On the distribution of the largest principal component

Iain M. Johnstone*
Department of Statistics,
Stanford University

August 8, 2000

Abstract

Let $x_{(1)}$ denote square of the largest singular value of an $n \times p$ matrix X, all of whose entries are independent standard Gaussian variates. Equivalently, $x_{(1)}$ is the largest principal component of the covariance matrix $X'X$, or the largest eigenvalue of a p variate Wishart distribution on n degrees of freedom with identity covariance.

Consider the limit of large p and n with $n/p = \gamma \geq 1$. When centered by $\mu_p = (\sqrt{n-1}+\sqrt{p})^2$ and scaled by $\sigma_p = (\sqrt{n-1}+\sqrt{p})(1/\sqrt{n-1}+1/\sqrt{p})^{1/3}$ the distribution of $x_{(1)}$ approaches the Tracy-Widom law of order 1, which is defined in terms of the Painlevé II differential equation, and can be numerically evaluated and tabulated in software. Simulations show the approximation to be informative for n and p as small as 5.

The limit is derived via a corresponding result for *complex* Wishart matrices using methods from random matrix theory. The result suggests that some aspects of large p multivariate distribution theory may be easier to apply in practice than their fixed p counterparts.

1 Introduction

The study of sample covariance matrices is fundamental in multivariate analysis. With contemporary data, the matrix is often large, with number of variables comparable to sample size. In this setting, relatively little is known about the distribution of the largest eigenvalue, or principal component, especially in null cases. A second impetus for this work comes from random matrix theory, a domain of mathematical physics and probability that has seen exciting recent development - for example the long sought asymptotic distribution of the length of the longest increasing subsequence in a random permutation (Baik et al., 1999). Some of these remarkable tools can be borrowed for covariance matrices. A surprise is that the results seem to give useful information about principal components for quite small values of n and p.

Let X be an n by p data matrix. Typically, one thinks of n observations or cases $\vec{x}_i$ of a p dimensional row vector which has covariance matrix Σ. For definiteness, assume that the

*Acknowledgements: Presented as the Laplace lecture at the 5th World Congress of Bernoulli Society and 63rd Annual meeting of the Institute of Mathematical Statistics, Guanajuato, Mexico. For discussions, references and software, many thanks to Persi Diaconis, Sun Holmes, Craig Tracy, Harold Widom, Joe Eaton, Amir Dembo and Peter Bickel. Supported in part by NSF DMS 95–05151 and 00–72661. **Key Words and Phrases.** Karhunen-Loève transform, empirical orthogonal functions, largest eigenvalue, Largest singular value, Laguerre ensemble, Laguerre polynomial, Wishart distribution, Plancherel-Rotach asymptotics, Painlevé equation, Tracy-Widom distribution, Random matrix theory, Fredholm determinant, Liouville-Green method.

rows $\vec{x}_i$ are independent Gaussian $N_p(0, \Sigma)$. In particular, the mean has been subtracted out. If we also don't worry about dividing by n, then we can call $X'X$ a covariance matrix. Under the Gaussian assumptions, $X'X$ is said to have a Wishart distribution $W_p(n, \Sigma)$. If $\Sigma = I$, the "null" case, we call it a white Wishart, in analogy with time series settings where a white spectrum is one with the same variance at all frequencies.

Large sample work in multivariate analysis has traditionally assumed that n/p, the number of observations per variable, is large. Today it is common for p to be large or even huge, and so n/p may be moderate to small, and in extreme cases less than one. For example,

a) *Climate studies:* n might be the number of time points, and p the number of observation stations. Principal components analysis is widely used under the name "empirical orthogonal functions". Preisendorfer (1988) is a book length treatment that emphasises n/p moderate.

b) *Financial data:* large covariance and correlation matrices, with $p \approx 400$ financial indicators, are publicly posted daily (e.g. `riskmetrics.com`) and used for value-at-risk calculations.

c) *IR/search engines:* A common search engine strategy forms huge term by document incidence matrices (n and p at least in the thousands) and then does a truncated singular value decomposition. This example is far from the Gaussian, but illustrates the huge matrices that arise.

d) *Functional data analysis:* Each data point is a curve, and so typically high dimensional. In the example of Figure 1, extracted from Buja et al. (1995), a small speech dataset consists 162 instances of a phoneme "dcl" spoken by about 50 males. Each instance is calculated as a periodogram on 256 points. So here $n = 162$, and $p = 256$.

Basic notation and phenomena. The eigenvalue-eigenvector decomposition of the sample covariance matrix S

$$S = X'X = ULU' = \sum l_j u_j u_j',$$

with eigenvalues in the diagonal matrix L and orthonormal eigenvectors collected as the columns of U. There is a corresponding decomposition of the population covariance matrix $\Sigma = \Upsilon \Lambda \Upsilon'$ with eigenvalues λ_i.

A basic phenomenon is that the sample eigenvalues l_i are more spread out than the population λ_i. This effect is strongest in the null cases when all population eigenvalues are the same. As a simple example, consider one random draw of a 10×10 matrix X with i.i.d $N(0, 1)$ entries. The ordered sample eigenvalues l_i of S were:

$$(l_j) = 3.07,\ 1.40,\ 1.12,\ .78,\ .51,\ .30,\ .16,\ .095,\ .036,\ .003$$

In this case, extreme because $n = p$, the ratio of largest to smallest is about 1000.

While our focus is on eigenvalues, there are two essentially equivalent points of view that are of importance. In the *singular value decomposition* of a data matrix $X = UDV'$, the singular values $D = \mathrm{diag}(d_j)$ are just the square roots of the eigenvalues l_j. The condition number of X is just the ratio of largest to smallest singular value. The distribution of the smallest singular value was given in certain cases by Edelman (1988) and studied in detail by Forrester (2000).

Eigenvalues also occur in principal components analysis, also widely known as the Karhunen-Loève transform. One seeks the successively orthogonal directions that maxi-

mally explain the variation in the data. In this case,

$$l_j = \max\{\frac{u'Su}{u'u} : u \perp u_1, \ldots, u_{j-1}\}$$

Here a key practical question emerges: how many principal components should be retained as being 'significant'?

The "scree plot" (e.g. Mardia et al. (1979)) is one of the many graphical and informal methods that have been proposed. One plots the ordered sample eigenvalues or singular values, and looks for an "elbow", or other break between presumably significant and presumably unimportant components. In the phoneme example, Figure 1(c) there are clearly three large values, but what about the fourth, fifth etc.?

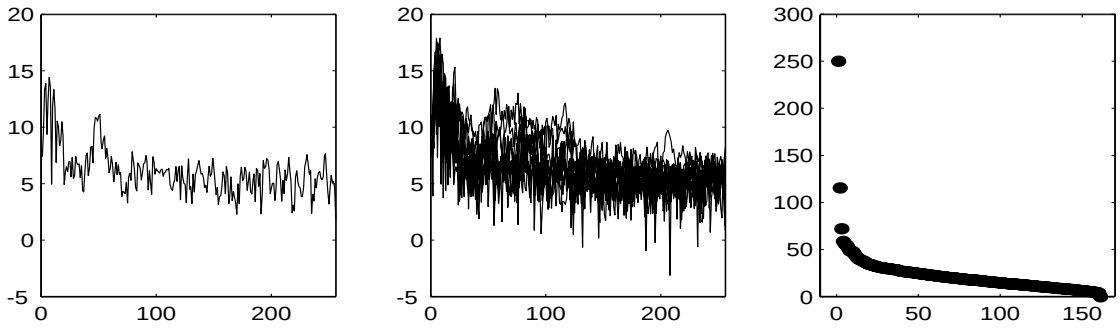

Figure 1: Panel (a): a single instance of a periodogram from the phoneme dataset, (b) ten instances, to indicate variability, (c) Screeplot of eigenvalues in phoneme example

In the study of eigenvalue distributions, two general areas can be distinguished: the *bulk*, which refers to the properties of the full set $l_1 > l_2 \ldots > l_p$, and the *extremes*, which addresses the (first few) largest and smallest eigenvalues. To provide context for later results, here is a brief and necessarily selective account.

1.1 Bulk spectrum

For square symmetric random matrices, the celebrated semi-circle law of Wigner (1955, 1958) describes the limiting density of eigenvalues. There is an analog for covariance matrices due to Marčenko and Pastur (1967), and independently, Stein (1969).

The Marčenko-Pastur result is stated here for Wishart matrices with identity covariance $\Sigma = I$, but is true more generally, including non-null cases. Suppose that both n and p tend to ∞, in some ratio $n/p \to \gamma \geq 1$. Then the empirical distribution of the eigenvalues converges almost surely:

$$G_p(t) = \frac{1}{p}\#\{l_i : l_i \leq nt\} \to G(t) \tag{1.1}$$

and the limiting distribution has a density $g(t) = G'(t)$:

$$g(t) = \frac{\gamma}{2\pi t}\sqrt{(b-t)(t-a)}, \quad a \leq t \leq b. \tag{1.2}$$

where $a = (1 - \gamma^{-1/2})^2$ and $b = (1 + \gamma^{-1/2})^2$. Compare Figure 2(a). Thus, the smaller n/p, the more spread the eigenvalues – even asymptotically, the spread of the empirical

eigenvalues doesn't disappear. For $n = p$, the largest normalized eigenvalue approaches 4, and the smallest approaches 0, which accounts for the large spread seen in the 10×10 example earlier. There has been a significant literature rediscovering and extending this theorem, with contributions, among others, by Bai, Grenander, Jonsson, Krishnaiaih, Silverstein, Wachter, and Yin. Bai (1999) provides details in a comprehensive recent survey.

