_{i,j}| \leq \mu\, 2\sigma_1^{1/2}(\Sigma)(\sigma_1^{1/2}(\Sigma) + p^{-\delta/2}) + \mu^2\sigma_1(\Sigma) \text{ a.s} \; .$$

We conclude that, for some constant C,

$$\|R\|_F^2 \leq Cn^2\mu^2 \leq Cn^2p^{-3-2\delta} \rightarrow \text{ a.s} \; .$$

Therefore $|||R|||_2 \rightarrow 0$ a.s . In other words, if we call S_U the matrix with i,j entry $U_i'\Sigma U_j/p$ off the diagonal and 0 on the diagonal,

$$|||S - S_U|||_2 \rightarrow 0 \text{ a.s} \; .$$

Now it is a standard result on Hadamard products (see for instance, Bhatia (1997), Problem I.6.13, or Horn and Johnson (1994), Theorems 5.5.1 and 5.5.15) that for two matrices A and B, $|||A\circ B|||_2 \leq |||A|||_2|||B|||_2$. Since the Hadamard product is commutative, we have

$$S\circ S - S_U\circ S_U = (S + S_U)\circ(S - S_U) \; .$$

We conclude that

$$|||S \circ S - S_U \circ S_U|||_2 \leq |||S - S_U|||_2(|||S|||_2 + |||S_U|||_2) \to 0 \text{ a.s } ,$$

since $|||S - S_U|||_2 \to 0$ a.s , and $|||S|||_2$ and hence $|||S_U|||_2$ stay bounded, a.s.

The conclusion of this study is that to approximate the second order term in operator norm, it is enough to work with S_U and not S, and hence, very importantly, with bounded random variables with zero mean. Further, the proof of Lemma A-3 makes clear that σ_U^2, the variance of the $U_{i,j}$'s, goes to 1, the variance of the $Y_{i,j}$'s very fast. So if we can approximate $U_i' \Sigma U_j / (p \sigma_U^2)$ consistently in operator norm by a matrix whose operator norm is bounded, this same matrix will constitute an operator norm approximation of $U_i' \Sigma U_j / p$.

In other words, we can assume that the random variables we will be working with have variance 1 without loss of generality, and that they have mean 0 and are bounded.

• **Control of the second order term**

As we just explained, we assume from now on in all the work concerning the second order term that the vectors Y_i have mean 0 and are bounded by $B_p = p^{1/2-\delta}$. This is because we just saw that replacing Y_i by U_i would not change (a.s and asymptotically) the operator norm of the matrix to be studied.

The control of the second order term turns out to be the most delicate part of the analysis, and the only place where we need the assumption that $X_i = \Sigma^{1/2} Y_i$. Let us call W the matrix with entries

$$W_{i,j} = \begin{cases} \frac{(X_i' X_j)^2}{p^2} \text{ , if } i \neq j \\ 0, \text{ if } i = j \end{cases}$$

Note that, when $i \neq j$,

$$\mathbf{E}\left(W_{i,j}\right) = \mathbf{E}\left(\text{trace}\left(X_i' X_j X_j' X_i\right)\right)/p^2 = \mathbf{E}\left(\text{trace}\left(X_j X_j' X_i X_i'\right)\right)/p^2 = \text{trace}\left(\Sigma^2\right)/p^2 .$$

Because we assume that trace $(\Sigma)/p$ has a finite limit, and n/p stays bounded away from 0, we see that the matrix $\mathbf{E}(W)$ has a largest eigenvalue that, in general, does not go to 0. Our aim is to show that W can be approximated in operator norm by this constant matrix. So let us consider the matrix $\widetilde{W}$ with entries

$$\widetilde{W}_{i,j} = \begin{cases} \frac{(X_i' X_j)^2}{p^2} - \text{trace}\left(\Sigma^2\right)/p^2 \text{ , if } i \neq j \\ 0, \text{ if } i = j \end{cases}$$

Simple computations show that the expected Frobenius norm squared of this matrix does not go to 0. Hence more subtle arguments are needed to control its operator norm. We will show that $\mathbf{E}\left(\text{trace}\left(\widetilde{W}^4\right)\right)$ goes to zero, which implies that $\mathbf{E}\left(|||\widetilde{W}|||_2^4\right)$ goes to zero, because $\widetilde{W}$ is real symmetric.

The elements contributing to trace $\left(\widetilde{W}^4\right)$ are generally of the form $\widetilde{W}_{i,j}\widetilde{W}_{j,k}\widetilde{W}_{k,l}\widetilde{W}_{l,i}$. We first focus on the case where all these indices (i,j,k,l) are different. Recall that $X_i = \Sigma^{1/2} Y_i$, where Y_i has i.i.d entries. We want to compute $\mathbf{E}\left(\widetilde{W}_{i,j}\widetilde{W}_{j,k}\widetilde{W}_{k,l}\widetilde{W}_{l,i}\right)$, so it is natural to focus first on

$$\mathbf{E}\left(\widetilde{W}_{i,j}\widetilde{W}_{j,k}\widetilde{W}_{k,l}\widetilde{W}_{l,i} \,|Y_i, Y_k\right)$$

Now, note that

$$\widetilde{W}_{i,j} = \frac{1}{p^2}\left\{Y_i' \Sigma Y_j Y_j' \Sigma Y_i - \text{trace}\left(\Sigma^2\right)\right\} = \frac{1}{p^2}\left\{Y_i' \Sigma (Y_j Y_j' - \text{Id}) \Sigma Y_i + \text{trace}\left(\Sigma^2 (Y_i Y_i' - \text{Id})\right)\right\} .$$

Hence, calling

$$M_j \triangleq Y_j Y_j' - \text{Id} ,$$

we have

$$p^4 \widetilde{W}_{i,j}\widetilde{W}_{j,k} = (Y_i' \Sigma M_j \Sigma Y_i Y_k' \Sigma M_j \Sigma Y_k) + (Y_i' \Sigma M_j \Sigma Y_i)\text{trace}\left(\Sigma^2 M_k\right) \\ + (Y_k' \Sigma M_j \Sigma Y_k)\text{trace}\left(\Sigma^2 M_i\right) + \text{trace}\left(\Sigma^2 M_i\right)\text{trace}\left(\Sigma^2 M_k\right) .$$

Now, of course, we have $\mathbf{E}\left(M_j\right) = \mathbf{E}\left(M_j | Y_i, Y_k\right) = 0$. Hence,

$$p^4 \mathbf{E}\left(\widetilde{W}_{i,j}\widetilde{W}_{j,k}\left|Y_i, Y_k\right.\right) = \left(Y_i'\Sigma \mathbf{E}\left(M_j \Sigma Y_i Y_k' \Sigma M_j \left|Y_i, Y_k\right.\right) \Sigma Y_k\right) + \operatorname{trace}\left(\Sigma^2 M_i\right) \operatorname{trace}\left(\Sigma^2 M_k\right) \ .$$

Now, note that if M is deterministic, we have, since $\mathbf{E}\left(Y_j Y_j'\right) = \mathrm{Id}$,

$$\mathbf{E}\left(M_j M M_j\right) = \mathbf{E}\left(Y_j Y_j' M Y_j Y_j'\right) - M \ .$$

If we now use Lemma A-1, and in particular Equation A-1, page 24, we finally have, recalling that here $\sigma^2 = 1$,

$$\begin{aligned}\mathbf{E}\left(M_j M M_j\right) &= (M + M') + (\mu_4 - 3)\mathrm{diag}(M) + \operatorname{trace}\left(M\right)\mathrm{Id} - M \\ &= M' + (\mu_4 - 3)\mathrm{diag}(M) + \operatorname{trace}\left(M\right)\mathrm{Id}\end{aligned}$$

In the case of interest here, we have $M = \Sigma Y_i Y_k' \Sigma$, and the expectation is to be understood conditionally on Y_i, Y_k, but because we have assumed that the indices are different, and the Y_m's are independent, we can do the computation of the conditional expectation as if M were deterministic. Therefore, we have

$$\begin{aligned}\left(Y_i'\Sigma \mathbf{E}\left(M_j \Sigma Y_i Y_k' \Sigma M_j \left|Y_i, Y_k\right.\right) \Sigma Y_k\right) &= Y_i'\Sigma\left[\Sigma Y_k Y_i' \Sigma + (\mu_4 - 3)\mathrm{diag}(\Sigma Y_i Y_k' \Sigma) + (Y_k' \Sigma^2 Y_i)\mathrm{Id}\right]\Sigma Y_k \\ &= \left[(Y_i'\Sigma^2 Y_k)^2 + (\mu_4 - 3) Y_i' \Sigma \mathrm{diag}(\Sigma Y_i Y_k' \Sigma)\Sigma Y_k + (Y_i'\Sigma^2 Y_k)^2\right]\end{aligned}$$

Naturally, we have $\mathbf{E}\left(\widetilde{W}_{i,j}\widetilde{W}_{j,k}\left|Y_i, Y_k\right.\right) = \mathbf{E}\left(\widetilde{W}_{k,l}\widetilde{W}_{l,i}\left|Y_i, Y_k\right.\right)$, and therefore, by using properties of conditional expectation, since all the indices are different,

$$p^8 \mathbf{E}\left(\widetilde{W}_{i,j}\widetilde{W}_{j,k}\widetilde{W}_{k,l}\widetilde{W}_{l,i}\right) = \mathbf{E}\left(\left[2(Y_i'\Sigma^2 Y_k)^2 + (\mu_4 - 3) Y_i'\Sigma \mathrm{diag}(\Sigma Y_i Y_k' \Sigma)\Sigma Y_k + \operatorname{trace}\left(\Sigma^2 M_i\right) \operatorname{trace}\left(\Sigma^2 M_k\right)\right]^2\right) \ .$$

Now by convexity, we have $(a+b+c)^2 \leq 3(a^2 + b^2 + c^2)$, so to control the above expression, we just need to control the square of each of the terms appearing in the above expression. Let us start by the term $\mathbf{E}\left((Y_i'\Sigma^2 Y_k)^4\right)$. A simple re-writing shows that

$$(Y_i'\Sigma^2 Y_k)^4 = Y_i'\Sigma^2 Y_k Y_k' \Sigma^2 Y_i Y_i' \Sigma^2 Y_k Y_k' \Sigma^2 Y_i \ .$$

Using Equation (A-1) in Lemma A-1, we therefore have, using the fact that $\Sigma^2 Y_i Y_i' \Sigma^2$ is symmetric,

$$\begin{aligned}\mathbf{E}\left((Y_i'\Sigma^2 Y_k)^4 \left|Y_i\right.\right) &= Y_i'\Sigma^2\left[2\Sigma^2 Y_i Y_i' \Sigma^2 + (\mu_4 - 3)\mathrm{diag}(\Sigma^2 Y_i Y_i' \Sigma^2) + \operatorname{trace}\left(\Sigma^2 Y_i Y_i' \Sigma^2\right)\mathrm{Id}\right]\Sigma^2 Y_i \\ &= 3(Y_i'\Sigma^4 Y_i)^2 + (\mu_4 - 3) Y_i'\Sigma^2 \mathrm{diag}(\Sigma^2 Y_i Y_i' \Sigma^2)\Sigma^2 Y_i \ .\end{aligned}$$

Finally, we have, using Equation (A-2) in Lemma A-1,

$$\begin{aligned}\mathbf{E}\left((Y_i'\Sigma^2 Y_k)^4\right) = 3&\left[2\operatorname{trace}\left(\Sigma^4\right) + (\operatorname{trace}\left(\Sigma^4\right))^2 + (\mu_4 - 3)\operatorname{trace}\left(\Sigma^4 \circ \Sigma^4\right)\right] \\ &+ (\mu_4 - 3)\mathbf{E}\left(Y_i'\Sigma^2 \mathrm{diag}(\Sigma^2 Y_i Y_i' \Sigma^2)\Sigma^2 Y_i\right) \ .\end{aligned}$$

Now, we have

$$Y_i'\Sigma^2 \mathrm{diag}(\Sigma^2 Y_i Y_i' \Sigma^2)\Sigma^2 Y_i = \operatorname{trace}\left(\Sigma^2 Y_i Y_i' \Sigma^2 \mathrm{diag}(\Sigma^2 Y_i Y_i' \Sigma^2)\right) = \operatorname{trace}\left(\Sigma^2 Y_i Y_i' \Sigma^2 \circ \Sigma^2 Y_i Y_i' \Sigma^2\right) \ .$$

