RECENT RESULTS ABOUT THE LARGEST EIGENVALUE OF RANDOM COVARIANCE MATRICES AND STATISTICAL APPLICATION*

Noureddine El Karoui

Department of Statistics, University of California
367 Evans Hall, Berkeley CA 94720-3860, USA
nkaroui@stat.berkeley.edu

(Received July 21, 2005)

This note is a short review of recent results concerning the fluctuation behavior of the largest eigenvalue of a class of random covariance matrices. We also present a concrete application of these results to a model checking problem in time series analysis to highlight their practical relevance.

PACS numbers: 02.50.–r

1. Introduction

Sample covariance matrices are a fundamental tool of multivariate statistics. After data collection, we get an $n \times p$ data matrix X. We will call n the number of observations and p the number of predictors. The rows of X are assumed to be realizations of a random variable whose covariance structure is Σ_p. For practical applications, one often wishes to estimate Σ_p in order to understand the dependence structure of the predictors, do various tests *etc.* Here the sample covariance matrix X^*X/n plays a key role. (Of course, in practice, it is often computed as $(X-\bar{X})^*(X-\bar{X})/(n-1)$, where $\bar{X}$ stands for the column-wise mean of the matrix X, but for the sake of this note we will assume that the entries are centered.)

One most important statistical application in which eigenvalues of the covariance matrix play a key role is Principal Component Analysis (PCA). It is a linear dimensionality reduction procedure, which can also be thought of as a model selection technique. The idea is as follows. We are interested in recovering as much of the total variance in the data as possible while reducing the dimensionality of the problem from p to k. In other words, we

* Presented at the Conference on Applications of Random Matrices to Economy and Other Complex Systems, Kraków, Poland, May 25–28, 2005.

are looking for k vectors $e_1, \ldots, e_k$ in $\mathbb{R}^p$ such that $\sum_{m=1}^{k} \mathrm{var}(\langle X_i, e_m \rangle)$ is maximal. $\{X_i\}_{i=1,\ldots,n}$ are assumed for simplicity to be i.i.d $\mathcal{N}(0, \Sigma_p)$. Of course, it is easy to see that one should choose for e_m's the eigenvectors associated with the first k eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ of Σ_p. The main question becomes how to choose k. To this end, one popular method in Statistics is the so-called "scree plot". An illustration and explanations follow. (The interested reader can find an interesting account on PCA in [19], Chapter 8.)

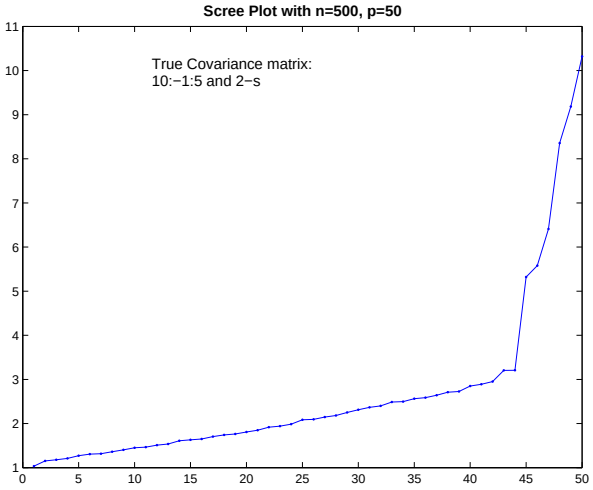

Fig. 1. **Scree Plot :** we compute $l_1 \geq l_2 \geq \ldots \geq l_p$ the eigenvalues of $X'X/n$, where X is a 500×50 matrix whose rows are i.i.d $\mathcal{N}(0, \Sigma_p)$ and Σ_p is a diagonal matrix with eigenvalues (10,9,8,7,6,5,2,2,... ,2). The eigenvalues of $X'X/n$ are ordered. The scree plot is a plot of the pairs $(51 - i, l_i)$ where i is the rank of the corresponding eigenvalue. In general one looks for an "elbow" on this plot to decide what k they should use in PCA. We would keep the first 6 eigenvalues by using the scree plot shown above. In this example, it corresponds to what we would do if we had perfect information, *i.e.* we knew the true covariance structure.

PCA has enjoyed wide popularity among practitioners of multivariate statistics for a long time now. A key rationale is a set of results first obtained by [1] regarding the asymptotic behavior of $l_1, \ldots, l_p$, the eigenvalues of $X'X/n$. A good reference is the classic text [2], especially Chapters 7, 11 and 13. The asymptotic results presented in [2] are valid under the following assumptions: (1) normality of the entries of X, (2) i.i.d-ness of its rows, (3) the eigenvalues of Σ_p all have multiplicity one, (4) Σ_p and hence p are fixed. Then [1] showed among many other things, that l_1 is a $\sqrt{n}$-consistent estimator of λ_1, the largest eigenvalue of Σ_p. Namely, as $n \to \infty$,

$$\sqrt{n}\left(l_1(X'X/n) - \lambda_1\right) \Longrightarrow \mathcal{N}(0, 2\lambda_1^2)\,.$$

We refer the reader to Theorem 13.5.1 in [2] for more details. Here "$\Rightarrow$" stands for convergence in distribution. This result and the other ones given in Theorem 13.5.1 of [2] mean, in particular, that the l_i's are consistent estimators of the λ_i's (*i.e.* the former quantities converge in probability to the latter ones). This is of course a desirable property and gives theoretical ground for using PCA.