Wachter (1976) advocated a nice data-analytic use of this result to yield a simple, but informative modification of the screeplot: make a probability plot of the ordered observed eigenvalues l_{p+1-i} against the quantiles $G^{-1}(\frac{i-1/2}{p})$ of the predicted "semi-circle" type distribution (1.1) - (1.2). Figure 2(b) shows the phoneme data (actually on the singular value scale). One sees the three large values as before, but it is notable that the bulk of the distribution follows the semi-circle law. There is an uptick at the right hand edge, which looks like there is extra variance in the directions corresponding to the 4th through 12th eigenvalues. Without variability information on the null distribution, one cannot say with rigor whether this is real.

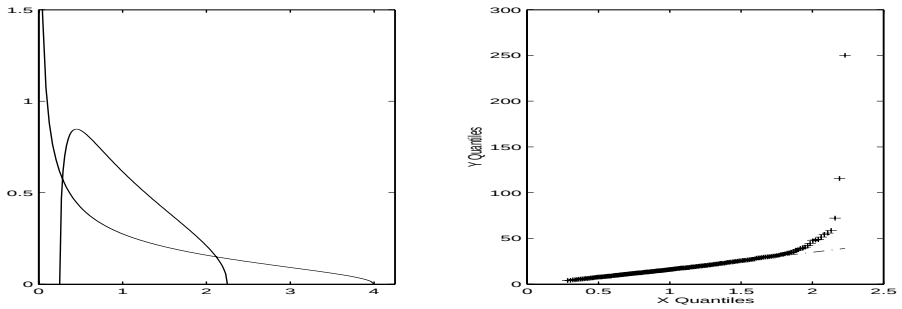

Figure 2: Panel (a). Limiting densities (1.2) corresponding to $n = 4p, \gamma = 4$ and $n = p, \gamma = 1$ (monotone line). Panel (b) Wachter plot of the empirical singular values of the phoneme data.

1.2 Largest Eigenvalue

Consider now the right hand edge, and particularly the largest eigenvalue. Why the interest in extremes? In the estimation of a sparse mean vector, the maximum of n i.i.d Gaussian noise variables plays a key role. Similarly, in distinguishing a "signal subspace" of higher variance from many noise variables, one expects the largest eigenvalue of a null (or white) sample covariance matrix to play a basic role.

The bulk limit (1.2) points to a strong law for the largest eigenvalue. Indeed, Geman (1980) shows that

$$n^{-1} l_1 \to (1 + \gamma^{-1/2})^2 \quad \text{a.s.},$$

i.e. $l_1 \sim (\sqrt{n} + \sqrt{p})^2$. Later Bai, Krishnaiah, Silverstein and Yin established that strong convergence occurred iff the parent distribution had zero mean and finite 4th moment. Bai (1999) has more details, and results on smallest eigenvalue.

However these results say nothing about the variability of the largest eigenvalue, let alone about its distribution. Muirhead (1982, Sec.9.7) surveys existing results. For example, Constantine (1963, p.1284) gives an exact expression in terms of a zonal polynomial series for a matrix hypergeometric function:

$$P(l_1 \le nt) = d_{p,n} t^{pn/2} {}_1F_1(\tfrac{1}{2}n; \tfrac{1}{2}(n + p + 1); -\tfrac{1}{2}ntI_p),$$

[cf. also Muirhead (1982, p. 421)]. There are explicit evaluations for $p = 2, 3$ (Sugiyama, 1972), but in general the alternating series converges very slowly, even for small n and p, and so is difficult to use in practice. For *fixed* p and large n, the classic paper by Anderson (1963) gives the limiting joint distribution of the roots, but the marginal of l_1 is hard to extract in the null case $\Sigma = I$. Muirhead (1974) gives a series approximation again for $p = 2, 3$. In general, there are upper bounds on the d.f. using p independent $\chi^2_{(n)}$. Overall, there is little that helps numerically with approximations for large p.

1.3 Main result

We now turn to what can be derived from random matrix theory (RMT) methods. Suppose that $X = (X_{jk})_{n \times p}$ has entries which are i.i.d. $X_{jk} \sim N(0, 1)$. Denote the sample eigenvalues of the Wishart matrix $X'X$ by $l_1 > \ldots > l_p$. Define center and scaling constants

$$\mu_{np} = (\sqrt{n-1} + \sqrt{p})^2, \tag{1.3}$$

$$\sigma_{np} = (\sqrt{n-1} + \sqrt{p}) \big(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}} \big)^{1/3}. \tag{1.4}$$

The *Tracy-Widom* law of order 1 has distribution function defined by

$$F_1(s) = \exp\{-\tfrac{1}{2} \int_s^\infty q(x) + (x-s) q^2(x) dx\}, \qquad s \in \mathbb{R},$$

where q solves the (non-linear) Painlevé II differential equation

$$\begin{cases} q''(x) = x q(x) + 2q^3(x) & \\ q(x) \sim \mathrm{Ai}(x) & \text{as } x \to +\infty, \end{cases} \tag{1.5}$$

and $\mathrm{Ai}(x)$ denotes the Airy function. This distribution was found by Tracy and Widom (1996) as the limiting law of the largest eigenvalue of an n by n Gaussian symmetric matrix.

The main result states that the same limiting distribution applies to covariance matrices XX' derived from *rectangular* data matrices X where both dimensions n and p are large. In fact, we assume that $n = n(p)$ increases with p in such a way that both μ_{np} and σ_{np} are increasing in p.

Theorem 1.1. *Under the above conditions, if* $n/p \to \gamma \geq 1$, *then*

$$\frac{l_1 - \mu_{np}}{\sigma_{np}} \xrightarrow{\mathcal{D}} W_1 \quad \sim F_1$$

The theorem is stated for situations in which $n \geq p$. However, it applies equally well if $n < p$ are both large, simply by reversing the roles of n and p in (1.3) – (1.4). We write $TW_1(n, p)$ for the law of $\mu_{np} + \sigma_{np} W_1$, which we use to approximate the distribution of l_1.

The mean growth of l_1 is as described earlier, except for a slight adjustment in (1.3) which is suggested by the proof and makes a distinct improvement to the quality of approximation for small n.

The scale behavior is noteworthy. A sum of i.i.d variables with positive mean grows with mean of order n and standard deviation of order $\sqrt{n}$. Here, the lead eigenvalue of a $p \times p$ Wishart grows with mean of order p, but with SD about that mean only of order $p^{1/3}$. Thus its distribution is relatively much more tightly clustered about its mean than in the case of sums.

From numerical work, Tracy and Widom (2000) report that the F_1 distribution, plotted in Figure 3, has mean $\doteq -1.21$, and $SD \doteq 1.27$. The density is asymmetric and Appendix A1 shows that its left tail has exponential order of decay like $e^{-|s|^3/24}$, while its right tail is of exponential order $e^{-\frac{2}{3}s^{3/2}}$.

Numerical table look up for this distribution is analogous to using the traditional statistical distributions such as chi-square, normal or F. Work is in progress to provide publicly downloadable `MATLAB` and `S-PLUS` routines for the Tracy-Widom cumulative, density and quantile functions.

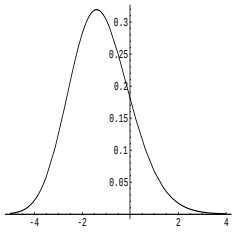

Figure 3: Density of the Tracy-Widom distribution F_1

Remarks. 1. The equation (1.5) is one of the six exceptional differential equations identified by Painlevé at the end of the 19th century. Further detail may be found in the expository article Deift (1999*a*).

2. An heuristic argument derives the $p^{1/3}$ scaling for l_1 from the bulk density (1.2). Indeed, let $f(t)$ denote the limiting density of $t_i = b - p^{-1}l_i$ (where $\asymp$ denotes exact order behavior). Suppose that the spacing of the smallest few t_i is of order $p^{-\alpha}$, so that the interval $[0, p^{-\alpha})$ contains $O(1)$ out of p eigenvalues. The cumulative of f, namely F, then satisfies $F(p^{-\alpha}) \asymp p^{-1}$. But from (1.2) we have $f(t) \asymp t^{1/2}$ for $t \to 0$, so also $F(p^{-\alpha}) \asymp p^{-3\alpha/2}$. This fixes $\alpha = 2/3$, so that the largest few l_i should have spacing $p \cdot p^{-2/3} = p^{1/3}$.

Theorem 1.1 relies heavily on the RMT literature. But first we focus on some of its statistical implications.

1.4 Statistical Implications

Quality of approximation for moderate n and p. As a check on the practical applicability of Theorem 1.1, some simulations were done, first for square cases $n = p = 5, 10$ and 100, using $R = 10,000$ replications, with results shown in Table 1 and Figure 4.

Even for 5×5 and 10×10, the approximation seems to be quite good in the right hand tail at conventional significance levels of 10%, 5% and 1%. At 100×100, the approximation seems to be reasonable throughout the range.