Calling $v_i = \Sigma^2 Y_i$, we note that the matrix under the trace is $(v_i v_i') \circ (v_i v_i') = (v_i \circ v_i)(v_i \circ v_i)'$ (see Horn and Johnson (1990), p. 458 or Horn and Johnson (1994), p. 307). Hence,

$$Y_i'\Sigma^2 \mathrm{diag}(\Sigma^2 Y_i Y_i' \Sigma^2)\Sigma^2 Y_i = \|v_i \circ v_i\|_2^2 \ .$$

Now let us call m_k the k-th column of the matrix Σ^2. Using the fact that Σ^2 is symmetric, we see that the k-th entry of the vector v_i is $v_i(k) = m_k' Y_i$. So $v_i(k)^4 = Y_i' m_k m_k' Y_i Y_i' m_k m_k' Y_i$. Calling $\mathcal{M}_k = m_k m_k'$, we see using Equation (A-2) in Lemma A-1 that

$$\mathbf{E}\left(v_i(k)^4\right) = 2\operatorname{trace}\left(\mathcal{M}_k^2\right) + \left[\operatorname{trace}\left(\mathcal{M}_k\right)\right]^2 + (\mu_4 - 3)\operatorname{trace}\left(\mathcal{M}_k \circ \mathcal{M}_k\right) \ .$$

Using the definition of $\mathcal{M}_k$, we finally get that

$$\mathbf{E}\left(v_i(k)^4\right) = 3\|m_k\|_2^4 + (\mu_4 - 3)\|m_k \circ m_k\|_2^2 \; .$$

Now, note that if A is a generic matrix, and A_k is its $k-th$ column, denoting by e_k the k-th vector of the canonical basis, we have $A_k = Ae_k$ and hence $\|A_k\|_2^2 = e_k' A' A e_k \leq \sigma_1^2(A)$, where $\sigma_1(A)$ is the largest singular value of A. So in particular, if we call $\lambda_1(B)$ the largest eigenvalue of a positive semi-definite matrix B, we have $\|m_k\|_2^4 \leq \lambda_1(\Sigma^4)\|m_k\|_2^2$.

After recalling the definition of m_k, and using the fact that $\sum_k \|m_k \circ m_k\|_2^2 = \|\Sigma^2 \circ \Sigma^2\|_F^2$, we deduce that

$$\begin{aligned}\mathbf{E}\left(\|v_i \circ v_i\|_2^2\right) &= 3\sum_k \|m_k\|_2^4 + (\mu_4 - 3)\sum_k \|m_k \circ m_k\|_2^2 \\ &\leq 3\lambda_1(\Sigma^4)\text{trace}\left(\Sigma^4\right) + (\mu_4 - 3)\text{trace}\left(\left[\Sigma^2 \circ \Sigma^2\right]^2\right) \; .\end{aligned}$$

Therefore, we can conclude that

$$\mathbf{E}\left((Y_i'\Sigma^2 Y_k)^4\right) \leq 3\lambda_1(\Sigma^4)\text{trace}\left(\Sigma^4\right) + (\mu_4 - 3)\text{trace}\left(\left[\Sigma^2 \circ \Sigma^2\right]^2\right) \; .$$

Now recall that, according to Theorem 5.5.19 in Horn and Johnson (1994), if A and B are positive semidefinite matrices, if $\lambda(A \circ B) \prec_w \text{d}(A) \circ \lambda(B)$, where $\lambda(B)$ is the vector of decreasingly ordered eigenvalues of B, and $\text{d}(A)$ denotes the vector of decreasingly ordered diagonal entries of A (because all the matrices are positive semidefinite, their eigenvalues are their singular values). Here $\prec_w$ denotes weak (sub)majorization. In our case, of course, $A = B = \Sigma^2$. Using the results of Example II.3.5 (iii) in Bhatia (1997), with the function $\phi(x) = x^2$, we see that

$$\text{trace}\left((\Sigma^2 \circ \Sigma^2)^2\right) = \sum \lambda_i^2(\Sigma^2 \circ \Sigma^2) \leq \sum d_i^2(\Sigma^2)\lambda_i^2(\Sigma^2) \leq \lambda_1(\Sigma^4)\text{trace}\left(\Sigma^4\right) \; .$$

Finally, we have

$$\mathbf{E}\left((Y_i'\Sigma^2 Y_k)^4\right) \leq (3 + |\mu_4 - 3|)\lambda_1(\Sigma^4)\text{trace}\left(\Sigma^4\right) \tag{1}$$

This bounds the first term in our upper bound.

Let us now turn to the third term. By independence of Y_i and Y_k, it is enough to understand $\mathbf{E}\left(\left[\text{trace}\left(\Sigma^2 M_i\right)\right]^2\right)$. Note that

$$\mathbf{E}\left(\left[\text{trace}\left(\Sigma^2 M_i\right)\right]^2\right) = \mathbf{E}\left(\left[Y_i'\Sigma^2 Y_i - \text{trace}\left(\Sigma^2\right)\right]^2\right) = \mathbf{E}\left(Y_i'\Sigma^2 Y_i Y_i'\Sigma^2 Y_i\right) - \text{trace}\left(\Sigma^2\right)^2 \; .$$

Using Equation (A-2) in Lemma A-1, we conclude that

$$\mathbf{E}\left(\left[\text{trace}\left(\Sigma^2 M_i\right)\right]^2\right) = 2\text{trace}\left(\Sigma^4\right) + (\mu_4 - 3)\text{trace}\left(\Sigma^2 \circ \Sigma^2\right) \; .$$

Using the fact that we know the diagonal of $\Sigma^2 \circ \Sigma^2$, we conclude that,

$$\mathbf{E}\left(\left[\text{trace}\left(\Sigma^2 M_i\right)\right]^2\left[\text{trace}\left(\Sigma^2 M_k\right)\right]^2\right) \leq \left\{2\text{trace}\left(\Sigma^4\right) + |\mu_4 - 3|\lambda_1(\Sigma^2)\text{trace}\left(\Sigma^2\right)\right\}^2 \; . \tag{2}$$

Finally, let us turn to the middle term. Before we square it, it has the form $Y_i'\Sigma\text{diag}(\Sigma Y_k Y_i'\Sigma)\Sigma Y_k$. Call $u_k = \Sigma Y_k$. Making the same computations as above, we find that

$$\begin{aligned}Y_i'\Sigma\text{diag}(\Sigma Y_k Y_i'\Sigma)\Sigma Y_k &= \text{trace}\left(\text{diag}(\Sigma Y_k Y_i'\Sigma Y_k Y_i'\Sigma\right) \\ &= \text{trace}\left((\Sigma Y_k Y_i'\Sigma) \circ (\Sigma Y_k Y_i'\Sigma)\right) \\ &= \text{trace}\left((u_k u_i') \circ (u_k u_i')\right) = \text{trace}\left((u_k \circ u_k)(u_i \circ u_i)'\right) = (u_i \circ u_i)'(u_k \circ u_k)\end{aligned}$$

We deduce, using independence and elementary properties of inner products that

$$\mathbf{E}\left(\left[Y_i'\Sigma\text{diag}(\Sigma Y_k Y_i'\Sigma)\Sigma Y_k\right]^2\right) \leq \mathbf{E}\left(\|u_i \circ u_i\|_2^2\right)\mathbf{E}\left(\|u_k \circ u_k\|_2^2\right) \; .$$

Note that to arrive at Equation (1), we studied expressions similar to $\mathbf{E}\left(\|u_i \circ u_i\|_2^2\right)$. So we can conclude that

$$\mathbf{E}\left(\left[Y_i'\Sigma \mathrm{diag}(\Sigma Y_k Y_i'\Sigma)\Sigma Y_k\right]^2\right) \leq \left\{(3+|\mu_4-3|)\lambda_1(\Sigma^2)\mathrm{trace}\left(\Sigma^2\right)\right\}^2 \tag{3}$$

With our assumptions, the terms (1), (2) and (3) are $\mathrm{O}(p^2)$. Note that in the computation of the trace, there are $\mathrm{O}(n^4)$ such terms. Finally, note that the expectation of interest to us corresponds to the sum of the three quadratic terms divided by p^8. So the total contribution of these terms is in expectation $\mathrm{O}(p^{-2})$. This takes care of the contribution of the terms involving four different indices, as it shows that

$$0 \leq \mathbf{E}\left(\sum_{i\neq j\neq k\neq l} \widetilde{W}_{i,j}\widetilde{W}_{j,k}\widetilde{W}_{k,l}\widetilde{W}_{l,i}\right) = \mathrm{O}(p^{-2}) \ .$$

Let us now focus on the other terms. Note that because $\widetilde{W}_{i,i}=0$, terms involving 3 different indices with a non-zero contribution are necessarily of the form $(\widetilde{W}_{i,j})^2(\widetilde{W}_{i,k})^2$, since terms with a cycle of length 3 all involve a term of the form $\widetilde{W}_{i,i}$ and hence contribute 0. Let us now focus on those terms, assuming that $j\neq k$. Note that we have $\mathrm{O}(n^3)$ such terms, and that it is enough to focus on the $W_{i,j}^2 W_{i,k}^2$, since the contribution of the other terms is, in expectation, of order $1/p^4$, and because we have only n^3 terms in the sum, this extra contribution is asymptotically zero. Now, we clearly have $\mathbf{E}\left(W_{i,j}^2 W_{i,k}^2|Y_i\right) = \left[\mathbf{E}\left(W_{i,j}^2|Y_i\right)\right]^2$, by conditional independence of the two terms. The computation of $\mathbf{E}\left(W_{i,j}^2|Y_i\right)$ is similar to the ones we have made above, and we have

$$p^4\mathbf{E}\left(W_{i,j}^2|Y_i\right) = 2(Y_i'\Sigma^2 Y_i)^2 + (\mu_4-3)Y_i'\Sigma\mathrm{diag}(\Sigma Y_i Y_i'\Sigma)\Sigma Y_i + (\mathrm{trace}\left(\Sigma Y_i Y_i'\Sigma\right))^2 \ .$$

Using the fact that $\mathcal{K}_i = \Sigma Y_i Y_i'\Sigma$ is positive semidefinite, and hence its diagonal entries are non-negative, we have $\mathrm{trace}\left(\mathcal{K}_i \circ \mathcal{K}_i\right) \leq (\mathrm{trace}\left(\mathcal{K}_i\right))^2$, we conclude that

$$p^4\mathbf{E}\left(W_{i,j}^2|Y_i\right) \leq (3+|\kappa_4-3|)(Y_i'\Sigma^2 Y_i)^2 \leq (3+|\kappa_4-3|)\sigma_1(\Sigma)^4\|Y_i\|_2^4 \ .$$

Hence,

$$\mathbf{E}\left(W_{i,j}^2 W_{i,k}^2\right) \leq \frac{1}{p^8}(3+|\kappa_4-3|)^2\sigma_1(\Sigma)^8\|Y_i\|_2^8 \ .$$

Now, the application F which takes a vector and returns its Euclidian norm is trivially a convex 1-Lipschitz function, with respect to Euclidian norm. Because the entries of Y_i are bounded by B_p, we see that, according to Corollary 4.10 in Ledoux (2001), $\|Y_i\|_2$ satisfies a concentration inequality, namely $P(|\|Y_i\|_2 - m_F| > r) \leq 4\exp(-r^2/16B_p^2)$, where m_F is a median of F. A simple integration (see for instance the proof of Proposition 1.9 in Ledoux (2001), and change the power from 2 to 8) then shows that

$$\mathbf{E}\left(|\|Y_i\|_2 - m_F|^8\right) = \mathrm{O}(B_p^8) \ .$$

Note, we know, according to Proposition 1.9 in Ledoux (2001), that if μ_F is the mean of F, μ_F exists and $|m_F-\mu_F| = \mathrm{O}(B_p)$. Since $\mu_F^2 \leq \mu_{F^2} = \mathbf{E}\left(\|Y_i\|_2^2\right) = p$, we conclude that, if C denotes a generic constant that may change from display to display,

$$\begin{aligned}\mathbf{E}\left(\|Y_i\|_2^8\right) &\leq \mathbf{E}\left(|\|Y_i\|_2 - m_F + m_F|^8\right) \leq 2^7(\mathbf{E}\left(|\|Y_i\|_2 - m_F|^8\right) + m_F^8) \\ &\leq C(\mathbf{E}\left(|\|Y_i\|_2 - m_F|^8\right) + |m_F-\mu_F|^8 + \mu_F^8 \leq C(B_p^8 + p^4)\end{aligned}$$