1.1. The issue of large n, large p

Because of advances in data collection mechanisms, statisticians now often encounter datasets for which both sample size and number of predictors are large. In genetic studies, for instance, it is not uncommon to have p (the number of genes in this context) of order a few 1,000's and n (the number of patients) of order 100. In financial applications, when working with a year worth of say log-returns for companies belonging to the S&P 500, one has $n \simeq 250$ (number of days) and $p \simeq 500$ (number of companies). See also [13], pp. 485–493 for examples of applications to other fields, such as image recognition.

In these situations, it is not so clear that classical results yield relevant information. The fundamental assumption underlying the classical results, namely p fixed and n going to infinity, does not seem to fit very well these new practical situations. The potential pitfalls we face are depicted in figure 2.

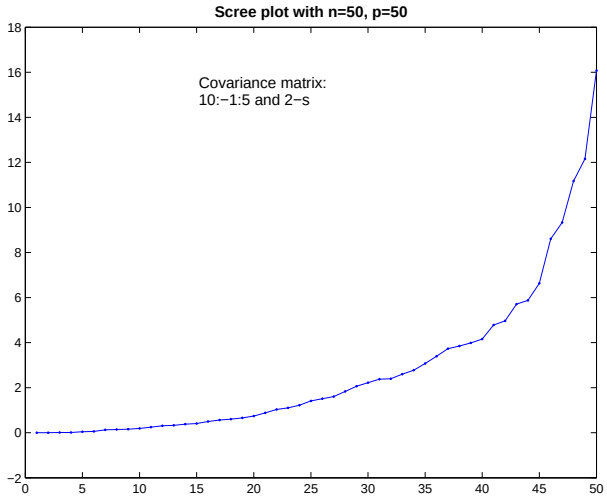

Fig. 2. **Large n, large p problem:** this picture represents a scree plot obtained with the same parameters as in figure 1, except for n which is now 50. While the covariance structure Σ_p is the same, we now have $n = p = 50$. The striking feature is the disappearance of the "elbow".

In light of figure 2, we would therefore like to understand how the classical tools used in multivariate statistics behave under another type of asymptotics: a situation in which both p and n go to infinity. We will see that this change of assumptions lead to a very different asymptotic behavior for the largest eigenvalue of X^*X/n. Ideally, we would also like to have estimates of distance between the finite p and n quantities of interest and their asymptotic counterparts so as to be able to meaningfully choose the relevant type of asymptotics in a specific practical situation. At this point in time, this type of estimates still seem to be a long way away.

2. Results about large random covariance matrices

The first results obtained under the "large p, large n" asymptotics framework date back from work of Marčenko and Pastur [18]. The first statistical application of these ideas, to which we will come back in Section 3, can be found in [26]. These results were mostly concerned with the behavior of the empirical spectral distribution of X^*X, *i.e.* the distribution that puts mass $1/p$ at all the eigenvalues of the $p \times p$ matrix X^*X.

2.1. Marčenko–Pastur law

The statement of this important result is the following (see [3]):

Theorem 1 (Marčenko–Pastur law) *Let X be an $n \times p$ matrix with independent, identically distributed entries $X_{i,j}$. We assume that $E(X_{i,j}) = 0$ and $var(X_{i,j}) = 1$. Suppose $l_1 \geq l_2 \geq ... \geq l_p$ are the p eigenvalues of $\frac{1}{n}X'X$. Suppose $\rho_n = p/n \to \rho \in (0, 1]$.*
If $F_p(x) = \frac{1}{p}\{\#l_i \leq x\}$, then

$$F_p \Longrightarrow F_\rho \quad \textit{almost surely (a.s)},$$

where F_ρ has a known density. ($f_\rho(x) = \sqrt{(b-x)(x-a)}/(2\pi x\rho)$, with $a = (1-\rho^{1/2})^2$, $b = (1+\rho^{1/2})^2$).

We refer to the review paper [3] for more details about ways of proving this result and a complete description of the sequence of papers that led to this stronger version of the original theorem found in [18].

While this is of course an extremely interesting result both from theoretical and practical standpoints (see *e.g.* [16]), it has possible drawbacks when one considers using it, for instance, for hypothesis testing. First, the Marčenko–Pastur result does not provide fluctuation information since it is an almost sure result. Second, there is some extra difficulty encountered in testing hypotheses about random measures as opposed to, say, real random variables. Third, while we have some information about what happens

when the covariance within a row is Σ_p and the rows are independent, the results are quite subtle and not extremely explicit (see Theorem 3.4 in [3] and [20, 21]). Moreover, the numerical inversion of the Stieltjes transforms that appear there is in itself a difficult problem (see *e.g.* [5], p. 4; note that equation (28) in [5] is equivalent to, for instance, equation 1.4 in [20]). These are some of the problems that hinder the use of the Marcčenko–Pastur law for hypothesis testing.