The same general picture holds for n/p in the ratio $4:1$. Even for 5×20 matrices, the approximation is reasonable, if not excellent, at the conventional upper significance levels.

A further summary message from these computations is that in the null Wishart case, about 80 % of the distribution lies below μ_{np}, and 95 % below μ_{np} plus one σ_{np}.

Non-null cases: empirical results. In practice, as in the phoneme example, there are often one or more large eigenvalues clearly separated from the bulk. This raises the question: if there were, say, only one or a small number of non-unit eigenvalues in the population, would they pull up the other values? Consider, therefore, a "spiked" covariance model,

Percentile	TW	**5×5**	**10×10**	**100×100**	**5×20**	**10×40**	**100×400**	2 * SE
-3.90	.01	.000	.001	.007	.002	.003	.010	(.002)
-3.18	.05	.003	.015	.042	.029	.039	.049	(.004)
-2.78	.10	.019	.049	.089	.075	.089	.102	(.006)
-1.91	.30	.211	.251	.299	.304	.307	.303	(.009)
-1.27	.50	.458	.480	.500	.539	.524	.508	(.010)
-0.59	.70	.697	.707	.703	.739	.733	.714	(.009)
0.45	**.90**	**.901**	**.907**	**.903**	**.919**	**.918**	**.908**	(.006)
0.98	**.95**	**.948**	**.954**	**.950**	**.960**	**.961**	**.957**	(.004)
2.02	**.99**	**.988**	**.991**	**.991**	**.992**	**.993**	**.992**	(.002)

Table 1: First column shows the probabilities of the F_1 limit distribution corresponding to fractions in second column. Next three columns show estimated cumulative probabilities for l_1 in $R = 10,000$ repeated draws from $W_p(n, I)$ with $n = p = 5, 10$ and 100. The following three cases have $n : p$ in the ratio $4 : 1$. Final column gives approximate standard errors based on binomial sampling. Bold font highlights some conventional significance levels. The Tracy-Widom distribution F_1 was evaluated on a grid of 121 points $-6(.1)6$ using the `Mathematica` package `p2Num` written by Craig Tracy. Remaining computations were done in `MATLAB`, with with percentiles obtained by inverse interpolation, and using `randn()` for normal variates and `norm()` to evaluate largest singular values.

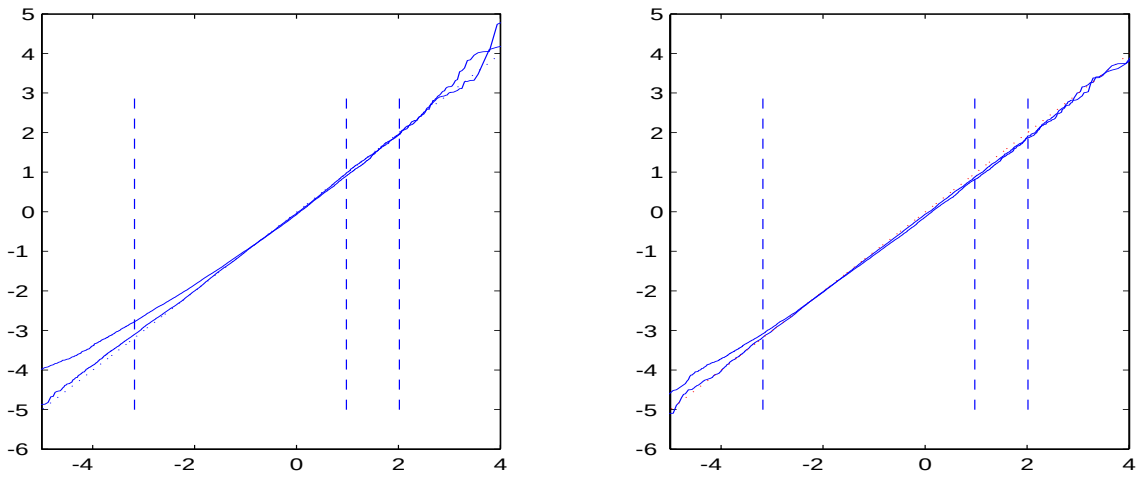

Figure 4: First panel: probability plots of $R = 10,000$ observed replications of l_1 drawn from $W_p(n, I)$ for $n = p = 10$ and 100. That is, the $10,000$ ordered observed values of l_1 are plotted against $F_1^{-1}((i - .5)/R), i = 1, \ldots, R$. The line for $n = p = 10$ is the one elevated in the left tail. The vertical dashed lines show 5th, 95th and 99th percentiles. The dotted line is the 45 degree line of perfect agreement of empirical law with asymptotic limit. Second panel: same plots for $n = 40, p = 10$ and $n = 400, p = 100$.

with a fixed number, say r, eigenvalues greater than 1:

$$\Sigma_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_r^2, 1, \ldots, 1). \tag{1.6}$$

In fact, the $(r+1)^{\mathrm{st}}$ eigenvalue in the spiked model is stochastically smaller than the *largest* eigenvalue in the null model with $p - r$ variables.

Proposition 1.2. $\mathcal{L}(\, l_{r+1} \,|n, p, \Sigma_\tau) \overset{st}{<} \mathcal{L}(\, l_1 \,|n, p-r, I_{p-r})$.

This non-asymptotic result follows directly from interlacing properties of eigenvalues (Appendix A2). Use of $\mathcal{L}(\, l_1 \,|n, p-r, I_{p-r})$ provides a conservative $p-$value for testing $H_0 : \tau_{r+1}^2 = 1$ in the spiked model, in the sense that the correct $p-$value for l_{r+1} is always smaller. And Theorem 1.1 and the simulations say that the $p-$value for $\mathcal{L}(\, l_1 \,|n, p-r, I_{p-r})$ can be numerically approximated using $TW_1(n, p-r)$.

In fact, empirical evidence from Figure 5 below, suggests that if τ_r^2 well separated from 1, then the distributions in Proposition 1.2 are approximately shifted by a constant $c = c_{n,p,\tau}$:

$$\mathcal{L}(\, l_{r+1} \,|n, p, \Sigma_\tau) \approx \mathcal{L}(\, l_1 \,|n, p-r, I_{p-r}) - c_{n,p,\tau}.$$

In the phoneme data example, of particular interest is a spiked model corresponding $p = 161$ and the three obviously separated singular values. Thus, for Σ_τ, set the top values τ_1, τ_2, τ_3 equal to the observed data, and the remaining τ_i all equal to one. The dashed line is corresponds to the fourth largest eigenvalue from this distribution, and the solid one to the largest eigenvalue from the 158 x 256 white case.

In addition, Proposition 1.2 shows that the 4th (and presumably down to 12th) eigenvalues in the data are significantly larger than would be expected under the null model. In other words, this uptick in the Wachter plot is a genuine departure from equal variances.

PCA on correlations. Principal components analysis is not invariant to changes in scale of the variables. For this reason, it is often recommended that principal components be calculated after standardizing the variables to have unit standard deviation. Equivalently, one performs PCA on the sample correlation matrix. This is problematic for distribution theory (e.g. Anderson (1963)). In particular, Theorem 1.1 does not directly apply.

An *ad hoc* construction may, however, be based on the Tracy-Widom approximation. Let the $n \times p$ data matrix X consist of n i.i.d draws of a row vector $\vec{x}_i$ with distribution $N_p(0, \Sigma)$. View the data matrix in terms of its columns: $X = [x_1 \cdots x_p]$ and standardize:

$$w_j = \frac{x_j}{s_j}, \qquad s_j = \|x_j\|.$$

Set $W = [w_1 \cdots w_p]$. Performing PCA on $W'W$ amounts to PCA on the sample correlations of the original data, with population correlations $R = (\rho_{jk})$.

To create a test of $H_0 : R = I$ based on W, observe that under H_0, the vectors w_j are i.i.d. on the unit sphere S^{n-1}. Now synthesize a standard Gaussian data matrix $\tilde{X} = [\tilde{x}_1 \cdots \tilde{x}_p]$ by multiplying each w_j by an independent chi-distributed length:

$$\tilde{x}_j = r_j w_j \qquad r_j^2 \overset{indep}{\sim} \chi_n^2.$$

Under H_0 the Tracy-Widom approximation then applies to the largest sample eigenvalue of $\tilde{S} = \tilde{X}'\tilde{X}$: so that $l_1(\tilde{S})$ has approximately the $TW_1(n, p)$ distribution.