Now, our original assumption about the number of moments of the random variables of interest imply that $B_p = \mathrm{O}(p^{1/2-\delta})$. Consequently,

$$\mathbf{E}\left(\|Y_i\|^8\right) = \mathrm{O}(p^4)$$

Therefore,

$$\mathbf{E}\left(W_{i,j}^2 W_{i,k}^2\right) = \mathrm{O}(p^{-4})$$

and

$$\sum_i \sum_{j\neq i, k\neq i, j\neq k} \mathbf{E}\left(W_{i,j}^2 W_{i,k}^2\right) = \mathrm{O}(p^{-1}) \; .$$

The last terms we have to focus on to control $\mathbf{E}\left(\text{trace}\left(\widetilde{W}^4\right)\right)$ are of the form $\widetilde{W}_{i,j}^4$. Note that we have n^2 terms like this. Since by convexity, $(a+b)^4 \leq 8(a^4+b^4)$, we see that it is enough to understand the contribution of $W_{i,j}^4$ to show that $\sum_{i,j} \mathbf{E}\left(\widetilde{W}_{i,j}^4\right)$ tends to zero. Now, let us call for a moment $v = \Sigma Y_i$ and $u = Y_j$. The quantity of interest to us is basically of the form $\mathbf{E}\left((u'v)^8\right)$. Let us do computations conditional on v. We note that since the entries of u are independent and have mean 0, in the expansion of $(u'v)^8$, the only terms that will contribute a non-zero quantity to the expectation have entries of u raised to a power greater than 2. We can decompose the sum representing $\mathbf{E}\left((u'v)^8|v\right)$ into subterms, according to what powers of the terms are involved. There are 6 terms: (2,2,2,2) (i.e all terms are raised to the power 2), (3,3,2) (i.e two terms are raised to the power 3, and one to the power 2), (4,2,2), (4,4), (5,3), (6,2) and (8). For instance the subterm corresponding to (2,2,2,2) is, before taking expectations,

$$\sum_{i_1\neq i_2\neq i_3\neq i_4} u_{i_1}^2 u_{i_2}^2 u_{i_3}^2 u_{i_4}^2 (v_{i_1} v_{i_2} v_{i_3} v_{i_4})^2 \; .$$

After taking expectations conditional on v, we see that it is obviously non-negative and contributes

$$(\sigma^2)^4 \sum_{i_1\neq i_2\neq i_3\neq i_4} (v_{i_1} v_{i_2} v_{i_3} v_{i_4})^2 \leq (\sum v_i^2)^4 = (Y_i' \Sigma^2 Y_i)^4 \leq \sigma_1(\Sigma)^8 \|Y_i\|_2^8 \; .$$

Note that we just saw that $\mathbf{E}\left(\|Y_i\|_2^8\right) = \mathrm{O}(p^4)$ in our context. Similarly, the term $(3,3,2)$ will contribute

$$\mu_3^2 \sigma^2 \sum_{i_1\neq i_2\neq i_3} v_{i_1}^3 v_{i_2}^3 v_{i_3}^2 \; .$$

In absolute value, this term is less than

$$\mu_3^2 \sigma^2 (\sum |v_i|^3)^2 (\sum v_i^2) \; .$$

Now, note that if z is such that $\|_{\|} 2 z_2 = 1$, we have, for $p \geq 2$, $\sum |z_i|^p \leq \sum z_i^2 = 1$. Applied to $z = v/\|v\|_2$, we conclude that $\sum |v_i|^p \leq \|v\|_2^p$. Consequently, the term (3,3,2) contributes in absolute value less than

$$\mu_3^2 \sigma^2 \|v\|_2^8 \; .$$

The same analysis can be repeated for all the other terms, which are all found to be less than, $\|v\|_2^8$ times the moments of u involved. Because we have assumed that our original random variables had $4+\epsilon$ moments, the moments of order less than 4 cause no problem. The moments of order higher than 4, say $4+k$, can be bounded by $\mu_4 B_p^k$. Consequently, we see that

$$\mathbf{E}\left(W_{i,j}^4\right) = \mathbf{E}\left(\mathbf{E}\left(W_{i,j}^4 | Y_i\right)\right) \leq C B_p^4 \mathbf{E}\left(\frac{\|Y_i\|^8}{p^8}\right) = \mathrm{O}(B_p^4/p^4) = \mathrm{O}(p^{-(2+4\delta)}) \; .$$

Since we have n^2 such terms, we see that

$$\sum_{i,j} \mathbf{E}\left(W_{i,j}^4\right) \to 0 \text{ as } p \to \infty \; .$$

We have therefore established control of the second order term and seen that the largest singular value of $\widetilde{W}$ goes to 0 in probability, using Chebyshev's inequality. Note that we have also shown that the operator norm of W is bounded in probability and that

$$|||W - \frac{\text{trace}\left(\Sigma^2\right)}{p^2}(11' - \text{Id})|||_2 \to 0 \text{ in probability.}$$

• **Control of the third order term**

We note that the third order term is of the form $f^{(3)}(\xi_{i,j})\frac{X_i'X_j}{p}W_{i,j}$. We first make the remark that if M is a symmetric matrix with non-negative entries, and E is a symmetric matrix such that $\max_{i,j}|E_{i,j}|=\zeta$, then

$$\sigma_1(E\circ M)\leq \zeta\sigma_1(M)\;.$$

As a matter of fact, since the matrices are symmetric,

$$\sigma_1(E\circ M)=\lim_{k\to\infty}(\text{trace}\left((E\circ M)^{2k}\right))^{1/2k}\;.$$

Now note that

$$|\text{trace}\left((E\circ M)^{2k}\right)|\leq \zeta^{2k}\text{trace}\left(M^{2k}\right)\;,$$

by upper bounding each term in the expansion of trace $\left((E\circ M)^{2k}\right)$ by ζ^{2k} times the corresponding term involving only the entries of M, which are all non-negative. Now summing all the terms involving the entries of M only gives trace $\left(M^{2k}\right)$. This shows the result concerning $\sigma_1(E\circ M)$.

So all we have to show is that $\max_{i\neq j}|X_i'X_j/p|$ goes to 0. We are going to make use of Lemma A-3, p.26 in the Appendix. In our setting, we have $B_p=p^{1/2-\delta}$, or $2/m=1/2-\delta$. The Lemma hence implies, for instance, that

$$\max_{i\neq j}|X_i'X_j/p|\leq p^{-\delta}\log(p)\;\;\text{a.s}\;.$$

So $\max_{i\neq j}|X_i'X_j/p|\to 0$ a.s . Note that this implies that $\max_{i\neq j}|\xi_{i,j}|\to 0$ a.s . Since we have assumed that $f^{(3)}$ exists and is continuous and hence bounded in a neighborhood of 0, we conclude that

$$\max_{i,j}|f^{(3)}(\xi_{i,j})X_i'X_j/p|=\text{o}(p^{-\delta/2})\;\;\text{a.s}\;.$$

If we call E the matrix with entry $E_{i,j}=f^{(3)}(\xi_{i,j})X_i'X_j/p$ off-the diagonal and 0 on the diagonal, we see that E satisfies the conditions put forth in our discussion earlier in this section and we conclude that

$$|||E\circ W|||_2\leq\max_{i,j}|E_{i,j}|\;|||W|||_2=\text{o}(p^{-\delta/2})\;\;\text{a.s}\;\;.$$

Hence, the operator norm of the third order term goes to 0 almost surely. (To maybe clarify our arguments, let us repeat that we analyzed the second order term by replacing the Y_i's by, in the notation of the truncation and centralization discussion, U_i. Let us call $W_U=S_U\circ S_U$, again using notation introduced in the truncation and centralization discussion. As we saw, $|||W-W_U|||_2\to 0$ a.s , so showing, as we did, that $|||W_U|||_2$ remains bounded (a.s) implies that $|||W|||_2$ does, too, and this is the only thing we need in our argument showing the control of the third order term.)

B) Control of the diagonal term The proof here is divided into two parts. First, we show that the error term coming from the first order expansion of the diagonal is easily controlled. Then we show that the terms added when replacing the off-diagonal matrix by $XX'/p+\text{trace}\left(\Sigma^2\right)/p^2 11'$ can also be controlled. Recall the notation $\tau=\text{trace}\left(\Sigma\right)/p$.

• **Errors induced by diagonal approximation**

Note that Lemma A-3 guarantees that for all i, $|\xi_{i,i}-\tau|\leq p^{-\delta/2}$, a.s. Because we have assumed that f' is continuous and hence bounded in a neighborhood of τ, we conclude that $f'(\xi_{i,i})$ is uniformly bounded in p. Now Lemma A-3 also guarantees that

$$\max_i\left|\frac{\|X_i\|_2^2}{p}-\tau\right|\leq p^{-\delta}\;\;\text{a.s}\;\;.$$

Hence, the diagonal matrix with entries $f(\|X_i\|_2^2/p)$ can be approximated consistenly in operator norm by $f(\tau)\text{Id}$.

•**Errors induced by off-diagonal approximation**

When we replace the off-diagonal matrix by $f'(0)XX'/p+[f(0)+f''(0)\text{trace}\left(\Sigma^2\right)/2p^2]11'$, we add a diagonal matrix with (i,i) entry $f(0)+f'(0)\|X_i\|_2^2/p+f''(0)\text{trace}\left(\Sigma^2\right)/2p^2$, which we need to subtract eventually. We note that $0\leq\text{trace}\left(\Sigma^2\right)/p^2\leq\sigma_1^2(\Sigma)/p\to 0$ when $\sigma_1(\Sigma)$ remains bounded in p. So this term does not create any problem. Now, we just saw that the diagonal matrix with entries $\|X_i\|_2^2/p$ can be consistently approximated in operator norm by $(\text{trace}\,(\Sigma)\,/p)\,\text{Id}$. So the diagonal matrix with (i,i) entry $f(0)+f'(0)\|X_i\|_2^2/p+f''(0)\text{trace}\left(\Sigma^2\right)/2p^2$ can be approximated consistently in operator norm by $(f(0)+f'(0)\text{trace}\,(\Sigma)\,/p)\text{Id}$.

This finishes the proof. □

2.3 Kernel random matrices of the type $\mathbf{f}(\|\mathbf{X_i}-\mathbf{X_j}\|_2^2/\mathbf{p})$

As is to be expected, the properties of such matrices can be deduced from the study of inner product kernel matrices, with a little bit of extra work. We need to slightly modify the distributional assumptions under which we work, and consider the case where we have $5+\epsilon$ moments; we also need to assume that f is regular is the neighborhood of different points. Otherwise, the assumptions are the same as that of Theorem 1. We have the following theorem:

Theorem 2 (Spectrum of Euclidian distance kernel matrices). *Let us call*

$$\tau = 2\frac{\textit{trace}\,(\Sigma)}{p} \;.$$

Let us call ψ the vector with entries $\|X_i\|_2^2/p-\textit{trace}\,(\Sigma)\,/p$. Consider the kernel matrix M with entries

$$M_{i,j}=f\left(\frac{\|X_i-X_j\|_2^2}{p}\right) \;.$$

Suppose that the assumptions of Theorem 1 hold, but that conditions e) and f) are replaced by

e') The entries of Y_i have $5+\epsilon$ moments.

f') f is C^3 in a neighborhood of τ.

Then M can be approximated consistently in operator norm (and in probability) by the matrix K, defined by

$$K=f(\tau)11'+f'(\tau)\left[1\psi'+\psi1'-2\frac{XX'}{p}\right]+\frac{f''(\tau)}{2}\left[1(\psi\circ\psi)'+(\psi\circ\psi)1'+2\psi\psi'+4\frac{\textit{trace}\,(\Sigma^2)}{p^2}11'\right]+\upsilon_p\text{Id}\;,$$
$$\upsilon_p=f(0)+\tau f'(\tau)-f(\tau)\;.$$

In other words,

$$|||M-K|||_2\to 0 \quad \textit{in probability.}$$

Proof. Note that here the diagonal is just $f(0)\text{Id}$ and it will cause no trouble. The work therefore focuses on the off-diagonal matrix. In what follows, we call $\tau=2\frac{\text{trace}(\Sigma)}{p}$. Let us define

$$A_{i,j}=\frac{\|X_i\|_2^2}{p}+\frac{\|X_j\|_2^2}{p}-\tau\;,$$

and

$$S_{i,j}=\frac{X_i'X_j}{p}\;.$$

With these notations, we have, off the diagonal,

$$M_{i,j}=f(\tau)+\left[A_{i,j}-2S_{i,j}\right]f'(\tau)+\frac{1}{2}\left[A_{i,j}-2S_{i,j}\right]^2f''(\tau)+\frac{1}{6}f^{(3)}(\xi_{i,j})\left[A_{i,j}-2S_{i,j}\right]^3\;.$$

We note that the matrix A with entries $A_{i,j}$ is a rank 2 matrix. As a matter of fact, it can be written, if ψ is the vector with entries $\psi_i=\frac{\|X_i\|_2^2}{p}-\tau/2$, $A=1\psi'+\psi1'$. Using the well-known identity (see e.g Gohberg et al. (2000), Chapter 1, Theorem 3.2)

$$\det(I+uv'+vu')=\det\begin{pmatrix}1+u'v & \|u\|_2^2\\ \|v\|_2^2 & 1+u'v\end{pmatrix}\;,$$

we see immediately that the non-zero eigenvalues of A are

$$1'\psi \pm \sqrt{n}\|\psi\|_2 \ .$$

After these preliminary remarks, we are ready to start the proof per se.