In other respects, a natural question stemming from the Marčenko–Pastur law, in particular in light of applications to PCA, is to understand how the largest eigenvalue of X^*X/n behaves, as opposed to its whole spectrum. The first result in this direction was obtained by Geman in [12], and subsequently refined in a series of paper (see [3], Section 2.2.2). They state that:

Theorem 2 (a.s limit of largest eigenvalue) *Under the assumptions of Theorem 1 and assuming that the entries of the matrix X have finite 4th moment, we have*

$$l_1(X^*X/n) \to (1+\sqrt{\rho})^2 \, , a.s \; .$$

Hence, $l_1(X^*X/n)$ is an inconsistent estimator of λ_1 in the large p, large n setting. This is very markedly different from the classical situation (p fixed, n goes to ∞) that we recalled in the introduction. Following recent developments in random matrix theory (RMT), there has been an important renewed focus in recent years on understanding the behavior of the largest eigenvalue of random covariance matrices. As we will see, these new results allow us to address the three points that we raised above about the difficulty of doing hypothesis testing when using only the Marčenko–Pastur law. At this point in time, most of these recent results are formulated in the case where the entries of X are complex normal random variables.

2.2. Fluctuation behavior of the largest eigenvalue of random covariance matrices

The results of the previous subsection showed that we have a lot of information about the almost sure behavior of the largest eigenvalue of large covariance matrices. For a number of practical applications, we often would like to be able to build confidence intervals, and hence a good understanding of the fluctuation behavior of the statistics of interest is required. This is the topic to which we now turn.

A note on notation before we proceed. In what follows, we will say that a vector is distributed as complex normal with covariance Σ_p and write $V \sim \mathcal{N}_{\mathbb{C}}(0, \Sigma_p)$, to say that $V = Y + iZ$, where Y and Z are independent real multivariate normal variables with $Y \sim \mathcal{N}(0, \Sigma_p/2)$ and $Z \sim \mathcal{N}(0, \Sigma_p/2)$.

2.2.1. The case of $\Sigma_p = \mathrm{Id}_p$

The first result in this direction was obtained by Forrester in [11] (in the case $n - p = C$, where C is a constant), and later extended by Johansson in [14]. The latter showed that:

Theorem 3 (Johansson) *Let us assume that $X_{i,j}$, the entries of the $n \times p$ matrix X are i.i.d $\mathcal{N}_{\mathbb{C}}(0,1)$. Then, if $\rho \in (0,1]$, and as $n \to \infty$, $p/n = \rho + O(n^{-2/3})$, and l_1 is the largest eigenvalue of X^*X/n, we have*

$$n^{2/3} \frac{l_1 - (1+\sqrt{p/n})^2}{(1+\sqrt{p/n})(1+\sqrt{n/p})^{1/3}} \Rightarrow TW_2 \ ,$$

where TW_2 has the Tracy–Widom distribution appearing in the study of the Gaussian Unitary Ensemble (GUE).

Turning to the real case, Johnstone [15] then showed that:

Theorem 4 (Johnstone) *Let us assume that $X_{i,j}$ are i.i.d $\mathcal{N}(0,1)$. Then, if $p/n \to \rho \in (0,1]$, as $n \to \infty$, and if l_1 is the largest eigenvalue of $X'X/n$, we have*

$$n^{2/3} \frac{l_1 - (\sqrt{1-1/n}+\sqrt{p/n})^2}{(\sqrt{1-1/n}+\sqrt{p/n})(\sqrt{1+1/(n-1)}+\sqrt{n/p})^{1/3}} \Rightarrow TW_1 \ ,$$

where TW_1 has the Tracy–Widom distribution appearing in the study of the Gaussian Orthogonal Ensemble.

Note that Johnstone also showed that Johansson's result held when assuming only that $n/p \to \rho$. For information about Tracy–Widom distributions, we refer the reader to [23, 24] and [7].

One naturally wonders about whether the somewhat restrictive Gaussian assumption can be removed. Soshnikov showed in [22] that assuming that (a) $n - p = O(p^{1/3})$, (b) $X_{i,j}$ are symmetric random variables and (c) a growth condition on the moments of $X_{i,j}$, the Tracy–Widom limit was universal. In other words, one could remove the normality assumption in the Johansson–Johnstone theorem (and replace it by the three conditions (a), (b) and (c)).

Note that because X and X^* have the same singular values (up to a number of 0's), the previous results also hold in the case $p/n \to \rho \in (0, \infty)$. (For centering and scaling purposes, one just needs to invert the roles of p and n.) A natural question was therefore to try to understand the situation in the case $p/n \to 0$. From a practical standpoint, this is interesting since if the result were to break down for natural relationships between n and p (for instance, $p = n^{1-\beta}$, $\beta > 0$), and other limiting distributions were to appear in these situations, the practical relevance of the Tracy–Widom approximation would be reduced. In [8], it is shown that:

Theorem 5 (NEK) *The Johansson-Johnstone theorem holds even when $p/n \to 0$, as long as both n and p go to ∞. Since X and X^* have the same largest singular value, it is also the case when $p/n \to \infty$.*

One interesting aspect of the previous result is that it is independent of the speed of convergence of p/n to 0 (or ∞).

A remarkable aspect of these convergence results is that they provide very good approximations to the finite-dimensional distributions of interest. This was first remarked by Johnstone in [15]. The following table illustrates this claim:

TABLE I

Quality of Tracy–Widom approximation, real case: each column is generated by simulating 10,000 matrices X, with entries i.i.d $\mathcal{N}(0,1)$, of dimensions $n \times p$ indicated at the top of the column. For each matrix, we extract the largest eigenvalue of X^*X/n, recenter and rescale it according to Theorem 4. We then compute the percentage of data points that are to the left of a given quantile of the limiting Tracy–Widom distribution (1st column) and compare it to the theoretical prediction (2nd column). The last column is twice the standard error of the binomial distribution with probability according to the 2nd column and number of repetitions 10,000. (Quantiles courtesy of Iain Johnstone).