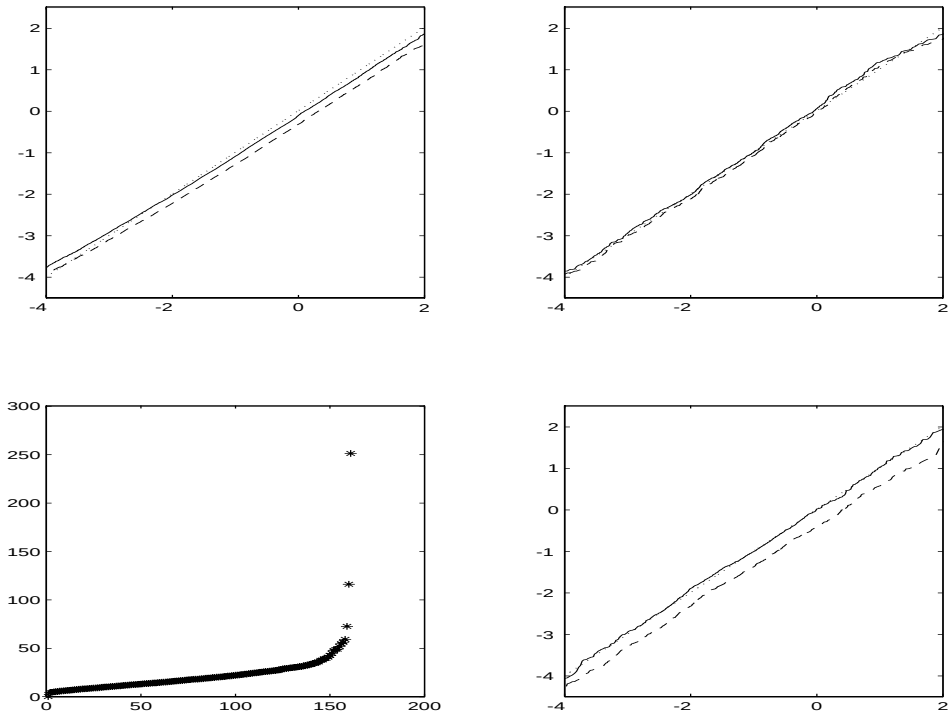

Figure 5: (a) 10 unit roots and one with $\tau = 10$ in model (1.6). $p = 10, n = 40$, with $N = 10,000$ replications. The dashed line is a qqplot of the second largest value against the TW distribution. For comparison, the solid line is the simulated null distribution for 10×40 white Wishart case. The two lines essentially differ by a vertical shift. Dotted line is the 45 degree line of perfect agreement. (b) 99 unit roots and one with $\tau = 10$ in model (1.6). $N = 1,000$ replications. Dashed line is second largest from this distribution, and the solid is the 100×100 white case. (c) Singular values of phoneme data $n = 256, p = 161$. (d) Dashed line is qqplot of fourth largest eigenvalue from spiked covariance model with top three values set at the observed values in the phoneme data. Solid line is qqplot of largest eigenvalue from a null Wishart with $p = 158$ and $n = 256$. $N = 1,000$ replications.

1.5 Complex Matrices

We now return to the main theorem. Data matrices with *complex* Gaussian entries are also of interest in statistics and signal processing. Suppose now that $X = (X_{ij})_{n\times p}$, with $\mathrm{Re}X_{ij}$, $\mathrm{Im}X_{ij} \sim N(0, \frac{1}{2})$ all independently of one another. The matrix $S = X^*X$ has the complex Wishart distribution, and we again suppose that its (real) eigenvalues are ordered $l_1 > \ldots > l_p$.

While studying a random growth model of interest in probability, Kurt Johansson (2000) stated without detailed proof a limit theorem which could be reinterpreted as giving the limit behavior of the largest eigenvalue of a complex Wishart matrix. First, define slight modifications of (1.3) - (1.4):

$$\mu'_{np} = (\sqrt{n} + \sqrt{p})^2,$$
$$\sigma'_{np} = (\sqrt{n} + \sqrt{p})\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}}\big)^{1/3}.$$

Assume that $n = n(p)$ increases with p so that both μ'_{np} and σ'_{np} are increasing in p.

Theorem 1.3 (Johansson, 2000). *Under the above conditions, if $n/p \to \gamma \geq 1$ then*

$$\frac{l_1 - \mu'_{np}}{\sigma'_{np}} \xrightarrow{\mathcal{D}} W_2 \quad \sim F_2$$

The center and scale are essentially the same as in the real case (but see Remark 4.1), however the limit distribution is now

$$F_2(s) = \exp\{-\int_s^\infty (x-s)q^2(x)dx\}, \tag{1.7}$$

where q is still the Painlevé II function defined at (1.5). This distribution was also first found by Tracy and Widom in the Wigner matrix setting to be recalled below.

Remark. The definition (1.7) implies $\frac{d^2}{ds^2}\log F_2(s) = -q^2(s)$, and $q^2(s)$ is monotone decreasing, asymptotic to $|s|/2$ as $s \to -\infty$, and to $e^{-\frac{4}{3}s^{2/3}}/4\pi\sqrt{s}$ as $s \to \infty$. (Appendix A1). This should be compared with the Pearson family of distributions (which contains most of the familiar laws) and the extreme value family, given respectively by

$$\frac{d}{ds}\log f(s) = -\frac{a+s}{c_0 + c_1 s + c_2 s^2}, \qquad \text{and}$$
$$\log F(s) = -e^{-s}; \text{or } -|s|^k.$$

1.6 Remarks on Proof and Outline of Paper

A central problem in random matrix theory has been the study of so-called Wigner matrices. In the real case, these are symmetric with i.i.d. elements, up to modifications on the diagonal. For example, $p \times p$ Gaussian Wigner matrices have

$$Y = Y' \qquad Y_{jk} \overset{ind}{\sim} N(0, \tfrac{1}{2} + \tfrac{\delta_{jk}}{2}), \qquad j \leq k.$$

In fact, it is simpler to derive results for complex matrices first. A complex Gaussian Wigner matrix is Hermitian, $Y = Y^*$, with Y_{jj} and $\mathrm{Re}Y_{jk}$, $\mathrm{Im}Y_{jk}, j < k$ all i.i.d. $N(0, \frac{1}{2})$. It was here that Tracy and Widom (1994) identified the F_2 distribution, and showed that

$$W_p = \frac{l_1 - 2\sqrt{p}}{p^{1/6}} \xrightarrow{\mathcal{D}} F_2.$$

This was based on determinant representations of the distributions for finite p, followed by asymptotics of Hermite polynomials $H_p(\mu_p + \sigma_p s)$, where $\mu_p = 2\sqrt{p}$ and $\sigma_p = p^{1/6}$. Finally, Tracy and Widom showed that the limiting distribution satisfied the Painleve II equation.

Later, Tracy and Widom (1996) derived the F_1 distribution for real matrices as a consequence of the complex case. One cannot do justice to their method in a single sentence, but essentially they derive the *square* of F_1 as a low rank perturbation of the complex setting.

For the covariance matrices $S = X'X$ studied in this paper, the same broad strategy works. Although the result for complex matrices is again the easier to derive, the conclusion for real matrices is what is needed for most statistical settings. A new aspect, also important in statistics, is that in the $n \times p$ data matrix X, both p and n are separately large, which leads to nonstandard asymptotics in Laguerre polynomials $L_p^{\alpha}(\mu_{np} + \sigma_{np} s)$ – large degree p, large order $\alpha = n - p$ and argument near the largest zero. Along the way, we give a separate proof of Johansson's result, based on Liouville-Green asymptotics of differential equations, rather than steepest descent methods applied to contour integrals.

Section 2 first describes some of the remarkable determinant formulas that have been developed in RMT and then makes some heuristic remarks about the Laguerre asymptotics. Section 3 assembles the operator theoretic tools need to complete our alternate proof of the complex case, Theorem 1.3. Section 4 establishes the main result, Theorem 1.1 by sketching how the arguments of Tracy and Widom (1996) are extended from the Gaussian to the Laguerre ensemble. Section 5 gives details of the Laguerre polynomial asymptotics.

1.7 Discussion

The main conclusion of this paper is that the Tracy-Widom distribution F_1 provides a useable numerical approximation to the null distribution of the largest principal component from Gaussian data even for quite moderate values of n and p. In particular, we have the following simple approximate rules of thumb:

1) about 83% of the distribution is less than $\mu_{np} = (\sqrt{n-1} + \sqrt{p})^2$;

2) about 95% and 99% lie below $\mu_{np} + \sigma_{np}$, and $\mu_{np} + 2\sigma_{np}$ respectively.

A second important conclusion is that in non-null cases in which the population covariance Σ has precisely r eigenvalues greater than 1, then the distribution of the $(r+1)$st sample eigenvalue l_{r+1} can be approximately bounded above by the Tracy-Widom law appropriate to an $n \times (p-r)$ matrix, leading to approximately conservative $P-$values.

Practical problems of data analysis often have covariance matrices with much more structure than assumed here. It may be, however, that such structure can be decomposed into subparts to which the Tracy-Widom approximation is relevant. For example, separate spherical Gaussian models can be appropriate for subsets of coefficients in an orthogonal wavelet decomposition of non-white Gaussian process data. In this respect, it is encouraging that the sizes of the subparts would not need to be particularly large. Specific examples must, however, await future work.

This paper raises other issues for future work, among which we mention:

a) what happens if the elements of the data matrix X are i.i.d. from a non-Gaussian distribution? Soshnikov (1999) established "universality" of the Tracy-Widom limit for square Wigner matrices. Does the same hold for $X'X$ for large n and p? Preliminary simulations are encouraging, though the quality of approximation (if not the actual rate of convergence in n and p) naturally appears to depend on the specific parent distribution.

b) RMT has a formalism for deriving the distribution of the kth largest eigenvalue (k fixed) - see for example Tracy and Widom (1998, 1994), where the latter carries this

through for Gaussian Hermitian Wigner matrices. Presumably this can be extended to null Gaussian covariance matrices.

c) Many techniques of classical multivariate analysis are based on the roots of determinantal equations such as $\det[A_1 - l(A_1 + A_2)] = 0$, with $A_i \sim W_p(n_i, I)$ independently. Thus, one might ask, for example, for an approximation to the distribution of the largest canonical correlation when p, n_1 and n_2 are large.