• **Truncation and centralization**

Since we assume $5+\epsilon$ moments, we see, using Lemma 2.2 in Yin et al. (1988), that we can truncate the Y_i's at level $B_p = p^{2/5-\delta}$, with $\delta > 0$ and a.s not change the data matrix. We then need to centralize the vectors truncated at $p^{2/5-\delta}$. Note that because we work with $X_i - X_j = \Sigma^{1/2}(Y_i - Y_j)$ centralization creates absolutely no problem here, since it is absorbed in the difference. So in what follows we can assume without loss of generality that we are working with vectors $X_i = \Sigma^{1/2}Y_i$, where the entries of Y_i are bounded by $p^{2/5-\delta}$ and $\mathbf{E}\left(Y_i\right) = 0$. The issue of variance 1 is addressed as before, so we can assume that the entries of Y_i have variance 1.

• **Concentration of $\|\mathbf{X_i} - \mathbf{X_j}\|_2^2/\mathbf{p}$**

By plugging-in the results of Corollary A-2, with $2/m = 2/5 - \delta$, we get that

$$\max_{i\neq j}\left|\frac{\|X_i - X_j\|_2^2}{p} - 2\frac{\mathrm{trace}\left(\Sigma\right)}{p}\right| \leq \log(p)p^{-1/10-\delta} \ .$$

Also, using the result of Lemma A-3, we have

$$\max_i |\psi_i| = \max_i \left|\frac{\|X_i\|_2^2}{p} - \frac{\mathrm{trace}\left(\Sigma\right)}{p}\right| \leq \log(p)p^{-1/10-\delta} \ .$$

Note that, as explained in the proof of Lemma A-3, these results are true whether we work with Y_i or their truncated and centralized version.

• **Control of the second order term**

Let us call T the matrix with 0 on the diagonal and off-diagonal entries $T_{i,j} = (A_{i,j} - 2S_{i,j})^2$. In other words, if $i \neq j$,

$$T_{i,j} = \left(\frac{\|X_i - X_j\|_2^2 - 2\mathrm{trace}\left(\Sigma\right)}{p}\right)^2 \ .$$

We simply write $(A_{i,j} - 2S_{i,j})^2 = A_{i,j}^2 - 4A_{i,j}S_{i,j} + 4S_{i,j}^2$. In the notation of the proof of Theorem 1, the matrix with entries $S_{i,j}^2$ off the diagonal and 0 on the diagonal is what we called W. We have already shown that

$$|||W - \frac{\mathrm{trace}\left(\Sigma^2\right)}{p^2}(11' - \mathrm{Id})|||_2 \rightarrow \ 0 \text{ in probability} \ .$$

Now, let us focus on the term $A_{i,j}S_{i,j}$. Let us call H the matrix with

$$H_{i,j} = (1 - \delta_{i,j})A_{i,j}S_{i,j} \ .$$

Let us denote by $\widetilde{S}$ the matrix with off-diagonal entries $S_{i,j}$ and 0 on the diagonal. Now note that $A_{i,j} = \psi_i + \psi_j$. Therefore, we have, if $\mathrm{diag}(\psi)$ is the diagonal matrix with (i,i) entry ψ_i,

$$H = \widetilde{S}\mathrm{diag}(\psi) + \mathrm{diag}(\psi)\widetilde{S} \ .$$

We just saw that under our assumptions, $\max_i |\psi_i| \rightarrow 0$ a.s . Because for any $n \times n$ matrices L_1, L_2, $|||L_1L_2|||_2 \leq |||L_1|||_2|||L_2|||_2$, we see that to show that $|||H|||_2$ goes to 0, we just need to show that $|||\widetilde{S}|||_2$ remains bounded. If we call $S = XX'/p$, we have

$$\widetilde{S} = S - \mathrm{diag}(S) \ .$$

Now we clearly have, $|||S|||_2 \leq |||\Sigma|||_2|||Y'Y/p|||_2$. We know from Yin et al. (1988), that $|||Y'Y/p|||_2 \rightarrow \sigma^2(1+\sqrt{n/p})^2$, a.s. Under our assumptions on n and p, this is bounded. Now

$$\mathrm{diag}(S) = \mathrm{diag}(\psi) + \frac{\mathrm{trace}\left(\Sigma\right)}{p}\mathrm{Id} \ ,$$

so our concentration results once again imply that $|||\text{diag}(S)|||_2 \leq \text{trace}\left(\Sigma\right)/p + \eta$ a.s , for any $\eta > 0$. Because $|||\cdot|||_2$ is subadditive, we finally conclude that

$$|||\widetilde{S}|||_2 \text{ is bounded a.s .}$$

Therefore,

$$|||H|||_2 \rightarrow 0 \text{ a.s .}$$

Putting together all these results, we see that we have shown that

$$|||T - (A \circ A - diag(A \circ A)) - 4\frac{\text{trace}\left(\Sigma^2\right)}{p^2}(11' - \text{Id})|||_2 \rightarrow 0 \text{ in probability .}$$

- **Control of the third order term**

The third order term is the matrix L with 0 on the diagonal and off-diagonal entries

$$L_{i,j} = \frac{f^{(3)}(\xi_{i,j})}{6}\left(\frac{\|X_i - X_j\|_2^2 - 2\text{trace}\left(\Sigma\right)}{p}\right)^3 \triangleq E \circ T ,$$

where T was the matrix investigated in the control of the second order term. On the other hand E is the matrix with entries

$$E_{i,j} = (1 - \delta_{i,j})\frac{f^{(3)}(\xi_{i,j})}{6}\left(\frac{\|X_i - X_j\|_2^2 - 2\text{trace}\left(\Sigma\right)}{p}\right) .$$

We have already seen that through concentration, we have

$$\max_{i \neq j}\left|\frac{\|X_i - X_j\|_2^2}{p} - \frac{2\text{trace}\left(\Sigma\right)}{p}\right| \leq \log(p)p^{-1/10-\delta} \quad \text{a.s .}$$

This naturally implies that

$$\max_{i \neq j}\left|\xi_{i,j} - \frac{2\text{trace}\left(\Sigma\right)}{p}\right| \leq \log(p)p^{-1/10-\delta} \text{ a.s .}$$

So if $f^{(3)}$ is bounded in a neighborhood of τ, we see that with high-probability so is $\max_{i \neq j}|f^{(3)}(\xi_{i,j})|$. Therefore,

$$\max_{i \neq j}|E_{i,j}| \leq K \log(p)p^{-1/10-\delta} .$$

We are now in position to apply the Hadamard product argument we used for the control of the third order term in the proof of Theorem 1. To show that the third order term tends in operator norm to 0, we hence just need to control $|||T|||_2$ remains small compared to the bound we just gave on $\max_{i,j}|E_{i,j}|$. Of course, this is equivalent to showing that the matrix that approximates T has the same property in operator norm.

Clearly, because $\sigma_1(\Sigma)$ stays bounded, $\text{trace}\left(\Sigma^2\right)/p$ stays bounded and so does $|||\text{trace}\left(\Sigma^2\right)/p^2(11' - \text{Id})|||_2$. So we just have to focus on $A \circ A - \text{diag}(A \circ A)$. Recall that $A_{i,i} = 2(\|X_i\|_2^2/p - \text{trace}\left(\Sigma\right)/p)$, and so $A_{i,i} = 2\psi_i$. We have already seen that our concentration arguments implies that $\max_i|\psi_i| \rightarrow 0$ a.s . So $|||\text{diag}(A \circ A)|||_2 = \max_i \psi_i^2$ goes to 0 a.s . Now,

$$A = 1\psi' + \psi 1' ,$$

and hence, elementary Hadamard product computations (relying on $ab' \circ uv' = (a \circ u)(b \circ v)'$) give that

$$A \circ A = 1(\psi \circ \psi)' + 2\psi\psi' + (\psi \circ \psi)1' .$$

Therefore,

$$|||A \circ A|||_2 \leq 2(\sqrt{n}\|\psi \circ \psi\|_2 + \|\psi\|_2^2) .$$

Using Lemma A-1, and in particular Equation (A-2), we see that

$$\mathbf{E}\left(\psi_i^2\right) = 2\sigma^4\frac{\text{trace}\left(\Sigma^2\right)}{p^2} + (\kappa_4 - 3\sigma^4)\frac{\text{trace}\left(\Sigma \circ \Sigma\right)}{p^2} ,$$

and therefore, $\mathbf{E}\left(\|\psi\|_2^2\right)$ remains bounded. On the other hand, using Lemma 2.7 of Bai and Silverstein (1998), we see that if we have $5+\epsilon$ moments,

$$\mathbf{E}\left(\psi_i^4\right) \leq C\left(\frac{(\mu_4 \text{trace}\left(\Sigma^2\right))^2}{p^4} + \mu_{5+\epsilon} B_p^{3-\epsilon}\frac{\text{trace}\left(\Sigma^4\right)}{p^4}\right) .$$

Now recall that we can take $B_p = p^{2/5-\delta}$. Therefore $n\mathbf{E}\left(\|\psi \circ \psi\|_2^2\right)$ is at most of order $B_p^{3-\epsilon}/p$. We conclude that

$$P(|||A \circ A|||_2 > \log(p)\sqrt{B_p^{3-\epsilon}/p}) \to 0 .$$

Note that this implies that

$$P(|||T|||_2 > \log(p)\sqrt{B_p^{3-\epsilon}/p}) \to 0 .$$

Now, note that the third order term is of the form $E \circ T$. Because we have assumed that we have $5+\epsilon$ moments, we have already seen that our concentration results imply that

$$\max_{i \neq j} |E_{i,j}| = \mathrm{O}\left(\log(p)\sqrt{\frac{B_p^2}{p}}\right) = \mathrm{O}\left(\log(p)p^{-1/10-\delta}\right) \quad \text{a.s} .$$

Using the fact that T has positive entries and therefore (see the proof of Theorem 1) $|||E \circ T|||_2 \leq \max_{i,j}|E_{i,j}|\ |||T|||_2$, we conclude that with high-probability,

$$|||E \circ T|||_2 = \mathrm{O}\left((\log(p))^2\sqrt{\frac{B_p^{5-\epsilon}}{p^2}}\right) = \mathrm{O}\left((\log(p))^2 p^{-\delta'}\right) \quad \text{where } \delta' > 0 .$$

Hence,

$$|||E \circ T|||_2 \to 0 \text{ in probability} .$$

- **Adjustment of the diagonal**

To obtain the compact form of the approximation announced in the theorem, we need to include diagonal terms that are not present in the matrices resulting from the Taylor expansion. Here, we show that the corresponding matrices are easily controlled in operator norm.

When we replace the zeroth and first order terms by

$$f(\tau)11' + f'(\tau)\left[1\psi' + \psi 1' - 2\frac{XX'}{p}\right] ,$$

we add to the diagonal the term $f(\tau) + f'(\tau)(2\psi_i - 2\|X_i\|_2^2/p) = f(\tau) - 2f'(\tau)\frac{\text{trace}(\Sigma)}{p}$. In the end, we need to subtract it.