Qtiles	TW	10×10^3	10×4000	10×10^4	100×4000	30×5000	2*SE
−3.9	.01	0.009	0.010	0.015	0.012	0.013	.002
−3.18	.05	0.047	0.050	0.060	0.053	0.055	.004
−2.78	.10	0.102	0.107	0.112	0.103	0.105	.006
−1.91	.30	0.303	0.308	0.316	0.304	0.303	.009
−1.27	.50	0.506	0.506	0.522	0.508	0.503	.010
−0.59	.70	0.705	0.704	0.723	0.706	0.702	.009
0.45	0.9	0.904	0.904	0.913	0.901	0.904	.006
0.98	.95	0.953	0.951	0.958	0.951	0.953	.004
2.02	.99	0.992	0.990	0.992	0.991	0.991	.002

It is natural to try to understand why this approximation is so good and to see if one can improve upon it in practice. This is one of the reasons for which the centering and scaling in Theorem 4 differ from that in Theorem 3, as Johnstone found empirically that the centering and scaling of Theorem 4 improved the numerical quality of the approximation. This sort of consideration is of course of interest to anyone who uses the asymptotic distribution as a proxy for the finite dimensional one (for *e.g.* confidence interval building or hypothesis testing), because the greater the quality of the approximation the greater the accuracy of our conclusions or inferences.

Some theoretical results driven by these practical considerations were obtained in [9]. These results concern the rate of convergence of the finite dimensional distributions to their asymptotic limit, which is shown to be at least 2/3. More precisely, it is shown in [9] that:

Theorem 6 (NEK) *Denote by F_2 the cumulative distribution function of TW_2. When the entries of the matrix X are i.i.d $\mathcal{N}_{\mathbb{C}}(0,1)$, let l_1 denote the largest eigenvalue of X^*X/n. Then, there exists $\tilde{\mu}_{n,p}$, $\tilde{\sigma}_{n,p}$, and a function M such that as n, p tend to $+\infty$, and $n/p \to \rho \in \mathbb{R}_+^*$ we have: for all s_0, there exists $N(s_0)$, such that for all $s \geq s_0$, and $n \geq N(s_0)$,*

$$(n \wedge p)^{2/3} \left| P\left(n^{2/3} \frac{l_1 - \tilde{\mu}_{n,p}}{\tilde{\sigma}_{n,p}} \leq s \right) - F_2(s) \right| \leq M(s_0)$$

M can be chosen to be non-increasing.

The centering and scaling sequences $\tilde{\mu}_{n,p}$ and $\tilde{\sigma}_{n,p}$ have explicit expressions. Since they are quite involved and we will not be using them, we refer the interested reader to [9] for the details. The main feature we would like to mention is that they are easy to implement on a computer and are therefore practically very usable. Note that the 2/3 rate obtained in the previous theorem is better than the traditional 1/2 rate obtained in the Berry–Esséen theorem, a rate of convergence result for the classical central limit theorem. This "good" rate and properties of the function M (see [9]; the article is currently under final revision before publication and will contain a better bound than $M(s_0)$ on the right hand side of the inequality) help explain, at least at an intuitive level, why the approximation works well in practice.

While it is important to understand the situation in the case where the covariance structure of the data is the identity matrix, it is clear that in practice this is rarely the case. To compute the power of our tests, or to understand how biased our statistics might be, we need to investigate the case where the covariance is not Id.

2.2.2. The case of $\Sigma_p \neq \mathrm{Id}_p$

Investigations of this situation have only started recently, the first results being obtained by Baik, Ben Arous and Péché in [4]. The only known results assume that the entries of the matrix X are complex normal. Note that even this situation is not fully understood as there is a wide range of possible behavior for Σ_p. [4] investigates the situation where Σ_p is a finite perturbation of Id_p. By this we mean that the eigenvalues of Σ_p are all equal to 1, except for a finite number, k, that is fixed in the asymptotic analysis. In other words, we have $\lambda_1(\Sigma_p) \geq \lambda_2(\Sigma_p) \geq \ldots \geq \lambda_k(\Sigma_p) > \lambda_{k+1}(\Sigma_p) = \ldots = \lambda_p(\Sigma_p) = 1$.

The authors of [4] discovered — among many other things — a very interesting phase transition picture for the behavior of the largest eigenvalue of X^*X/n under the aforementioned assumptions. Essentially, when λ_1 is "far enough" from 1, l_1, properly centered and scaled, behaves like the largest eigenvalue of a matrix belonging to the $m_1 \times m_1$ Gaussian Unitary Ensemble, where m_1 is the multiplicity of λ_1. When λ_1 is "close enough" to 1, then l_1 is asymptotically Tracy–Widom, with the same centering and scaling as in Theorem 3. In between, it behaves according to new distributions found in [4]. To make this statement precise, we present a simple version of the main theorem in [4]. We focus only on the case where $m_1 = 1$, as it leads to the most familiar distributions. (It is a very reduced version of the main theorem in [4], where more information can be found.)