2 Determinant Formulas, Laguerre heuristics

The joint density of the latent roots of a real Wishart matrix was found in 1939, in a remarkable instance of simultaneous publication, independently by each of Fisher, Girshick, Hsu, Mood and Roy – see Wilks (1962) for citations and Section 4 below. For reasons just given, we start with the complex version of this density. So, let X be an $n \times N$ complex normal matrix with $\mathrm{Re}\, X_{ij}$, $\mathrm{Im}\, X_{ij}$ all independently and identically distributed as $N(0, \frac{1}{2})$ (Eaton, 1983). The cross products matrix X^*X then has the complex Wishart distribution with identity covariance matrix. The eigenvalues $x = (x_1, \ldots, x_N)$ of X^*X are real and non-negative, and have density (James, 1964)

$$P_N(x_1, \ldots, x_N) = c_{N,n}^{-1} \prod_{1 \le j < k \le N} (x_j - x_k)^2 \prod_{j=1}^{N} x_j^{\alpha} e^{-x_j}, \qquad \alpha = n - N. \tag{2.1}$$

[Warning! Notation change: to accord with the RMT literature, we henceforth write the sample eigenvalues as $x_1 > \ldots > x_N$, rather than $l_1 > \ldots > l_p$.]

Efforts to use density (2.1) directly to get information on the largest eigenvalue are frustrated by the high dimensionality and the Jacobian term $\prod_{j<k}(x_j - x_k)^2$. Random matrix theory (RMT) addresses this by starting with the Vandermonde determinant identity

$$\prod_{1 \le j,k \le p} (x_j - x_k) = \det_{1 \le j,k \le N} [x_j^{k-1}]. \tag{2.2}$$

One sees polynomials in each of the eigenvalues with degrees up to $N - 1$. Let $w(x) = x^{\alpha} e^{-x}$ be the Laguerre weight function, and $\phi_k(x) = \phi_k(x; \alpha)$ be functions obtained by orthonormalizing the sequence $x^k w^{1/2}(x)$ in $L^2(0, \infty)$: in fact

$$\phi_k(x) = \sqrt{\frac{k!}{(k+\alpha)!}} x^{\alpha/2} e^{-x/2} L_k^{\alpha}(x) \tag{2.3}$$

where $L_k^{\alpha}(x)$ are the Laguerre polynomials, defined as in, e.g. Szegö (1967).

A standard argument [e.g. Mehta (1991, Ch. 5), Deift (1999*b*, Ch. 5.)] yields a remarkable determinant representation for the joint density

$$P_N(x_1, \ldots, x_N) = \frac{1}{N!} \det_{1 \le j,k \le N} S(x_j, x_k),$$

where the bilinear kernel $S = S_N$ is given by

$$S(x, y) = \sum_{k=0}^{N-1} \phi_k(x) \phi_k(y). \tag{2.4}$$

A kernel $A(x,y)$ defines an operator A on functions g as usual via $(Ag)(y) = \int A(x,y)g(y)dy$. For suitable functions f, denote by Sf the operator with kernel $S(x,y)f(y)$. A further key RMT formula [e.g. Tracy and Widom (1998), Deift (1999*b*))], valid in particular for (2.1), states that

$$E_N \prod_{j=1}^{N}[1 + f(x_j)] = \det(I + Sf), \tag{2.5}$$

where the right side is a Fredholm determinant of the operator Sf [Riesz and Sz.-Nagy (1955), Gohberg and Krein (1969, Ch. 4.)].

The choice $f = -\chi$, where $\chi = I\{x : x \geq t\}$, yields the determinantal expression for the distribution of $x_{(1)}$:

$$F_{N2}(t) = P\{\max_{1\leq j\leq N} x_j \leq t\} = \det(I - S\chi). \tag{2.6}$$

[The subscript 2 recalls the exponent 2 in (2.1). In fact $F_{N2} = F_{N,n,2}$ depends on n also, but this will not always be shown explicitly.]

Tracy and Widom showed that these operator determinants satisfy differential equations which in the large N limit involve the Painlevé functions. We refer to Tracy and Widom (1994, 1996) (and the expositions in Tracy and Widom (1999, 2000) for the full story, and turn to indicating where the centering and scaling constants come from, as well as the Airy functions.

Laguerre Heuristics. Consider the mode x^* of the density P_N, with components $x_1^* > x_2^* > \ldots > x_N^*$ in decreasing order. According to Stieltjes' "electrostatic interpretation," (Szegö, 1967, p. 141) these components are precisely the zeros $L_N^{\alpha-1}(x_i^*) = 0$ of the degree N Laguerre polynomial of order $\alpha - 1$. Thus, the "typical" positions of the eigenvalues correspond to the zeros of orthogonal polynomials, and in particular the largest eigenvalue should be sought in the neighborhood of the largest zero of $L_N^{\alpha-1}$. The largest zero of an orthogonal polynomial of high degree marks a transition in the behavior of the polynomial from oscillation ($x < x_1^*$) to rapid growth ($x > x_1^*$). In turn, this can be studied using the differential equations that characterize the Laguerre polynomials.

Begin with the Airy equation: $\mathrm{Ai}''(\zeta) = \zeta \mathrm{Ai}(\zeta)$ – this has a *turning point* at 0: ζ positive corresponds to the region of exponential behaviour of the solution and ζ negative to the oscillatory zone.

Figure 6 shows that an appropriately weighted Laguerre polynomial $w_N(x/\kappa) = x^{(\alpha+1)/2}e^{-x/2}L_N^{\alpha}(x)$ looks very similar to the Airy function near largest zero: the polynomial passes from oscillation to exponential decay (because of e^{-x} damping factor). Although the similarity does not extend throughout the range, it is the neigborhood of the largest zero that is of interest here. Note also the relatively small values $n = N = 10$.

The center and scaling constants μ_N, σ_N arise through aligning these two functions. Laguerre polynomials and hence the function w_N satisfy a second order differential equation (e.g. Szegö (1967, Ch. V) and Section 5 below). We write this in a form with a large parameter $\kappa = N + \frac{\alpha+1}{2}$:

$$w_N''(\xi) = [\kappa^2 f(\xi) + g(\xi)]w,$$

where $\xi = x/\kappa$ and

$$f(\xi) = \frac{(\xi - \xi_1)(\xi - \xi_2)}{4\xi^2}, \qquad g(\xi) = -\frac{1}{4\xi^2}.$$

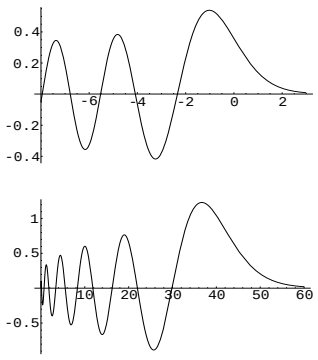

Figure 6: Top panel: Airy function $\mathrm{Ai}(\zeta)$. Lower panel: Laguerre polynomial $w_N(x/\kappa) = x^{(\alpha+1)/2}e^{-x/2}L_N^\alpha(x)$ as a function of x. Here $n = N = 10$.

The function $f(\xi)$ has upper turning point at $\xi_2 = 2 + \sqrt{4 - (\alpha/\kappa)^2}$ and it is easily verified that $\mu_N \sim (\sqrt{n} + \sqrt{N})^2$. On the x scale, this occurs at $x = \kappa\xi_2 = \mu_N$, and this fixes the centering constant. Now, transform the independent variable ξ into ζ so that if g is ignored, then the Laguerre equation turns into the Airy equation, using $\zeta^{1/2}d\zeta = f^{1/2}(\xi)d\xi$. Then w_N can be approximated by the Airy function in the new variable ζ:

$$w_N(\xi) = c\mathrm{Ai}(\kappa^{2/3}\zeta) + \text{ error.}$$

On the x scale, we are interested in values $x = \mu_N + \sigma_N s$ close to the turning point. On the ξ scale, $\xi = \xi_2 + (\sigma_N/\kappa)s$ and we linearize $\zeta(\xi)$ about $\zeta(\xi_2) = 0$. The scale σ_N is then chosen so that $\kappa^{2/3}\zeta(\xi) \doteq s$— calculations at (5.20) and (5.21) below show that $\sigma_N = \kappa^{1/3}/\dot{\zeta}(\xi_2)$, and so in particular σ_N is of order $N^{1/3}$. This is certainly not a stochastic argument, but the picture does give some of the flavor of an approximation valid for quite small N.

There is an explicit error term from the Liouville-Green method for asymptotic approximation of ordinary differential equations (Olver, 1974) and this allows rigorous operator theoretic proofs to go through as in the Wigner matrix case.