When we replace the second order term by $\frac{1}{2}f''(\tau)[1(\psi\circ\psi)' + 2\psi\psi' + (\psi\circ\psi)1' + 4\frac{\text{trace}(\Sigma^2)}{p^2}11']$, we add to the diagonal the diagonal matrix with (i,i) entry

$$2f''(\tau)[\psi_i^2 + \frac{\text{trace}\left(\Sigma^2\right)}{p^2}] .$$

With our assumptions, $\max_i |\psi_i| \to 0$ a.s and $\frac{\text{trace}(\Sigma^2)}{p}$ remains bounded, so the added diagonal matrix has operator norm converging to 0 a.s . We conclude that we do not need to add it to the correction in the diagonal of the matrix approximating our kernel matrix. □

An interpretation of the proofs of Theorems 1 and 2 is that they rely on a local "multiscale" approximation of the original matrix. However, globally, there is a bit of a mixture between the scales which creates the difficulties we had to deal with to control the second order term.

2.3.1 A note on the Gaussian Kernel

The Gaussian kernel corresponds to $f(x) = \exp(-\gamma x)$ in the notation of Theorem 2. We would like to discuss it a bit more because of its widespread use in applications.

The result of Theorem 2 gives accurate limiting eigenvalue information for the case where we renormalize the distances by the dimension, which seems to be, implicitly or explicitly what is often done in practice.

However, it is possible that information about the non-renormalized might be of interest, too in some situations. Let us assume now that trace (Σ) grows to infinity at least as fast as $p^{1/2+2/m+\delta}$, where $\delta > 0$ is such that $1/2 + 2/m + \delta < 1$, which is possible since $m \geq 5 + \epsilon$ here. We of course still assume that its largest singular value, $\sigma_1(\Sigma)$ remains bounded. Then, Corollary A-2 guarantees that

$$\min_{i \neq j} \frac{\|X_i - X_j\|_2^2}{p} > \frac{\text{trace}\,(\Sigma)}{p} \quad \text{a.s} \ .$$

Hence,

$$\max_{i \neq j} \exp(-\|X_i - X_j\|_2^2) \leq \exp(-\text{trace}\,(\Sigma)) \leq \exp(-p^{1/2+2/m+\delta}) \ \ \text{a.s} \ .$$

Hence, in this case, if M is our kernel matrix with entries $\exp(-\|X_i - X_j\|_2^2)$, we have,

$$|||M - \text{Id}|||_2 \leq n \exp(-p^{1/2+2/m+\delta}) \ , \ \text{a.s} \ \ ,$$

and the upper bound tends to zero extremely fast.

2.4 More general models

In this subsection, we consider more general models that the ones considered above. In particular, we will here focus on data models for which the vectors X_i satisfy a so-called dimension-free concentration inequality. As was shown in El Karoui (2007a), under these conditions, the Marčenko-Pastur equation holds (as well as generalized versions of it). Note that these models are more general than the one considered above (the proofs in the Appendix illustrate why the standard random matrix models can be considered as subcases of this more general class of matrices), and can describe various interesting objects, like vectors with certain log-concave distributions, or vectors sampled in a uniform manner from certain Riemannian submanifolds of $\mathbb{R}^p$, endowed with the canonical Riemannian metric inherited from $\mathbb{R}^p$. We are now ready to state the theorem

Theorem 3. *Suppose the vectors $X_i \in \mathbb{R}^p$ are i.i.d and have the property that for 1-Lipschitz functions (with respect to Euclidian norm),*

$$P(|F - m_F| > r) \leq C \exp(-cr^2) \ ,$$

where C is independent of p and c may depend on p, but is required to satisfy $c \geq p^{-1/2+\epsilon}$.

Consider the kernel random matrix M with $M_{i,j} = f(X_i' X_j/p)$. Call Σ the covariance matrix of the X_i's and assume that $\sigma_1(\Sigma)$ stays bounded and trace $(\Sigma)/p$ has a limit. Suppose that f is a real valued function, which is C^2 around 0 and C^1 around trace $(\Sigma)/p$.

The spectrum of this matrix is asymptotically non-random and has, a.s, the same limiting spectral distribution as that of

$$\widetilde{M} = f(0)11' + f'(0)\frac{XX'}{p} + \upsilon_p \text{Id}_n \ ,$$

where $\upsilon_p = f(\frac{trace(\Sigma)}{p}) - f(0) - f'(0)\frac{trace(\Sigma)}{p}$.

We note that the term $f(0)11'$ does not affect the limiting spectral distribution of $\widetilde{M}$, since finite rank perturbations do not have any effect on limiting spectral distributions (see e.g Bai (1999), Lemma 2.2). Therefore, it could be removed from the approximating matrix, but since it will clearly be present in numerical work and simulations, we chose to leave it in our approximation.

The first step in the proof is the following lemma.

Lemma 1. *Suppose K_n is an $n \times n$ symmetric matrix with a limiting spectral distribution. Suppose M_n is an $n \times n$ symmetric matrix.*

1. *Suppose M_n is such that $\|M_n - K_n\|_F = \mathrm{o}(\sqrt{n})$. Then, M_n and K_n have the same limiting spectral distribution.*

2. *Suppose M_n is such that $|||M_n - K_n|||_2 \to 0$. Then, M_n and K_n have the same limiting spectral distribution.*

Proof of Lemma. We call St_{K_n} and St_{M_n} the Stieltjes transforms of the spectral distributions of these two matrices. Suppose $z = u + iv$. Let us call $l_i(M_n)$ the i-th largest eigenvalue of M_n.

We first focus on the Frobenius norm part of the lemma. We have

$$|\mathrm{St}_{K_n}(z) - \mathrm{St}_{M_n}(z)| = \frac{1}{n}\left|\sum_{i=1}^{n} \frac{1}{l_i(K_n) - z} - \frac{1}{l_i(M_n) - z}\right| \leq \frac{1}{n}\sum_{i=1}^{n} \frac{|l_i(M_n) - l_i(K_n)|}{v^2}.$$

Now, by Holder's inequality, $\sum |l_i(M_n) - l_i(K_n)| \leq \sqrt{n}\sqrt{\sum |l_i(M_n) - l_i(K_n)|^2}$. Now using Lidskii's theorem (i.e the fact that, since M_n and K_n are hermitian, the vector with entries $l_i(M_n) - l_i(K_n)$ is majorized by the vector $l_i(M_n - K_n))$), with, in the notation of Bhatia (1997), Theorem III.4.4 $\Phi(x) = x^2$, we have

$$\sum |l_i(M_n) - l_i(K_n)|^2 \leq \sum l_i^2(M_n - K_n) = \|M_n - K_n\|_F^2 \ .$$

We conclude that

$$|\mathrm{St}_{K_n}(z) - \mathrm{St}_{M_n}(z)| \leq \frac{\|M_n - F_n\|_F}{\sqrt{n}v^2} \ .$$

Under the assumptions of the lemma, we therefore have

$$|\mathrm{St}_{K_n}(z) - \mathrm{St}_{M_n}(z)| \to 0 \ .$$

Therefore the Stieltjes transform of the spectral distribution of M_n converges pointwise to the Stieltjes transform of the limiting spectral distribution of K_n. Hence, the spectral distribution of M_n converges in distribution to the limiting spectral distribution of K_n.

Let us now turn to the operator norm part of the lemma. By the same computations as above, we have

$$|\mathrm{St}_{K_n}(z) - \mathrm{St}_{M_n}(z)| = \frac{1}{n}\left|\sum_{i=1}^{n} \frac{1}{l_i(K_n) - z} - \frac{1}{l_i(M_n) - z}\right| \leq \frac{1}{n}\sum_{i=1}^{n} \frac{|l_i(M_n) - l_i(K_n)|}{v^2} \leq \frac{|||M_n - K_n|||_2}{v^2} \ .$$

Hence if $|||M_n - K_n|||_2 \to 0$, it is clear that the two Stieljtes transforms are asymptotically equal, and the conclusion follows. □

We now turn to the proof of the theorem.

Proof of theorem. For the weaker statement required for the proof of Theorem 3, we will show that in the δ-method we need to keep only the first term of the expansion, as long as f has a second derivative that is bounded in a neighborhood of 0, and a first derivative that is bounded in a neighborhood of $\mathrm{trace}\,(\Sigma)\,/p$. In other words, we will split the problem into two parts: off the diagonal, we write

$$f\left(\frac{X_i'X_j}{p}\right) = f(0) + f'(0)\frac{X_i'X_j}{p} + \frac{f''(\xi_{i,j})}{2}\left(\frac{X_i'X_j}{p}\right)^2 \ ;$$

on the diagonal, we write

$$f\left(\frac{X_i'X_i}{p}\right) = f\left(\frac{\mathrm{trace}\,(\Sigma)}{p}\right) + f'\left(\xi_{i,i}\right)\left(\frac{X_i'X_i}{p} - \frac{\mathrm{trace}\,(\Sigma)}{p}\right) \ .$$

• **Control of the off-diagonal error matrix** The strategy is going to be to control the Frobenius norm of the matrix

$$W_{i,j} = \begin{cases} \left(\frac{X_i'X_j}{p}\right)^2 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \ .$$

According to Lemma 1, it is enough for our needs to show that the Frobenius norm of this matrix is $\mathrm{o}(\sqrt{n})$ a.s to have the result we wish. Hence, the result will be shown, if we can for instance show that

$$\max_{i,j} W_{i,j} \leq p^{-(1/2+\epsilon)}(\log(p))^{1+\delta} \text{ a.s } .$$

Now Lemma A-4 gives for instance,

$$\max_{i\neq j} \left|\frac{X_i' X_j}{p}\right| \leq (pc(p))^{-1/2} \log(p) \text{ a.s } .$$

Therefore, with our assumption on $c(p)$, we have

$$\max_{i,j} W_{i,j} \leq p^{-(1/2+\epsilon)}(\log(p))^{2} \text{ a.s } .$$

Now, $\|W\|_F \leq n \max_{i,j} |W_{i,j}|$, so we conclude that in this situation, with our assumptions that $n \asymp p$,

$$\|W\|_F = \mathrm{o}(\sqrt{n}) \ a.s$$

(We note that given a sequence model of matrices, the Borel-Cantelli lemma would apply and give an almost sure statement in the above expression.) Since $\|W\|_F^2 \leq n^2 \max_{i,j} W_{i,j}^2$, we conclude that with very high-probability,

$$\|W\|_F = \mathrm{O}(p^{1/2-\epsilon/2}) \ .$$

Now let us focus on

$$\widetilde{W}_{i,j} = f''(\xi_{i,j}) W_{i,j} \ ,$$

where $\xi_{i,j}$ is between 0 and $X_i'X_j/p$. We just saw that with very high-probability, this latter quantity was less than $p^{-(1/4+\epsilon/2)}$, if $c \geq p^{-1/2+\epsilon}$, therefore is f'' is bounded by K in a neighborhood of 0, we have, with very high probability that

$$\|\widetilde{W}\|_F \leq K \|W\|_F = \mathrm{o}(\sqrt{n}) \ .$$

- **Control of the diagonal matrix**

We first note that when we replace the off-diagonal matrix by $f(0)11' + f'(0)XX'/p$, we add to the diagonal certain terms that we need to subtract eventually.

Hence, our strategy here is to show that we can approximate (in operator norm) the diagonal matrix D with entries

$$D_{i,i} = f\left(\frac{\mathrm{trace}\,(\Sigma)}{p}\right) + f'\left(\xi_{i,i}\right)\left(\frac{X_i'X_i}{p} - \frac{\mathrm{trace}\,(\Sigma)}{p}\right) - f'(0)\frac{X_i'X_i}{p} - f(0) \ ,$$

by $v_p \mathrm{Id}_p$. To do so, we just have to show that the diagonal error matrix Z, with entries

$$Z_{i,i} = \left(f'\left(\xi_{i,i}\right) - f'(0)\right)\left(\frac{X_i'X_i}{p} - \frac{\mathrm{trace}\,(\Sigma)}{p}\right)$$

goes to zero in operator norm.

As seen in Lemma A-4, if $c \geq p^{-1/2+\epsilon}$, with very high-probability,

$$\max_{i} \left|\frac{X_i'X_i}{p} - \frac{\mathrm{trace}\,(\Sigma)}{p}\right| \leq p^{-(1/4+\epsilon/2)} \ .$$

If f' is continuous and hence bounded around $\frac{\mathrm{trace}(\Sigma)}{p}$, we therefore see that the operator (or spectral) norm of Z satisfies with high-probability

$$|||Z|||_2 \leq K p^{-(1/4+\epsilon/2)} \ .$$

- **Final step**

We clearly have

$$\widetilde{M} - M = W + Z \ .$$

It is also clear that $\widetilde{M}$ has a limiting spectral distribution, satisfying, up to centering and scaling, the Marčenko-Pastur equation; this was shown in El Karoui (2007a). By Lemma 1, we see that $\widetilde{M}$ and $\widetilde{M}-Z$ have the same limiting spectral distribution, since their difference is Z and $|||Z|||_2 \to 0$. Using the same lemma, we see that M and $\widetilde{M}-Z$ have (in probability) the same limiting spectral distribution, since their difference is W and we have established that the Frobenius norm of this matrix is (in probability) $\mathrm{o}(\sqrt{n})$. Hence, M and $\widetilde{M}$ have (in probability) the same limiting spectral distribution. □

We finally treat the case of kernel matrices computed from Euclidian norms, in this more general distributional setting.