Theorem 7 (Baik–Ben Arous–Péché) *Let the rows of the $n \times p$ matrix X be i.i.d $\mathcal{N}_{\mathbb{C}}(0, \Sigma_p)$. Let l_1 be the largest eigenvalue of X^*X/n. Assume that Σ_p is a finite perturbation of Id_p.*
Assume that $\lambda_1(\Sigma_p)$, the largest eigenvalue of Σ_p has multiplicity 1. Let us call $\rho_n = p/n$ and assume that, as n goes to ∞, $\eta \leq \rho_n \leq 1$, for some $\eta > 0$. If for some $\varepsilon > 0$, $\lambda_1 > 1 + \sqrt{\rho_n} + \varepsilon$, and λ_1 is uniformly bounded as p and n go to ∞,

$$\sqrt{n}\,\frac{l_1/\lambda_1 - (1 + p/(n(\lambda_1 - 1)))}{\sqrt{1 - p/(n(\lambda_1 - 1)^2)}} \Rightarrow \mathcal{N}(0,1)\,.$$

If for some $\varepsilon > 0$, $\lambda_1 < 1 + \sqrt{\rho_n} - \varepsilon$,

$$n^{2/3}\frac{l_1 - (1 + \sqrt{p/n})^2}{(1 + \sqrt{p/n})(1 + \sqrt{n/p})^{1/3}} \Rightarrow TW_2\,.$$

If $\lambda_1 = 1 + \sqrt{\rho_n}$, then

$$n^{2/3}\frac{l_1 - (1 + \sqrt{p/n})^2}{(1 + \sqrt{p/n})(1 + \sqrt{n/p})^{1/3}} \Rightarrow F_1\,,$$

where F_1 is a generalized Tracy–Widom distribution.

This result raises many questions: how "natural" is the Tracy–Widom limit for the largest eigenvalue of random covariance matrices? In other words, when the limit is Tracy–Widom, does the large amount of symmetry still found in the problem when assuming that the covariance structure is a finite dimensional perturbations of Id play an important role? Naturally, we may also wonder how important is the assumption that Σ_p is a finite perturbation of Id_p. Or try to understand the bias in l_1 when Σ_p is not a finite perturbation of Id. These types of questions are investigated in [10]. The main result of this paper can be stated as follows:

Theorem 8 (NEK) *Let us consider X, an $n \times p$ matrix whose rows are i.i.d $\mathcal{N}_{\mathbb{C}}(0, \Sigma_p)$. Let λ_1 be the largest eigenvalue of Σ_p and λ_p the smallest one. Let H_p be the spectral distribution of Σ_p. Let c be the unique solution in $[0, 1/\lambda_1(\Sigma_p))$ of the equation*

$$c = c(\Sigma_p, n, p)\,,\ c \in [0, 1/\lambda_1(\Sigma_p))\ :\ \int \left(\frac{\lambda c}{1-\lambda c}\right)^2 dH_p(\lambda) = \frac{n}{p}\,.$$

We assume that $n/p \geq 1$ is uniformly bounded, and that $\limsup \lambda_1 < \infty$, $\liminf \lambda_p > 0$, and $\limsup \lambda_1 c < 1$. We denote by $\mathcal{G}$ the class of models $\{(\Sigma_p, n, p)\}$ for which these conditions hold. We call

$$\mu = \frac{1}{c}\left(1 + \frac{p}{n}\int \frac{\lambda c}{1-\lambda c} dH_p(\lambda)\right)\ ,\ \textit{and}$$
$$\sigma^3 = \frac{1}{c^3}\left(1 + \frac{p}{n}\int \left(\frac{\lambda c}{1-\lambda c}\right)^3 dH_p(\lambda)\right)\ .$$

Let l_1 be the largest eigenvalue of X^*X/n. Then we have, when the covariance model is in $\mathcal{G}$ and n goes to ∞,

$$n^{2/3}\ \frac{l_1(X^*X/n) - \mu}{\sigma} \Rightarrow TW_2\ .$$

(Both μ and σ are functions of (Σ_p, n, p) but we dropped this dependence for the sake of clarity.)

The relevance of this result depends, of course, on how large $\mathcal{G}$ is. As shown in [10], $\mathcal{G}$ contains a large number of models found in applications. Among others, the situation where H_p is a finite sum of atoms (*i.e.* diracs) having a mass that does not go to 0 when p tends to ∞ is covered by the theorem. This means that the $\Sigma_p = \mathrm{Id}_p$ is a subcase of Theorem 8.

It also means that asymptotics done at "fixed spectral distribution" are in the realm of application of the theorem. By asymptotics at fixed spectral distribution we mean that H_p is fixed in the asymptotic analysis. (Concretely, in this situation, we assume that Σ_{2p} has the same eigenvalues as Σ_p, and their multiplicity in Σ_{2p} is twice what it is in Σ_p. It is then clear that $H_{2p}(\Sigma_{2p}) = H_p(\Sigma_p)$.) This type of procedure is practically appealing as it is a minimal assumption made about the sequence H_p and one that, practitioners hope, lead to good approximations to the finite dimensional behavior of the statistics of interest. In particular, this assumption is often made, at least implicitly, in numerical implementations of results in [18, 20], and [21] relating the Stieltjes transform of the limiting spectral distribution of X^*X/n to certain integrals computed against the limit H of H_p as p tends to ∞.