3 Complex case

The goal of this section is to establish Theorem 1.3, the Tracy-Widom limit for the largest eigenvalue of a large complex Wishart matrix. First, the strategy. The starting point is the fixed N formula (2.6) which we write as $\det(I - S)$, regarding S as an operator on $L^2(t, \infty)$. To show the existence of a limit, adopt the scaling $t = \tau(s) = \mu_N + \sigma_N s$ suggested earlier. Define therefore

$$S_\tau(x, y) = \sigma_N S(\mu_N + \sigma_N x, \mu_N + \sigma_N y).$$

As an operator on $L^2(s, \infty)$, S_τ has the same eigenvalues as does S on $L^2(t, \infty)$. Hence

$$F_{N2}(\mu_N + \sigma_N s) = \det(I - S_\tau).$$

Let $\bar{S}$ denote the Airy operator on $L^2(s, \infty)$ with kernel

$$\bar{S}(x, y) = \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y) - \mathrm{Ai}(y)\mathrm{Ai}'(x)}{x - y}. \tag{3.1}$$

Tracy and Widom (1994) showed that the distribution F_2 satisfies $F_2(s) = \det(I - \bar{S})$, so that Theorem 1.3 follows if we show

$$\det(I - S_\tau) \to \det(I - \bar{S}). \tag{3.2}$$

Since the Fredholm determinant $\det(I - A)$ is a continuous function of A in the trace class norm on operators (e.g. Gohberg and Krein (1969, p. 160)), it suffices to show that $S_\tau \to \bar{S}$ in trace class norm.

Now for the details. Let D denote the differentiation operator, $(Df)(x) = f'(x)$. Widom (1999) derives a formula for the commutator $[D, S] = DS - SD$ in a class of unitary ensembles: for the Laguerre ensemble (2.1), this operator has kernel

$$(\frac{\partial}{\partial x} + \frac{\partial}{\partial y})S(x, y) = -\phi(x)\psi(y) - \psi(x)\phi(y). \tag{3.3}$$

Here ϕ is a function required to belong to $\mathcal{H}_N = \text{span}\{\phi_0, \phi_1, \ldots, \phi_{N-1}\}$ and ψ turns out then to be orthogonal to $\mathcal{H}_N$ and to satisfy $\int_0^\infty \psi = 0$. Let $\xi_N(x) = \phi_N(x)/x$ and $a_N = \sqrt{N(N + \alpha_N)}$. For the Laguerre ensemble, Widom finds that

$$\phi(x) = (-1)^N \sqrt{\frac{a_N}{2}} \{\sqrt{N + \alpha_N}\, \xi_N(x) - \sqrt{N}\, \xi_{N-1}(x)\} \tag{3.4}$$

$$\psi(x) = (-1)^N \sqrt{\frac{a_N}{2}} \{\sqrt{N}\, \xi_N(x) - \sqrt{N + \alpha_N}\, \xi_{N-1}(x)\} \tag{3.5}$$

[of course $\phi(x) = \phi(x; \alpha_N, N)$ and $\psi(x) = \psi(x; \alpha_N, N)$ depend on N and α_N, but this is not shown explicitly].

From (3.3) follows a useful integral representation

$$S(x, y) = \int_0^\infty \phi(x + z)\psi(x + z) + \psi(x + z)\phi(y + z)dz, \tag{3.6}$$

this is proved in the same manner as the formula for Gaussian unitary ensembles Tracy and Widom (1996, Sec. VII).

Laguerre asymptotics. The large N behavior of ϕ and ψ in the scaling $x = \mu_N + s\sigma_N$ is determined by the asymptotics of the Laguerre polynomials $L_N^{\alpha_N}$ and the corresponding weighted polynomials ϕ_N near the turning point μ_N. Using the notation $\phi_\tau(s) = \sigma_N \phi(\mu_N + s\sigma_N)$ in Section 5 we show for each fixed s, that as $N \to \infty$,

$$\phi_\tau(s), \quad \psi_\tau(s) \to \frac{1}{\sqrt{2}} \text{Ai}(s) \tag{3.7}$$

and, uniformly in N and in intervals of s that are bounded below,

$$\phi_\tau(s), \quad \psi_\tau(s) = O(e^{-s}). \tag{3.8}$$

Operator convergence. It follows from (3.6)–(3.8) that

$$S_\tau(x, y) = \int_0^\infty \phi_\tau(x + u)\psi_\tau(y + u) + \psi_\tau(x + u)\phi_\tau(y + u)du, \tag{3.9}$$

$$\to \int_0^\infty \text{Ai}(x + u)\text{Ai}(y + u)du = \bar{S}(x, y), \tag{3.10}$$

where $\bar{S}$ is shown in Tracy and Widom (1994) to be the Airy kernel (3.1).

In terms of operators on $L^2(s,\infty)$, the integral formulas (3.9) and (3.10) may be written

$$S_\tau = G_\tau H_\tau + H_\tau G_\tau, \qquad \bar{S} = 2G^2,$$

where the corresponding kernels

$$G_\tau(x,y) = \phi_\tau(x+y-s), \qquad G(x,y) = 2^{-1/2}\mathrm{Ai}(x+y),$$
$$H_\tau(x,y) = \psi_\tau(x+y-s).$$

Write $|A|_1$ for the trace class norm on operators (ℓ_1 norm on singular values), and $|A|_2$ for the Hilbert-Schmidt norm (ℓ_2 norm on singular values). Then by a standard inequality (Cauchy-Schwartz on singular values)

$$|S_\tau - \bar{S}|_1 \le 2|G_\tau|_2|H_\tau - G|_2 + 2|G_\tau - G|_2|G|_2 \to 0$$

since (3.7) and (3.8) show that G_τ and $H_\tau \to G$ in Hilbert-Schmidt norm on $L^2(s,\infty)$. This completes the proof of (3.2) and hence of Theorem 1.3.

Remark. The scaling limit (3.1) for $S_\tau(x,y)$ was stated by Johansson (2000), and for fixed α by Forrester (1993).

4 Real Case

In this section, suppose that X is an $N \times n$ *real* normal matrix with X_{ij} all independently and identically distributed as $N(0,1)$. The cross products matrix $A = XX'$ has the real Wishart distribution $W_N(n,I)$. The eigenvalues $(x_1,\dots,x_N)$ of A are real and non-negative, and according to the classical formula of Fisher, Girshick, Hsu, Mood and Roy have density

$$P_N(x_1,\dots,x_N) = d_{N,n}^{-1} \prod_{1\le j<k\le N} |x_j - x_k| \prod_{j=1}^N x_j^{\alpha/2} e^{-x_j/2}, \qquad \alpha = n-1-N, \tag{4.1}$$

This is again an example of a Laguerre ensemble, now corresponding to the orthogonal case (sometimes abbreviated LOE), with the differences $x_j - x_k$ raised to power $\beta = 1$, in contrast with the complex Wishart case, which corresponds to the Laguerre unitary ensemble (LUE) with $\beta = 2$.

A determinant representation for $\beta = 1$ analogous to (3.3) was developed by Dyson (1970). We refer to Tracy and Widom (1998), which gives a self-contained derivation of the formula

$$E_N \prod_{j=1}^N [1 + f(x_j)] = \sqrt{\det(I + K_N f)}, \tag{4.2}$$

which has as consequence for $f = -\chi$,

$$F_{N1}(t) = P\{\max_{1\le j\le N} x_j \le t\} = \sqrt{\det(I - K_N\chi)}. \tag{4.3}$$

Remark 4.1. As in the complex case, F_{N1} depends on n also, and should be written more carefully as $F_{N,n-1,1}$ to emphasize that the exponent in (4.1) $\alpha = n-1-N$. Here, the appropriate complex distribution is $F_{N,n-1,2}$ and so in determining the centering and scaling constants μ_N and σ_N, the indicated value for $\gamma = (n-1)/N$.

The major differences from (3.3) are the appearance of the square root, and the fact that K_N is now an operator with kernel a 2×2 matrix. Building on Tracy and Widom (1998), Widom (1999) shows that $K_N f : L_2^2(0,\infty) \to L_2^2(0,\infty)$ can be represented as a matrix of operators

$$K_N f = \begin{pmatrix} S + \psi \otimes \epsilon\phi & SD - \psi \otimes \phi \\ \epsilon S - \epsilon + \epsilon\psi \otimes \epsilon\phi & S + \epsilon\phi \otimes \psi \end{pmatrix} f \tag{4.4}$$

Here S, ψ and ϕ are as defined earlier at (2.4), (3.5) and (3.4) respectively. The function $\epsilon(x) = \frac{1}{2}\mathrm{sgn}\, x$ and the operator ϵ denotes convolution with the kernel $\epsilon(x-y)$. The notation $\alpha \otimes \beta$ denotes the rank one operator defined by $(\alpha \otimes \beta) f = (\beta, f)\alpha$.

Formula (4.4) has exactly the same form as that derived for the Gaussian orthogonal ensemble in Tracy and Widom (1996, Section V), hereafter [TW96]. Indeed, it turns out that much of the limiting argument can be made to follow the pattern of [TW96] exactly, so only a brief summary is given here.