Theorem 4. *Let us call $\tau = 2trace\left(\Sigma\right)/p$, where Σ is the covariance matrix of the X_i's. Suppose that f is a real valued function, which is C^2 around τ and C^1 around 0.*

Under the assumptions of Theorem 3, the kernel matrix M with (i,j) entry

$$M_{i,j} = f\left(\frac{\|X_i - X_j\|_2^2}{p}\right)$$

has a non-random limiting spectral distribution, which is the same as that of the matrix

$$\widetilde{M} = f(\tau)11' - 2f'(\tau)\frac{XX'}{p} + \upsilon_p \mathrm{Id}_n \ ,$$

where $\upsilon_p = f(0) + \tau f'(\tau) - f(\tau)$.

We note once again that the term $f(\tau)11'$ does not affect the limiting spectral distribution of M. But we keep it for the same reasons as before.

Proof. Note that the diagonal term is simply $f(0)\mathrm{Id}$, so this term does not create any problem.

The rest of proof is similar to that of Theorem 3. In particular the control of the Frobenius norm of the second order term is done in the same way, by controlling the maximum of the off-diagonal term, using Corollary A-3 (and hence Lemma A-4).

Therefore, we only need to understand the first order term, in other words, the matrix with 0 on the diagonal and off diagonal entry

$$\begin{aligned} R_{i,j} &= \frac{\|X_i - X_j\|_2^2}{p} - \tau \\ &= \left[\frac{\|X_i\|_2^2}{p} - \frac{\mathrm{trace}\left(\Sigma\right)}{p}\right] + \left[\frac{\|X_j\|_2^2}{p} - \frac{\mathrm{trace}\left(\Sigma\right)}{p}\right] - 2\frac{X_i'X_j}{p} \end{aligned}$$

As in the proof of Theorem 2, let us call ψ the vector with i-th entry $\psi_i = \frac{\|X_i\|_2^2}{p} - \frac{\mathrm{trace}(\Sigma)}{p}$. Clearly,

$$R_{i,j} = \delta_{i,j}(1\psi' + \psi 1' - 2\frac{XX'}{p}) \ .$$

Simple computations show that

$$R - 2\frac{\mathrm{trace}\left(\Sigma\right)}{p}\mathrm{Id} = 1\psi' + \psi 1' - 2\frac{XX'}{p} \ .$$

Now, obviously, $1\psi' + \psi 1'$ is a matrix of rank at most 2. Hence, R has the same limiting spectral distribution as

$$2\frac{\mathrm{trace}\left(\Sigma\right)}{p}\mathrm{Id} - 2\frac{XX'}{p} \ ,$$

since finite rank perturbations do not affect limiting spectral distributions (see for instance Bai (1999), Lemma 2.2). This completes the proof. □

The results of Theorem 3 and Theorem 4 apply to a wide variety of distributions, and in particular ones for which the entries of the data vectors can have a fairly complicated dependence structure. For instance, they apply to the following type of distributions:

- log-concave distributions, with a density of the type $\exp(-U(x))$, with $\text{Hessian}(U(x)) \succeq c\text{Id}$, where $c > 0$. (Theorem 2.7 in Ledoux (2001).)
- For data sampled from certain Riemannian submanifolds of $\mathbb{R}^p$, the Riemannian metric at stake being the one inherited from the ambient space. The key parameter in the concentration function here is a certain type of curvature, called Ricci curvature. (See Theorem 2.4 in Ledoux (2001), and the fact that the geodesic distance on the manifold is greater, with this choice of Riemannian metric, than the Euclidian distance; this implies that Lipschitz functions with respect to the Euclidian metric are Lipschitz with respect to the geodesic distance on the manifold, with the same Lipschitz constant.)

2.5 Some consequences of the Theorems

In practice, it is often the case that slight variant of kernel random matrices are used. In particular, it is customary to center the matrices, i.e transform M so that its row sum, or column sum or both are 0. In these situations, our results still apply; the following Fact makes it clear.

Fact 1 (Centered kernel random matrices)**.** *Let H be the $n \times n$ matrix $\text{Id}_n - 11'/n$.*

1. *If the kernel random matrix M can be approximated consistently in operator norm by K, then, if $a, b \in \{0, 1\}$,*
$$H^a M H^b \textit{ can be approximated consistently in operator norm by } H^a K H^b \;.$$
2. *If the kernel random matrix M has the same limiting spectral distribution as the matrix K, then, if $a, b \in \{0, 1\}$,*
$$H^a M H^b \textit{ has the same limiting spectral distribution as } K \;.$$

A nice consequence of the first point is that the recent hard work on localizing the largest eigenvalues of sample covariance matrices (see Baik and Silverstein (2006), Paul (2007) and El Karoui (2007c)) can be transferred to kernel random matrices and used to give some information about the localization of the largest eigenvalues of HMH for instance. In the case of the results of El Karoui (2007c), Fact 2, the arguments of El Karoui (2007a), Subsection 2.2.4, show that it gives exact localization information. In other words, we can characterize the a.s limit of the largest eigenvalue of HMH (or HM or MH) fairly explicitly, provided Fact 2 in El Karoui (2007c) apply. Finally, let us mention the obvious fact that since for two square matrices A and B, AB and BA have the same eigenvalues, we see that HMH has the same eigenvalues as MH and HM, because $H^2 = H$.

Proof. The proofs are very simple. First note that H is positive semi-definite and $|||H|||_2 = 1$. Using the submultiplicativity of $||| \cdot |||_2$, we see that
$$|||H^a M H^b - H^a K H^b|||_2 \leq |||M - K|||_2 |||H^a|||_2 |||H^b|||_2 = |||M - K|||_2 \;.$$
This shows the first point of the Fact.

The second point follows from the fact that $H^a M H^b$ is a finite rank perturbation of M. Hence, using Lemma 2.2 in Bai (1999), we see that these two matrices have the same limiting spectral distribution, and since by assumption, K has the same limiting spectral distribution as M, we have the result of the second point. □

3 Conclusions

Beside the mathematical results which basically give both strong and weak approximation theorems, this study raises several statistical questions, both about the richness - or lack thereof - of models that are often studied in random matrix theory and about the effect of kernel methods in this context.

Limitations of standard random matrix models

In the study of spectral distribution of large dimensional sample covariance matrices, it has been somewhat forcefully advocated that the study should be done under the assumptions that the data are of the form $X_i = \Sigma^{1/2} Y_i$, where the entries of Y_i have finite fourth moment. At first sight, this idea is appealing, as it seems to allow a great variety of distributions and hence flexible modeling. A possible drawback however, is the assumption that the data are linear combinations of i.i.d random variables, or the necessary presence of independence in the model. This has however been recently addressed (see e.g El Karoui (2007a)) and it has been shown that one could go beyond models requiring independence in a lurking random vector which the data linearly depend on.

Data analytic consequences However, a serious limitation is still present. As the results of Lemmas A-3 and A-4 make clear, under the models for which the limiting spectral distribution of the sample covariance matrix has been shown to satisfy the Marčenko-Pastur equation, the norms of the data vectors are concentrated. More precisely, if one were to plot a histogram of $\{\|X_i\|_2^2/p\}_{i=1}^n$, this histogram would look tightly concentrated around a single value. Hence these data vectors, when properly renormalized, stay close to a sphere. Though the models are quite rich, the geometry that we can perceive by sampling n such vectors, with $n \asymp p$, is, arguably, relatively poor. These remarks should not be taken as aiming to discredit the rich and extremely interesting body of work that has emerged out of the study of such models. Their aim is just to warn possible users that in data analysis, a good first step would be to plot the histogram of $\{\|X_i\|_2^2/p\}_{i=1}^n$ and check whether it is concentrated around a single value. Similarly, one might want to plot the histogram of inner products $\{X_i' X_j/p\}$ and check that it is concentrated around 0. If this is not the case, then insights derived from random matrix theoretic studies would likely not be helpful in the data analysis.

We note however that recent random matrix work (see Boutet de Monvel et al. (1996), Burda et al. (2005), Paul and Silverstein (2007), El Karoui (2007a)) has been concerned with distributions which could be loosely speaking be called of "elliptical" type - though they are more general than what is usually called elliptical distributions in Statistics. In those settings, the data is, for instance, of the form $X_i = r_i \Sigma^{1/2} Y_i$, where r_i is a real-valued random variable, independent of Y_i. This allows the data vectors to not approximately live on spheres, and is a possible way to address the concerns we just raised. However, the characterization of the limiting expressions gets quite a bit more involved.

On kernel random matrices

Our study, motivated in part by numerical experiments we read about in the interesting Williams and Seeger (2000), has shown that in the asymptotic setting we considered, which is generally considered relevant for high-dimensional data analysis, kernel random matrices behave essentially like matrices closely connected to sample covariance matrices. This is in sharp contrast to the low dimensional setting where it was explained heuristically in Williams and Seeger (2000), and proved rigorously in Koltchinskii and Giné (2000), that the eigenvalues of kernel random matrices converged (under certain assumptions) to those of a canonically related operator. Under various assumptions on the distribution of our data, we have been able to show a strong approximation result (operator norm consistency) whose meaning is that to first order, the eigenvalues of kernel random matrices behave (up to centering and scaling) like the eigenvalues of the covariance matrix of the data. The same is true for the eigenvectors of the kernel matrix and those of the matrix XX'/p which are associated to separated eigenvalues. We have also characterized limiting spectral distributions of kernel random matrices for a broader class of distributions. This suggests that kernel methods could suffer from the same problems that affect linear statistical methods, such as Principal Component Analysis, in high-dimensions.

Our study also permits the transfer of some recent random matrix results concerning large dimensional sample covariance matrices to kernel random matrices.

APPENDIX

In this Appendix, we collect a few useful results that are needed in the proof of our Theorems, and whose content we thought would be more accessible if they were separated from the main proofs.

Some useful results

We have the following elementary facts.

Lemma A-1. *Suppose Y is a vector with i.i.d entries, and mean 0. Call its entries y_i. Suppose $\mathbf{E}\left(y_i^2\right) = \sigma^2$ and $\mathbf{E}\left(y_i^4\right) = \mu_4$. Then, if M is a deterministic matrix,*

$$\mathbf{E}\left(YY'MYY'\right) = \sigma^4(M+M') + (\mu_4 - 3\sigma^4)\,diag(M) + \sigma^4 trace\,(M)\,\mathrm{Id}\;. \tag{A-1}$$

Further, we have $(Y'MY)^2 = trace\,(MYY'MYY')$, and

$$\mathbf{E}\left(trace\left(MYY'MYY'\right)\right) = \sigma^4 trace\left(M^2 + MM'\right) + \sigma^4(trace\,(M))^2 + (\mu_4 - 3\sigma^4) trace\,(M \circ M)\;. \tag{A-2}$$

Here $diag(M)$ denotes the matrix consisting of the diagonal of the matrix M and 0 off the diagonal. The symbol $\circ$ denotes Hadamard multiplication between matrices.

Proof. Let us call $R = YY'MYY'$. The proof of the first part is elementary and consists merely in writing the (i,j)-th entry of the corresponding matrix. As a matter of fact, we have

$$R_{i,j} = y_i y_j \sum_{i,j} y_i y_j M_{i,j} = \sum_{k,l} y_i y_j y_k y_l M_{k,l}\;.$$

Using the fact that entries of Y are independent and have mean 0, we see that, in the sum, the only terms that will not be 0 in expectation are those for which each index appears at least twice. If $i \neq j$, only the terms of the form $y_i^2 y_j^2$ have this property. So if $i \neq j$,

$$\mathbf{E}\left(R_{i,j}\right) = \mathbf{E}\left(y_i^2 y_j^2 (M_{i,j} + M_{j,i})\right) = \sigma^4(M_{i,j} + M_{j,i})\;.$$

Let us now turn to the diagonal terms. Here again, only the terms $y_i^2 y_k^2$ matter. So on the diagonal,

$$\mathbf{E}\left(R_{i,i}\right) = \mu_4 M_{i,i} + \sigma^4 \sum_{j\neq i} M_{j,j} = (\mu_4 - \sigma^4) M_{i,i} + \mathrm{trace}\,(M)\;.$$

We conclude that

$$\mathbf{E}\left(R\right) = \sigma^4(M+M') + (\mu_4 - 3\sigma^4)\mathrm{diag}(M) + \mathrm{trace}\,(M)\,\mathrm{Id}\;.$$

The second part of the proof follows from the first result, after we remark that, if D is a diagonal and L is general matrix, $\mathrm{trace}\,(LD) = \mathrm{trace}\,(L \circ D)$, from which we conclude that $\mathrm{trace}\,(M\mathrm{diag}(M)) = \mathrm{trace}\,(M \circ \mathrm{diag}(M)) = \mathrm{trace}\,(M \circ M)$.