Another class of covariance models often found in applications that belongs to $\mathcal{G}$ is that of well-behaved Toeplitz matrices. A Toeplitz matrix T is a matrix for which $T_{i,j} = t(i-j)$, for a certain function t. It is shown in [10] that finite order Toeplitz matrices, *i.e.* Toeplitz matrices for which $t = 0$ is $|i-j| > l_0$ for a certain l_0, generally belong to $\mathcal{G}$. One can also check that AR(1) matrices, *i.e.* $t(k) = r^k$ for a certain r are also in $\mathcal{G}$, if $|r| \neq 1$. Finally, a number of finite perturbations of matrices in $\mathcal{G}$ are also in $\mathcal{G}$. For more general criteria for belonging to $\mathcal{G}$ and more details, we refer the reader to [10].

Statistically speaking, the last theorem gives us a way to compute the power of tests based on the largest eigenvalue of sample covariance matrices for a large class of alternatives. Note that at this point, the theorem is restricted to the complex case situation. Nevertheless, simulations seem to indicate that the same results might hold in the real case for normal variables, after we replace TW_2 by TW_1 as the limit appearing in the theorem. The centering and scaling sequences should be left unchanged.

3. A statistical application

In this last section, we revisit a fairly classical statistical problem by making use of the theoretical results we presented above. The problem we will look at from this new perspective is that of testing for white Gaussian noise in time series analysis.

White Gaussian noise is an essential driving element of most time series models. Often, the last step of a model building process will be to check whether residuals are white noise. Numerous approaches exist. A basic and often advocated one is the Ljung–Box statistic.

Let us describe the problem in slightly more detail. We call $\{\varepsilon_t\}$ the white noise time series, and ρ_k the lag k autocorrelation of the abstract time series we are considering. We have T points. To test the null hypothesis $H_0 : \rho_1 = \rho_2 = \ldots = \rho_m = 0$, Ljung and Box [17] (see [25], p. 25) propose to use the statistic

$$Q(m) = T(T+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{T-k} \quad .$$

Under the assumption that the series at stake is i.i.d and satisfies certain moment condition, it can be shown that Q is asymptotically chi-squared. According to [25], choosing $m \simeq \ln(T)$ gives better power in simulation studies. Note also that Q arose historically from modification of another statistic and was supposed to increase the power in small sample sizes.

Even though the Ljung–Box statistic is a basic tool and is widely used, it is also known to be rife with problems. For a quick introductory discussion

and the opinion of a practitioner, the reader is referred to [6], Section 4.7 on residual analysis, pp. 67–70. In his words, "[...] these tests have rather poor power properties and in my experience rarely give significant results [...]". The situation is strikingly well-suited for the tools we talked about in the previous two sections.

To check whether or not the residuals ϵ_t are effectively a sample from Gaussian white noise, we propose to use the methodology described in Table II.

TABLE II

Methodology for white noise test.

1. Organize the ϵ_t's in an $n \times p$ matrix, X, in row-major order
2. Compute l_1, the largest eigenvalue of $\tilde{X}'\tilde{X}$, where $\tilde{X}$ is a matrix derived from X(see below)
3. Estimate the variance $\hat{\sigma}^2$ of the ϵ_t's
4. Compare the behavior of l_1 with the one expected under the Tracy–Widom approximation, and use the Tracy–Widom rejection region

The description given in Table II is a little vague so let us make it more precise.

• CHOICE OF n AND p. There are two issues we need to pay heed to. First, it is clear that the procedure we propose will lead to discarding some data. Our first objective is to discard as few data points as possible. So np should be close to T, the total number of observations. Second, the ratio p/n (assuming from now on that $p < n$) will affect the power of our test. This is clear from either Theorems 7 or 8. Note that Theorem 7 essentially asserts that the indifference zone of a test of the type we propose is (in the complex case and finite perturbation of Id setting) $\sqrt{p/n}$. In other words, if the largest eigenvalue of the true covariance matrix is smaller than $1 + \sqrt{p/n}$, the fluctuation of the largest eigenvalue of the empirical covariance matrix is asymptotically indiscernible from its counterpart when the true covariance matrix is Id_p. So we want np close to T and p/n small.

• CHOICE OF $\tilde{X}$. If we kept X and there was correlation in the data, we would be at risk of having correlation across both rows and columns. Since it is unclear at this point how correlation between rows would affect the results, we prefer to be safe and hence we choose to discard, say, one out of two rows when going from X to $\tilde{X}$. This way it is unlikely that for a reasonable (*i.e.* large enough) choice of p, we will encounter correlation within one column. Of course, discarding so many data points is questionable if one does not have much data, but it will not be a problem in the example we give below.

• Estimation of the variance of the ϵ_t's. We can try at least two simple approaches. The first one, and the least favored, would be to compute the empirical variance of the ϵ_t's. This is a problem if one column has significantly more variance than others, or if there are big outliers. A more robust approach (if the user has reasonable confidence in the Gaussian and mean 0 assumptions) is to use a robust estimator of the type $\hat{\sigma} = \text{median}(|\epsilon_t|)/0.6745$.

• Comparison of behavior with Tracy–Widom approximation. Once we have l_1 and $\hat{\sigma}$, we can make a one sided test at level α (=5%, for instance) to check how likely it is that one would observe such a value of l_1 under the null hypothesis that the covariance matrix of the data is Id_p. The choice of the one-sided test is based on the fact that the quality of the Tracy–Widom approximation is really excellent in the upper tail of the distribution, whereas the lower tail match is not as good.