The strategy is to derive the real case from the complex. Thus the complex Wishart probability $F_{N2}(t) = \det(I - S\chi)$ is treated as 'known'. The same manipulations as in [TW96] convert the matrix determinant appearing in (4.3) and (4.4) into a scalar determinant. After factoring out $I - S\chi$, one obtains a rank two perturbation of the identity:

$$F_{N1}(t)^2 / F_{N2}(t) = \det(I - \alpha_1 \otimes \beta_1 - \alpha_2 \otimes \beta_2) = \det_{1 \le j,k \le 2}(\delta_{jk} - (\alpha_j, \beta_k)). \tag{4.5}$$

To write out this determinant explicitly, we define certain quantities associated with the operator $(I - S\chi)^{-1}$. The notation is patterned after [TW96], though the formulas now refer to the Wishart setting.

$$\begin{aligned} P &= (I - S\chi)^{-1}\psi, & Q_\epsilon &= (I - S\chi)^{-1}\epsilon\phi, \\ \tilde{v}_\epsilon &= (Q_\epsilon, \chi\psi), & q_\epsilon &= Q_\epsilon(t), \\ c_\phi &= \tfrac{1}{2}\textstyle\int_0^\infty \phi. \end{aligned} \tag{4.6}$$

Let $R(x,y)$ be the kernel of the resolvent operator for $S\chi$, thus $(I - S\chi)^{-1} = I + R$. Set

$$\mathcal{P} = \int_0^t P(x)dx, \qquad \mathcal{R} = \int_0^t R(x,t)dx. \tag{4.7}$$

Then (4.5) is given by

$$F_{N1}(t)^2 / F_{N2}(t) = (1 - \tilde{v}_\epsilon)(1 - \tfrac{1}{2}\mathcal{R}) - \tfrac{1}{2}(q_\epsilon - c_\phi)\mathcal{P}. \tag{4.8}$$

Note that all the quantities just defined depend on t (with the exception of c_ϕ), though this is not shown explicitly. The right side of (4.8) is a finite N formula, although we do not have a more explicit evaluation for it. Nevertheless, by using the same scaling $t = \mu_N + \sigma_N s$, s fixed, as in the complex case, we may evaluate the large N limit. Indeed, there is an invariance principle: (4.8) converges to the same limit as occurs in the GOE setting.

Define

$$a(x) = 2^{-1/2} - 2^{-1/2}\int_x^\infty A(y)dy, \qquad \bar{Q} = (I - \bar{S}\bar{\chi})^{-1}a.$$

Then, as $N \to \infty$, arguing as in [TW96],

$$q_\epsilon \to \bar{q} = \bar{Q}(s), \qquad \tilde{v}_\epsilon \to \bar{u} = 2^{-1/2}(\bar{Q}, \bar{\chi}A),$$

and similarly $\mathcal{R} \to \bar{\mathcal{R}}$, and $\mathcal{P} \to \bar{\mathcal{P}}$, where in view of (3.7) and (3.8) and what has been said, the barred quantities are exactly those occurring in the GOE case. A separate argument (A7) is required to establish that $c_\phi \to 2^{-1/2}$ as $N \to \infty$ through even values. It follows then that

$$\lim F_{N1}^2(t)/F_{N2}(t) = F_1^2(s)/F_2(s) = (1-\bar{u})(1-\tfrac{1}{2}\bar{\mathcal{R}}) - \tfrac{1}{2}(\bar{q} - 2^{-1/2})\bar{\mathcal{P}}.$$

Setting $\mu = \int_s^\infty q(x)dx$, [TW96] identify

$$2\bar{u} = \bar{\mathcal{R}} = -\sqrt{2}\bar{\mathcal{P}} = 1 - \sqrt{2}\bar{q} = 1 - e^{-\mu},$$

and so

$$F_1^2(s)/F_2(s) = e^{-\mu} = \exp\{-\int_s^\infty q(x)dx\}.$$

Remark. [TW96] derive systems of differential equations satisfied by the functions in (4.6) and (4.7). The scaling limit of these finite N equations is then shown to exist and is solved to yield explicit formulas for the scaling limit of (4.6). In the Wishart-Laguerre setting, similar finite N differential equations could be derived, though this is not necessary for this proof – all we need is to show that the limit of (4.6) is the same as the GOE limit, which is already identified. [Of course, if the finite N equations, including those for $F_{N2}(t)$ could be solved exactly or numerically, then one would have a direct evaluation of $F_{N1}(t)$.]

5 Laguerre polynomial asymptotics

In this section, we establish (3.7) and (3.8). The key to these results are certain asymptotics for Laguerre polynomials of both large order and large degree, in the neighborhood of the largest zero. These asymptotics are nonstandard since statistical applications require $\alpha_N = n - N$ large, as well as N.

Specifically, consider the Laguerre polynomials $L_N^\alpha(x)$, as defined in Szegö (1967), but with $\alpha = \alpha_N \sim (\gamma - 1)N$ for N large and $\gamma \geq 1$ fixed. With the abbreviations $N_+ = N + 1/2$ and $n_+ = n + 1/2 = N + \alpha_N + 1/2$, define a rescaling $x = \mu_N + \sigma_N s$ with

$$\mu_N = (\sqrt{N_+} + \sqrt{n_+})^2.$$
$$\sigma_N = (\sqrt{N_+} + \sqrt{n_+})\Big(\frac{1}{\sqrt{N_+}} + \frac{1}{\sqrt{n_+}}\Big)^{1/3}.$$

Writing $\rho = \gamma^{-1/6}(1 + \sqrt{\gamma})^{4/3}$, we show that, as $N \to \infty$,

$$(-1)^N N^{-1/6}\sqrt{N!/(N+\alpha_N)!}\; x^{(\alpha_N+1)/2}e^{-x/2}L_N^{\alpha_N}(x) \begin{cases} \to \sqrt{\rho}\,\mathrm{Ai}(s) & \forall s \in \mathbb{R} \\ = O(e^{-s}) & \text{uniformly in } [s_0, \infty), s_0 \in \mathbb{R}. \end{cases} \tag{5.1}$$

Compare Figure 6. Note that when $\alpha_N \equiv \alpha$ is fixed, the pointwise result reduces to the classical Plancherel-Rotach type formula of Szegö (1967, p. 201): for example $\mu_N = 4N + 2\alpha + 2 + O(N^{-1})$. Also helpful for fixed α is Erdélyi (1960).

We use a standard differential equation approach, based on heuristics sketched in Section 2 and exploiting existing literature. Indeed, pointwise convergence was studied by Temme

(1990), based on work of Dunster (1989), in turn building on Olver (1974). We give an account starting from Olver (1974) (i) to correct a consequential typographical error in the first two references, (ii) to derive the uniform bound in (5.1), which is needed for the operator theoretic arguments of Section 3, and (iii) to focus on real asymptotics for Laguerre polynomials rather than the more general setting of Whittaker functions in the complex domain of Dunster (1989).

The differential equation satisfied by $w_N(x) = x^{(\alpha+1)/2}e^{-x/2}L_N^\alpha(x)$ is, from Szegö (1967, p 100),

$$\frac{d^2w}{dx^2} = \left\{\frac{1}{4} - \frac{\kappa}{x} + \frac{\lambda^2 - 1/4}{x^2}\right\} w, \qquad \kappa = N + \frac{\alpha+1}{2}, \quad \lambda = \frac{\alpha}{2}. \tag{5.2}$$

Rescaling the x axis via $\xi = x/\kappa$ puts this equation into the form

$$\frac{d^2w}{d\xi^2} = \left\{\kappa^2 f(\xi) + g(\xi)\right\} w, \tag{5.3}$$

where

$$f(\xi) = \frac{(\xi - \xi_1)(\xi - \xi_2)}{4\xi^2}, \qquad g(\xi) = -\frac{1}{4\xi^2} \tag{5.4}$$

The turning points of the equation are given by zeros of f, namely

$$\xi_1 = 2 - \sqrt{4 - \omega^2}, \qquad \xi_2 = 2 + \sqrt{4 - \omega^2}, \tag{5.5}$$

with

$$\omega = \frac{2\lambda}{\kappa} = \frac{\alpha}{N + (\alpha+1)/2}. \tag{5.6}$$

Since N and $n = \gamma N$ are large, $\kappa \sim \frac{1}{2}(1+\gamma)N$ will be a large parameter, while ω approaches a finite limit, $\omega \to 2(\gamma-1)/(\gamma+1) \in [0, 2)$.

Liouville-Green. This classical method, modified to the case of turning points, describes how the solutions of equations such as (5.3) – (5.4) are approximated by Airy functions. This theory, *along with error bounds*, is described in detail by Olver (1974), referred to as [O] below. We summarize and specialize the part we need here. A change of independent variable in (5.3) is made by the Liouville-Green transformation $\zeta(\xi)$ defined on an interval containing ξ_2 (say $(2, \infty)$) by

$$\tfrac{2}{3}\zeta^{3/2} = \int_{\xi_2}^{\xi} f^{1/2}(t)\,dt. \tag{5.7}$$

Define also a new dependent variable W by $w = (d\zeta/d\xi)^{-1/2}W$. These choices put (5.3) into the form

$$\frac{d^2W}{d\zeta^2} = \{\kappa^2\zeta + \psi(\zeta)\}W \tag{5.8}$$

where the perturbation term $\psi(\zeta) = \hat{f}^{-1/4}(d^2/d\zeta^2)(\hat{f}^{1/4}) + g/\hat{f}$. Here $\hat{f}$ is defined by

$$\hat{f}(\xi) = \left(\frac{d\zeta}{d\xi}\right)^2 = \frac{f(\xi)}{\zeta}. \tag{5.9}$$

with the second equality following from (5.7). [O,Lemma 11.3.1] guarantees that $\zeta(\xi)/(\xi - \xi_2)$ is positive and twice continuously differentiable for $\xi \in (2, \infty)$.