□

Lemma A-2 (Concentration of quadratic forms)**.** *Suppose the vectors Z is a vector in $\mathbb{R}^p$, with i.i.d entries of mean 0 and variance σ^2. Suppose that their entries are bounded by B_p. Let M be a symmetric matrix, with largest singular value $\sigma_1(M)$. Call*

$$\zeta_p = \frac{128\exp(4\pi)\sigma_1(M)B_p^2}{p}$$
$$\nu_p = \sqrt{\sigma_1(\Sigma)}$$

Then we have, if $r/2 > \zeta_p$,

$$P\left(\left|\frac{Z'MZ}{p} - \sigma^2\frac{trace\,(M)}{p}\right| > r\right) \leq 8\exp(4\pi)\exp(-p(r/2-\zeta_p)^2/(32B_p^2(1+2\nu_p)^2\sigma_1(M))) \tag{A-3}$$
$$+\, 8\exp(4\pi)\exp(-p/(32B_p^2(1+2\nu_p)^2\sigma_1(M)))\;.$$

Proof. We can decompose, using the spectral decomposition of M, $M = M_+ - M_-$, where M_+ is positive semi-definite and M_- is positive definite (or 0 if M is itself positive semi-definite). We can do so by replacing the negative eigenvalues of M by 0 in the spectral decomposition and get M_+ in that way. Note that then, the largest singular values of M_+ and M_- are also bounded by $\sigma_1(M)$, since $\sigma_1(M)$ is absolute value of the largest eigenvalue of M in absolute value, and the non-zero eigenvalues of M_+ are a subset of the eigenvalues of M, and so are the eigenvalues of M_-, when M_- is not 0. Now it is clear that the function F which associates to a vector x in $\mathbb{R}^p$ the scalar $\sqrt{x'M_+x/p} = \|M_+^{1/2}x/\sqrt{p}\|_2$ is a convex, $\sqrt{\sigma_1(M)/p}$-Lipschitz function with respect to Euclidian norm. Calling m_F the median of this function, when x is sampled like Z, we have, using Corollary 4.10 in Ledoux (2001)

$$P(|F(Z) - m_F| > r) \leq 4\exp(-pr^2/(16B_p^2\sigma_1(M))) \; .$$

Let us call μ_F the mean of F (it exists according to Proposition 1.8 in Ledoux (2001)). Following the arguments given in the proof of this Proposition 1.8, and spelling out the constants appearing in the last result of Proposition 1.8 in Ledoux (2001), we see that

$$P(|F(Z) - \mu_F| > r) \leq 4\exp(4\pi)\exp(-pr^2/(32B_p^2\sigma_1(M))) \; .$$

(Using the notation of Proposition 1.8 in Ledoux (2001), we picked $\kappa_2 = 1/2$, and $C' = \exp(\pi C^2/4)$; showing that this is a valid choice just requires to carry out some of the computations mentioned in the proof of that Proposition.)

Let us call A, B, D the sets

$$\begin{aligned} A &\triangleq \left\{ \left| \frac{Z'M_+Z}{p} - \mu_F^2 \right| > r \right\} \; , \\ B &\triangleq \left\{ \sqrt{\frac{Z'M_+Z}{p}} + \mu_F \leq 1 + 2\mu_F \right\} = \left\{ \sqrt{\frac{Z'M_+Z}{p}} - \mu_F \leq 1 \right\} \quad \text{and} \\ D &\triangleq \left\{ \left| \sqrt{\frac{Z'M_+Z}{p}} - \mu_F \right| > r/(1+2\mu_F) \right\} \; . \end{aligned}$$

Of course, we have $P(A) \leq P(A \cap B) + P(B^c)$. Now note that $A \cap B \subseteq D$, simply because for positive reals, $a - b/(\sqrt{a} + \sqrt{b}) = \sqrt{a} - \sqrt{b}$. We conclude that

$$P(A) \leq 4\exp(4\pi)\left[\exp(-pr^2/(32B_p^2(1+2\mu_F)^2\sigma_1(M))) + \exp(-p/(32B_p^2\sigma_1(M)))\right] \; .$$

Let us know call σ^2 the variance of the each of the component of Z. We know, according to Proposition 1.9 in Ledoux (2001), that

$$\operatorname{var}(F) = \frac{\mathbf{E}\left(Z'M_+Z\right)}{p} - \mu_F^2 = \sigma^2\frac{\operatorname{trace}(M_+)}{p} - \mu_F^2 \leq \zeta_p = \frac{128\exp(4\pi)\sigma_1(M)B_p^2}{p} \; .$$

Hence, we conclude that,if $r > \zeta_p$,

$$\begin{aligned} P\left(\left| \frac{Z'M_+Z}{p} - \sigma^2\frac{\operatorname{trace}(M_+)}{p} \right| > r \right) &\leq 4\exp(4\pi)\exp(-p(r-\zeta_p)^2/(32B_p^2(1+2\mu_F)^2\sigma_1(M))) \\ &+ 4\exp(4\pi)\exp(-p/(32B_p^2(1+2\mu_F)^2\sigma_1(M))) \; . \end{aligned}$$

To get the announced result, we note that for the sum of two reals to be greater than r in absolute value, one needs to be greater than $r/2$, and that our bounds become conservative when we replace μ_F (and its counterpart for M_-) by ν_p. (Note that the get conservative bounds when replacing the μ_F's by $\max(\mathbf{E}\left(\sqrt{Z'M_+Z/p}\right), \mathbf{E}\left(\sqrt{Z'M_-Z/p}\right))$, and that this quantity is clearly bounded by $\sigma\sigma_1(\Sigma)$.) Hence, we have, as announced: if $r/2 > \zeta_p$,

$$\begin{aligned} P\left(\left| \frac{Z'MZ}{p} - \sigma^2\frac{\operatorname{trace}(M)}{p} \right| > r \right) &\leq 8\exp(4\pi)\exp(-p(r/2-\zeta_p)^2/(32B_p^2(1+2\mu_F)^2\sigma_1(M))) \\ &+ 8\exp(4\pi)\exp(-p/(32B_p^2(1+2\mu_F)^2\sigma_1(M))) \; . \end{aligned}$$

Finally, we note that the proof makes clear that the same result would hold for different choices of M_+ and M_-, as long as $\max(\sigma_1(M_+), \sigma_1(M_-)) \leq \sigma_1(M)$. □

We therefore have the following useful corollary:

Corollary A-1. *Let Y_i and Y_j be i.i.d random vectors as in Lemma A-2, with variance 1. Suppose that Σ is a positive semi-definite matrix. We have, with*

$$\begin{aligned}\zeta_p &= \frac{128\exp(4\pi)\sigma_1(\Sigma)B_p^2}{p} \text{ , and} \\ \nu_p &= \sqrt{\sigma_1(\Sigma)} \text{ ,}\end{aligned}$$

that if $r/2 > \zeta_p$, and $K = 8\exp(4\pi)$,

$$\begin{aligned}P\left(\left|\frac{Y_i'\Sigma Y_j}{p}\right| > r\right) &\leq K\exp(-p(r/2-\zeta_p)^2/(32B_p^2(1+2\nu_p)^2\sigma_1(\Sigma))) \\ &+ K\exp(-p/(32B_p^2(1+2\nu_p)^2\sigma_1(\Sigma))) \text{ .}\end{aligned} \tag{A-4}$$

Proof. The proof relies on the results of Lemma A-2. Remark that, since Σ is symmetric,

$$Y_i'\Sigma Y_j = \frac{1}{2}(Y_i' Y_j') \begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix} \begin{pmatrix} Y_i \\ Y_j \end{pmatrix} \text{ .}$$

Now the entries of the vector made by concatenating Y_i and Y_j are i.i.d. and so we fall back into the setting of Lemma A-2. Finally, here M_+ and M_- are known explicitly. A possible choice is $M_+ = 1/2\begin{pmatrix} \Sigma & \Sigma \\ \Sigma & \Sigma \end{pmatrix}$ and $M_- = 1/2\begin{pmatrix} \Sigma & 0 \\ 0 & \Sigma \end{pmatrix}$. ν_p is obtained by upper bounding the expectation of the square of F in the notation of the proof of the previous Lemma, for these explicit matrices. Note that their largest singular values are both smaller that $\sigma_1(\Sigma)$, so the results of the previous lemma apply. □

Lemma A-3. *Let $\{Y_i\}_{i=1}^n$ be i.i.d random vectors in $\mathbb{R}^p$, whose entries are i.i.d, mean 0, variance 1, and have bounded (in p) $m \geq 4$ moments. Suppose that $\{\Sigma_p\}$ is a sequence of positive semi-definite matrices, whose operator norms are uniformly bounded in p and n/p is asymptotically bounded. We have, for any given $\epsilon > 0$,*

$$\max_{i,j}\left|\frac{Y_i'\Sigma_p Y_j}{p} - \delta_{i,j}\frac{\text{trace}\,(\Sigma_p)}{p}\right| \leq p^{-1/2+2/m}\,(\log(p))^{(1+\epsilon)/2} \;\; a.s \text{ .}$$

Proof. In all the proof, we assume without loss of generality that $m < \infty$.

Call $t = 2/m$. According to Lemma 2.2 in Yin et al. (1988), the maximum of the array of $\{Y_i\}_{i=1}^n$ is a.s less than p^t. So to control the maximum of the inner products of interest, it is enough to control the same quantity when we replace Y_i by $\widetilde{Y_i}$, with $\widetilde{Y_{i,l}} \triangleq Y_{i,l}1_{|Y_{i,l}| \leq p^t}$. Now note that $\widetilde{Y_i}$ satisfies the assumptions of Corollary A-1, except for the fact that its mean is not necessarily zero. Note however, that all the entries of $\widetilde{Y_i}$ have the same mean, $\widetilde{\mu}$. Since Y_i has mean 0, we have

$$|\widetilde{\mu}| \leq \mathbf{E}\left(|Y_{1,1}|1_{|Y_{1,1}|>p^t}\right) \leq \mathbf{E}\left(|Y_{1,1}|^m p^{-t(m-1)}\right) \leq \mu_m p^{-2+t} \text{ .}$$

Similarly, if we call $\widetilde{\sigma}^2$ the variance of $\widetilde{Y}$, we have

$$\widetilde{\sigma}^2 = \mathbf{E}\left(|Y_{1,1}|^2 1_{|Y_{1,1}|\leq p^t}\right) - \widetilde{\mu}^2 = 1 - \left(\mathbf{E}\left(|Y_{1,1}|^2 1_{|Y_{1,1}|>p^t}\right) + \widetilde{\mu}^2\right) \text{ .}$$

Hence, $0 \leq 1 - \widetilde{\sigma}^2$, and

$$\begin{aligned}1 - \widetilde{\sigma}^2 &= \mathbf{E}\left(|Y_{1,1}|^2 1_{|Y_{1,1}|>p^t}\right) + \widetilde{\mu}^2 \\ &\leq \mathbf{E}\left(|Y_{1,1}|^m p^{-t(m-2)}\right) + \widetilde{\mu}^2 \\ &\leq \mu_m p^{-2+2t} + \mu_m^2 p^{-4+2t} = \mathrm{O}(p^{-2+2t}) \text{ .}\end{aligned}$$