3.1. A worked-out example

Here we present an application of the methodology outlined above on a real dataset. We fit a Gaussian innovation-GARCH (1,1) model to daily Hewlett–Packard log returns and check one aspect of the adequacy of the model by looking closely at the standardized residuals, which, of course, are supposedly i.i.d $\mathcal{N}(0,1)$. The data was found through [25]; it is provided on the book web site, datasets corresponding to Chapter 2. The range is January 1980 to December 1999 and we have 5056 observations. We obtained the standardized residuals through the `Matlab` commands given in the Appendix.

The standardized residuals are plotted in figure 3. Beside a few big outliers, it is visually not clear whether one should reject the null hypothesis that this series is actually i.i.d Gaussian noise. The standard procedure is then to check the Ljung–Box statistic, which we did through the `Matlab` commands given in the Appendix. Based on this analysis, it would seem that the model is adequate: we could not reject the hypothesis that there was no correlation in the data.

Motivated by the doubts raised in [6] and elsewhere, we performed an alternative analysis, using our random matrix methodology. We chose $n = 561$ and $p = 9$, as the Ljung–Box heuristic for choosing m (and described at the beginning of the section) seemed to indicate that $p = 9$ was a reasonable choice. We then discarded the even rows from X to get $\tilde{X}$.

The first diagnostic is then to look at the Wachter plot (first proposed in [26]): we plot the values of the ordered eigenvalues against the quantiles of the Marčenko–Pastur law. If the data were i.i.d, *i.e.* the covariance matrix was Id_9, we should observe a straight line. What we see in figure 4 is, again,

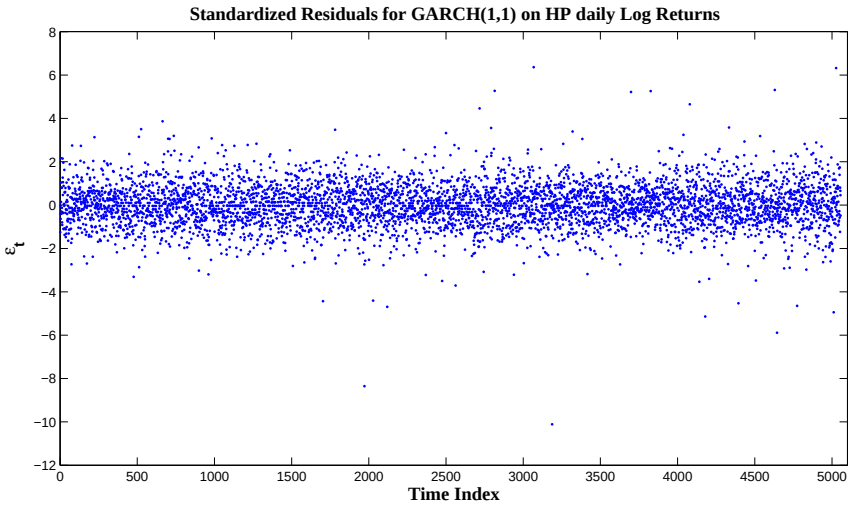

Fig. 3. Standardized Residuals for GARCH (1,1) model on HP Log Returns.

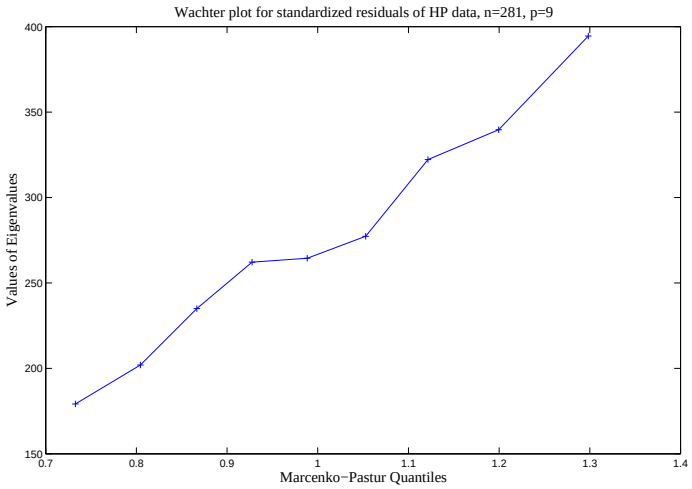

Fig. 4. Wachter plot for the HP residuals data.

hard to judge by Wachter-plot standards, and fluctuation behavior is really needed here to help us decide whether or not the model is acceptable. We therefore proceeded to compute the value of the natural test statistic under the Tracy–Widom approximation. Given the presence of two huge outliers that might have contributed up to 5% of the value of σ^2, we decided to use the median rule to measure $\hat{\sigma}$. We did this knowing that the innovations had to be Gaussian, and so it was consistent with our model building procedure. It also provided an independent "verification" of $\hat{\sigma}$ since the GARCH (1,1)

model forces a value of $\hat{\sigma} \simeq 1$ (using the standard estimate of variance), by construction of the fitting algorithm. We found that, for our data, $\hat{\sigma} = 0.8642$, and hence the "Tracy–Widom" score was 9.9554, while the 99-th percentile of the distribution is only 2.0234. Concerned that we might have been too harsh on the data, we also used the standard estimate of variance, found $\hat{\sigma} = 0.9822$ and a Tracy–Widom score of 1.1895. The 95-th percentile is at .9793, so we again would reject at level $\alpha = 5\%$. We arrived at the same conclusion when we re-centered the columns of the matrix (which, by standard statistical arguments found in [2], p. 76 does not change the nature of the covariance matrix and just requires adjusting the parameters in the test). This is in stark contrast with the conclusions reached by the Ljung–Box statistic.

A thorough statistical analysis would require further investigation about this discrepancy. In the absence of universality result in the case of $\Sigma_p = \mathrm{Id}_p$, we would have to look at departure from normality problems for the standardized residuals. Stationarity issues would also have to be investigated. A detailed statistical analysis is nevertheless not the point of this example; what we saw is that simple random matrix methodology was able to ring a warning bell about the adequacy of the model, in a situation where the classical procedures seemed to indicate that the model was acceptable.

While this is in some sense an isolated data analytic example, we hope that it will help convince the reader of the real world relevance of the results obtained recently about the behavior of the largest eigenvalue of large covariance matrices.

The author wishes to thank the organizers of the conference for their invitation. He is also grateful to Iain Johnstone, David Donoho, Persi Diaconis and Alice Whittemore for help, advice and support at different stages of the projects that are presented here. This work started when the author was a graduate student at the Department of Statistics, Stanford University and was supported in part by NSF grants DMS-0077621, ANI-008584(ITR) and DMS-0140698.

Appendix A

Matlab commands for data analysis

For the sake of reproducibility, we provide the simple commands we used to get the plots (and statistics) mentioned in Section 3. As we said earlier, the raw data is available online.

We got the standardized residuals by using the standard GARCH package in Matlab. The session reads:

```
>> load hpLogReturns
>> [Coeff,Errors,LLF,Innovations,Sigmas,Summary]
= garchfit(hpLogReturns);
>> plot(Innovations./Sigmas,'.')
>> stdRedHP=Innovations./Sigmas;
```

We then computed the Ljung–Box statistics, also using the standard Matlab functions. Here is the command and its output: (an H — read on even lines — of 0 corresponds to non rejection)

```
>> [H,P]=lbqtest(stdRedHP,[6:15]');
>> [H,P]'
ans =
         0         0         0         0         0
    0.1451    0.1776    0.2479    0.3154    0.3455

         0         0         0         0         0
    0.3486    0.4235    0.5041    0.5659    0.6284
```

REFERENCES

[1] T.W. Anderson, *Ann. Math. Statist.* **34**, 122 (1963).

[2] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis, Wiley Series in Probability and Statistics*, third edition, Wiley-Interscience, John Wiley & Sons, Hoboken, NJ, 2003.

[3] Z.D. Bai, *Statist. Sinica* **9**, 611 (1999).

[4] J. Baik, G. Ben Arous, S. Péché, `arXiv:math.PR/0403022`, March 2004.

[5] Z. Burda, J. Jurkiewicz, B. Wacław, *Phys. Rev.* **E71**, 026111 (2005).

[6] Ch. Chatfield, *The Analysis of Time Series, Chapman & Hall/CRC Texts in Statistical Science Series*, sixth edition, Chapman & Hall/CRC, Boca Raton, FL 2004. An introduction.

[7] M. Dieng, `arXiv:math.PR/0411421`, November 2004.

[8] N. El Karoui, `arXiv:math.ST/0309355`, September 2003.

[9] N. El Karoui, `arXiv:math.PR/0409610`, September 2004.

[10] N. El Karoui, `arXiv:math.PR/0503109`, March 2005.

[11] P.J. Forrester, *Nucl. Phys.* **B402**, 709 (1993).

[12] S. Geman, *Ann. Probab.* **8**, 252 (1980).

[13] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning, Springer Series in Statistics*, Springer-Verlag, New York 2001.

[14] K. Johansson, *Commun. Math. Phys.* **209**, 437 (2000).

[15] I.M. Johnstone, *Ann. Statist.* **29**, 295 (2001).

[16] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters. *Phys. Rev. Lett.* **83**, 1467 (1999).

[17] G. Ljung, G.E.P. Box. *Biometrika* **66**, 67 (1978).

[18] V.A. Marčenko, L.A. Pastur, *Mat. Sb. (N.S.)* **72**, 507 (1967).

[19] K.V. Mardia, J.T. Kent, J.M. Bibby, *Multivariate Analysis*, Academic Press, Harcourt Brace Jovanovich Publishers, London 1979.

[20] J.W. Silverstein, *J. Multivariate Anal.* **55**, 331 (1995).

[21] J.W. Silverstein, Z.D. Bai, *J. Multivariate Anal.* **54**, 175 (1995).

[22] A. Soshnikov, *J. Stat. Phys.* **108**, 1033 (2002).

[23] C.A. Tracy, H. Widom. *Commun. Math. Phys.* **159**, 151 (1994).

[24] C.A. Tracy, H. Widom. *Commun. Math. Phys.* **177**, 727 (1996).

[25] R.S. Tsay, *Analysis of Financial Time Series*, Wiley Inter-Science, 2002.

[26] K.W. Wachter, Proceedings of the Computer Science and Statistics 9th Annual Symposium on the Interface, pp. 299–308, 1976.