Now note that Corollary A-1 applies to the random variables $U_i = \widetilde{Y}_i - \widetilde{\mu} 1_p$, with $B_p = 2p^t$, when p is large enough. So $\zeta_p = \mathrm{O}(p^{1-2t})$. Let us now call, for some $\epsilon > 0$,

$$r(p) = p^{t-1/2}(\log(p))^{(1+\epsilon)/2} .$$

Since, for p large enough, $r(p)/2 > \zeta_p$, we can apply the conclusions of Corollary A-1, and plugging-in the different quantities, we see that

$$P(|U_i'\Sigma_p U_j/p| > r(p)) \leq \exp(-K(\log(p))^{1+\epsilon}) ,$$

where K denotes a generic constant. In particular, K is independent of p and is hence trivially bounded away from 0 as p grows. We note further that the arguments of Lemma A-2 show that, since $\widetilde{\sigma}^2$ is the variance of U_i,

$$P(|U_i'\Sigma_p U_i/p - \widetilde{\sigma}^2 \mathrm{trace}\left(\Sigma_p\right)/p| > r(p)) \leq \exp(-K(\log(p))^{1+\epsilon}) .$$

Now,

$$\frac{\widetilde{Y}_i'\Sigma_p\widetilde{Y}_j}{p} = \frac{U_i'\Sigma_p U_j}{p} + \mu\frac{(1'\Sigma_p U_j + U_i'\Sigma_p 1)}{p} + \mu^2\frac{1'\Sigma_p 1}{p} .$$

Remark that $1'\Sigma_p 1 \leq p\sigma_1(\Sigma_p)$, and $|1'\Sigma_p U_j| \leq \sqrt{1'\Sigma_p 1}\sqrt{U_j'\Sigma_p U_j}$. We conclude, using the results obtained in the proof of Lemma A-2 that with probability greater than $1 - \exp(-K(\log(p))^{1+\epsilon})$, the middle term is smaller than $2\sqrt{\sigma_1(\Sigma_p)}(\sqrt{\sigma_1(\Sigma_p)} + r(p))\mu$. As a matter of fact, $\sqrt{U_j'\Sigma_p U_j/p}$ is concentrated around its mean, which is smaller than $\widetilde{\sigma}\sqrt{\mathrm{trace}\left(\Sigma_p\right)/p}$, which is itself smaller than $\sqrt{\sigma_1(\Sigma_p)}$. Now recall that $\widetilde{\mu} = \mathrm{O}(p^{-2+t}) = \mathrm{o}(r(p))$. We can therefore conclude that,

$$P\left(\left|\frac{\widetilde{Y}_i'\Sigma_p\widetilde{Y}_j}{p} - \delta_{i,j}\widetilde{\sigma}^2\frac{\mathrm{trace}\left(\Sigma_p\right)}{p}\right| > 2r(p)\right) \leq 2\exp(-K(\log(p))^{1+\epsilon}) .$$

Now note, that $0 \leq 1-\widetilde{\sigma}^2 = \mathrm{O}(p^{-2+2t}) = \mathrm{o}(r(p))$, since $t \leq 1/2 < 3/2$. With our assumptions, $\mathrm{trace}\left(\Sigma_p\right)/p$ remains bounded, so we have finally

$$P\left(\left|\frac{\widetilde{Y}_i'\Sigma_p\widetilde{Y}_j}{p} - \delta_{i,j}\frac{\mathrm{trace}\left(\Sigma_p\right)}{p}\right| > 3r(p)\right) \leq 2\exp(-K(\log(p))^{1+\epsilon}) .$$

And therefore,

$$P\left(\max_{i,j}\left|\frac{\widetilde{Y}_i'\Sigma_p\widetilde{Y}_j}{p} - \delta_{i,j}\frac{\mathrm{trace}\left(\Sigma_p\right)}{p}\right| > 3r(p)\right) \leq 2n^2\exp(-K(\log(p))^{1+\epsilon}) .$$

Using the Borel-Cantelli Lemma, we reach the conclusion that

$$\max_{i,j}\left|\frac{\widetilde{Y}_i'\Sigma_p\widetilde{Y}_j}{p} - \delta_{i,j}\frac{\mathrm{trace}\left(\Sigma_p\right)}{p}\right| \leq 3r(p) = 3p^{2/m-1/2}\log(p) \quad \text{a.s} \quad .$$

Because the left-hand side is a.s equal to $\left|\frac{Y_i'\Sigma_p Y_j}{p} - \delta_{i,j}\frac{\mathrm{trace}(\Sigma_p)}{p}\right|$, we reach the announced conclusion, but with $r(p)$ replaced by $3r(p)$. Note that, of course, any multiple of $r(p)$, where the constant is independent of p, would work in the proof. In particular, by taking $\tilde{r}(p) = r(p)/3$, we reach the announced conclusion. □

Corollary A-2. *Under the same assumptions as that of Lemma A-3, if we call $X_i = \Sigma_p^{1/2}Y_i$, we also have*

$$\max_{i\neq j}\left|\frac{\|X_i - X_j\|_2^2}{p} - 2\frac{trace\left(\Sigma_p\right)}{p}\right| \leq p^{-1/2+2/m}(\log(p))^{(1+\epsilon)/2} \quad a.s \; .$$

Proof. The proof follows immediately from the results of Lemma A-3, after we write

$$\|X_i - X_j\|_2^2 - 2\text{trace}\,(\Sigma_p) = [Y_i\Sigma_p Y_i - \text{trace}\,(\Sigma_p)] + [Y_j\Sigma_p Y_j - \text{trace}\,(\Sigma_p)] - 2Y_i'\Sigma_p Y_j \ .$$

Note that as explained in the proof of Lemma A-3, the constants in front of the bounding sequence do not matter, so we can replace $3p^{-1/2+2/m}(\log(p))^{(1+\epsilon)/2}$ by $p^{-1/2+2/m}(\log(p))^{(1+\epsilon)/2}$, and the result still holds. (In other words, we are really using Lemma A-3 with upper bound $p^{-1/2+2/m}(\log(p))^{(1+\epsilon)/2}/3$.) □

Lemma A-4. *Let $\{X_i\}_{i=1}^n$ be i.i.d random vectors in $\mathbb{R}^p$, whose entries are i.i.d, mean 0, having the property that for 1-Lipschitz functions F, if we denote by m_F the median of F,*

$$P(|F - m_F| > r) \leq C\exp(-c(p)r^2) \ ,$$

where C is independent of p and c is allowed to vary with p. Call Σ_p the covariance matrix of the X_1. Assume that $\sigma_1(\Sigma_p)$ remains bounded in p. Then, we have

$$\max_{i,j}\left|\frac{X_i'X_j}{p} - \delta_{i,j}\frac{\textit{trace}\,(\Sigma_p)}{p}\right| \leq (pc(p))^{-1/2}\left(\log(p)\right)^{(1+\epsilon)/2} \ a.s \ .$$

Proof. The proof once again relies on concentration inequalities. First note that Proposition 1.11 combined with Proposition 1.7 in Ledoux (2001) show that if X_i and X_j are independent and satisfy concentration inequalities with concentration function $\alpha(r)$ (with respect to Euclidian norm), then the vector $\begin{pmatrix} Y_i \\ Y_j \end{pmatrix}$ also satisfies concentration inequalities, with concentration function $2\alpha(r/2)$ with respect to Euclidian norm in $\mathbb{R}^{2p}$. (We note that Proposition 1.11 is proved for the metric on $\mathbb{R}^{2p}$ $\|\cdot\|_2 + \|\cdot\|_2$, where each Euclidian norm is a norm in $\mathbb{R}^p$, but the same proof goes through for Euclidian norm on $\mathbb{R}^{2p}$. Another argument would be to say that the metric $\|\cdot\|_2 + \|\cdot\|_2$ is equivalent to the norm of the full $\mathbb{R}^{2p}$, with the constants in the inequalities being 1 and $\sqrt{2}$, simply because for $a, b > 0$, $\sqrt{a^2+b^2} \leq a + b \leq \sqrt{2}\sqrt{a^2+b^2}$.)

Therefore, the arguments of Lemma A-2 go through without any problems, with $\Sigma_p = \text{Id}$ and $B_p^2 = 4/c(p)$. So a result similar to Corollary A-1 holds and we can apply the same ideas as in the proof of Lemma A-3 and get the announced result.

□

Corollary A-3. *Under the assumptions of Lemma A-4, we have*

$$\max_{i\neq j}\left|\frac{\|X_i - X_j\|_2^2}{p} - 2\frac{\textit{trace}\,(\Sigma_p)}{p}\right| \leq (pc(p))^{-1/2}\left(\log(p)\right)^{(1+\epsilon)/2} \ a.s \ .$$

Proof. The proof is an immediate consequence of Lemma A-4, along the same lines as the proof of Corollary A-2. □

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

Bai, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica* **9**, 611–677. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.

Bai, Z. D., Miao, B. Q., and Pan, G. M. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab.* **35**, 1532–1572.

Bai, Z. D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.* **26**, 316–345.

Bai, Z. D. and Silverstein, J. W. (1999). Exact separation of eigenvalues of large-dimensional sample covariance matrices. *Ann. Probab.* **27**, 1536–1555.

Baik, J., Ben Arous, G., and Péché, S. (2005). Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. *Ann. Probab.* **33**, 1643–1697.

Baik, J. and Silverstein, J. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97**, 1382–1408.

Bhatia, R. (1997). *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.

Bogomolny, E., Bohigas, O., and Schmit, C. (2003). Spectral properties of distance matrices. *J. Phys. A* **36**, 3595–3616.

Bordenave, C. (2006). Eigenvalues of euclidean random matrices. Available at http://arxiv.org/abs/math/0606624.

Boutet de Monvel, A., Khorunzhy, A., and Vasilchuk, V. (1996). Limiting eigenvalue distribution of random matrices with correlated entries. *Markov Process. Related Fields* **2**, 607–636.

Burda, Z., Jurkiewicz, J., and Wacław, B. (2005). Spectral moments of correlated Wishart matrices. *Phys. Rev. E* **71**.

El Karoui, N. (2003). On the largest eigenvalue of Wishart matrices with identity covariance when n, p and $p/n \to \infty$. *arXiv:math.ST/0309355* To appear in *Bernoulli*.

El Karoui, N. (2007a). On the spectrum of correlation matrices and covariance matrices computed from elliptically distributed data. Technical Report 740, Department of Statistics, UC Berkeley.

El Karoui, N. (2007b). Operator norm consistent estimation of large dimensional sparse covariance matrices. Technical Report 734, Department of Statistics, UC Berkeley. To appear in *The Annals of Statistics*.

El Karoui, N. (2007c). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability* **35**, 663–714.

Forrester, P. J. (1993). The spectrum edge of random matrix ensembles. *Nuclear Phys. B* **402**, 709–728.

Geman, S. (1980). A limit theorem for the norm of random matrices. *Ann. Probab.* **8**, 252–261.

Geronimo, J. S. and Hill, T. P. (2003). Necessary and sufficient condition that the limit of Stieltjes transforms is a Stieltjes transform. *J. Approx. Theory* **121**, 54–60.

Gohberg, I., Goldberg, S., and Krupnik, N. (2000). *Traces and determinants of linear operators*, volume 116 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel.

Horn, R. A. and Johnson, C. R. (1990). *Matrix analysis.* Cambridge University Press, Cambridge. Corrected reprint of the 1985 original.

Horn, R. A. and Johnson, C. R. (1994). *Topics in matrix analysis.* Cambridge University Press, Cambridge. Corrected reprint of the 1991 original.

Johansson, K. (2000). Shape fluctuations and random matrices. *Comm. Math. Phys.* **209**, 437–476.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.* **29**, 295–327.

Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6**, 113–167.

Ledoux, M. (2001). *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)* **72 (114)**, 507–536.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* **17**. Available at `http://anson.ucdavis.edu/~debashis/techrep/techrep.html`.

Paul, D. and Silverstein, J. (2007). No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix Available at `http://www4.ncsu.edu/~jack/pub.html`.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels.* The MIT Press, Cambridge, MA.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55**, 331–339.

Tracy, C. and Widom, H. (1994). Level-spacing distribution and the Airy kernel. *Comm. Math. Phys.* **159**, 151–174.

Tracy, C. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.* **177**, 727–754.

Tracy, C. and Widom, H. (1998). Correlation functions, cluster functions and spacing distributions for random matrices. *J. Statist. Phys.* **92**, 809–835.

Voiculescu, D. (2000). Lectures on free probability theory. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pp. 279–349. Springer, Berlin.

Wachter, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probability* **6**, 1–18.

Wigner, E. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math. (2)* **62**, 548–564.

Williams, C. and Seeger, M. (2000). The effect of the input density distribution on kernel-based classifiers. *International Conference on Machine Learning* **17**, 1159–1166.

Yin, Y. Q., Bai, Z. D., and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probab. Theory Related Fields* **78**, 509–521.