If the perturbation term $\psi(\zeta)$ in (5.8) were absent, the equation $d^2W/d\zeta^2 = \kappa^2 \zeta W$ would have linearly independent solutions in terms of Airy functions, traditionally denoted by $\mathrm{Ai}(\kappa^{2/3}\zeta)$ and $\mathrm{Bi}(\kappa^{2/3}\zeta)$. Our interest is in approximating the *recessive* solution $\mathrm{Ai}(\kappa^{2/3}\zeta)$, so write the relevant solution of (5.8) as $W_2(\zeta) = \mathrm{Ai}(\kappa^{2/3}\zeta) + \eta(\zeta)$. In terms of the original dependent and independent variables ξ and w, the solution W_2 becomes

$$w_2(\kappa, \xi) = \hat{f}^{-1/4}(\xi)\{\mathrm{Ai}(\kappa^{2/3}\zeta) + \epsilon_2(\kappa, \xi)\}. \tag{5.10}$$

[O, Theorem 11.3.1] provides an explicit bound for $\eta(\zeta)$ and hence ϵ_2 in terms of the function $\mathcal{V}(\zeta) = \int_\zeta^\infty |\psi(v)v^{-1/2}|dv$. (A3 has more on $\mathcal{V}$.) To describe this error bound even in the oscillatory region of $\mathrm{Ai}(x)$, [O] introduces a modulus function $M(x)$ and a positive weight function $E(x)$ such that if $E^{-1}(x) = 1/E(x)$ then $\mathrm{Ai}(x) \le M(x)E^{-1}(x)$ for all x. In addition,

$$\mathrm{Ai}(x) = \frac{1}{\sqrt{2}} M(x)E^{-1}(x), \qquad x \ge c \doteq -0.37, \tag{5.11}$$

and the asymptotics as $x \to \infty$ are given by

$$E(x) \sim \sqrt{2}e^{\frac{2}{3}x^{3/2}}, \qquad M(x) \sim \pi^{-1/2}x^{-1/4}. \tag{5.12}$$

The key bound of [O, Theorem 11.3.1] is then

$$|\epsilon_2(\kappa, \xi)| \le M(\kappa^{2/3}\zeta)E^{-1}(\kappa^{2/3}\zeta)[\exp\{\frac{\lambda_0}{\kappa}\mathcal{V}(\zeta)\} - 1], \qquad \xi \in (2, \infty) \tag{5.13}$$

where $\lambda_0 \doteq 1.04$. For $\xi \ge \xi_2$, (5.11) shows that the coefficient is just $\sqrt{2}\mathrm{Ai}(\kappa^{2/3}\zeta)$.

Identifying Laguerre polynomials. (5.13) has a useful double asymptotic character in κ and ξ. First, suppose that N and hence κ are held fixed. As $\xi \to \infty$, we have $\zeta \to \infty$ also, since $f(\xi) \sim 1/4$. Consequently $\mathcal{V}(\zeta) \to 0$ and so from (5.13) and its following remark,

$$\epsilon_2(\kappa, \xi) = o(\mathrm{Ai}(\kappa^{2/3}\zeta)) \qquad \text{as} \quad \xi \to \infty. \tag{5.14}$$

Since the weighted Laguerre polynomial $w_N(x) = x^{(\alpha+1)/2}e^{-x/2}L_N^\alpha(x)$ is a recessive solution of (5.2), it must be proportional to w_2:

$$w_N(\kappa\xi) = c_\kappa w_2(\kappa, \xi), \tag{5.15}$$

and we can use (5.10), (5.14) and the asymptotics of $\mathrm{Ai}(x)$ to evaluate c_κ (see A4). This yields

$$c_\kappa = \frac{(-1)^N}{N!}(2\pi)^{1/2}\kappa^{\kappa+1/6}e^{\kappa c_0}, \tag{5.16}$$

where $c_0 = c_0(\omega)$ is given explicitly at (6.3) below.

Let us summarize the situation so far. Recalling from (2.3) that $\phi_N(x) = \sqrt{N!/(N+\alpha)!}$ $x^{\alpha/2}e^{-x/2}L_N^\alpha(x)$, observe that the left side of (5.1) becomes $(-1)^N N^{-1/6}x^{1/2}\phi_N(x)$. From (5.15) and (5.16) we have

$$x^{1/2}\phi_N(x) = \sqrt{\frac{N!}{(N+\alpha)!}}w_N(x) = (-1)^N\kappa^{1/6}r_N w_2(\kappa, \xi), \tag{5.17}$$

where

$$r_N := (-1)^N \frac{c_\kappa}{\kappa^{1/6}} \sqrt{\frac{N!}{(N+\alpha)!}} \to 1, \qquad \text{as } N \to \infty, \tag{5.18}$$

(and the convergence is shown in A5.) Bringing in (5.10), we have then, for fixed N and x,

$$(-1)^N N^{-1/6} x^{1/2} \phi_N(x) = r_N (\kappa_N/N)^{1/6} \hat{f}^{-1/4}(\xi) \{\mathrm{Ai}(\kappa^{2/3}\zeta) + \epsilon_2(\kappa, \xi)\}. \tag{5.19}$$

$N^{1/3}$ scaling about the largest zero. On the original x scale, we are interested in values $x = \mu_N + \sigma_N s$, where $\mu_N = \kappa\xi_2$ is the upper turning point. We now choose the scale σ_N so that $\mathrm{Ai}(\kappa^{2/3}\zeta) \to \mathrm{Ai}(s)$. Expand $\zeta(\xi)$ about the turning point ξ_2, at which $\zeta(\xi_2) = 0$, and put $\dot{\zeta} = \dot{\zeta}(\xi_2)$. For s fixed, we have the approximation

$$\kappa^{2/3}\zeta(\xi) = \kappa^{2/3}\zeta(\xi_2 + \sigma_N \kappa^{-1} s) \doteq \sigma_N \kappa^{-1/3} s \dot{\zeta}. \tag{5.20}$$

Equating the right side with s yields the $N^{1/3}$ scaling:

$$\sigma_N = \kappa^{1/3}/\dot{\zeta} \sim \rho N^{1/3}, \tag{5.21}$$

where the final limit follows from explicit evaluation of $\dot{\zeta}(\xi_2)$ in A6. As noted earlier, $\xi \to \zeta(\xi)/(\xi - \xi_2)$ is positive and C^2 near ξ_2, and from this it follows that uniformly in $s \le \delta N^{2/3}$, we have $\kappa^{2/3}\zeta = s + O(N^{-2/3})$. Since Ai is continuous, for each fixed s,

$$\lim_{N\to\infty} \mathrm{Ai}(\kappa^{2/3}\zeta) = \mathrm{Ai}(s). \tag{5.22}$$

Negligibility of error term. Return to the error term in Olver's bound (5.13). First note that $s \ge s_0$ is equivalent to $\xi \ge \xi_2 + s_0\sigma_N/\kappa = \xi_2 + O(N^{-2/3})$. On the ζ scale, this means for any fixed $\zeta_0 < 0$, and for N large enough, that $s \ge s_0$ entails $\zeta \ge \zeta_0$. Letting $c_1 = \lambda \mathcal{V}(\zeta_0)$, the final term in (5.13) can be bounded

$$\exp\{\lambda \mathcal{V}(\zeta)/\kappa\} - 1 \le e^{c_1/\kappa} - 1 =: \epsilon_N \to 0 \qquad \text{as } N \to \infty.$$

Consequently, (5.13) becomes, uniformly in $s \in [s_0, \infty)$,

$$|\epsilon_2(\kappa, \xi)| \le M(\kappa^{2/3}\zeta) E^{-1}(\kappa^{2/3}\zeta) \epsilon_N = o(\mathrm{Ai}(\kappa^{2/3}\zeta)). \tag{5.23}$$

Pointwise limit in (5.1). From (5.19), by combining (5.18), (5.22) and (5.23) with the limit $(\kappa_N/N)^{1/6}\hat{f}^{-1/4}(\xi) \to \sqrt{\rho}$, (see (6.11) in Appendix A6),

$$(-1)^N N^{-1/6} x^{1/2} \phi_N(x) \to \sqrt{\rho}\,\mathrm{Ai}(s). \tag{5.24}$$

As remarked earlier, the definition of ϕ_N shows that this is (5.1).

Uniform bound in (5.1). From (5.19) and (5.23), and recalling that $\mathrm{Ai} \le ME^{-1}$, we have

$$|(-1)^N N^{-1/6} x^{1/2} \phi_N(x)| \le C(\gamma) \hat{f}(\xi)^{-1/4} M(\kappa^{2/3}\zeta) E^{-1}(\kappa^{2/3}\zeta).$$

It remains to show that the right side, which we denote by $T_N(s)$, satisfies $T_N(s) \le Ce^{-s}$ for $s \ge s_0$. This is done in A8. We remark here only that for $\zeta > 0$, (5.12) shows that

$E^{-1}(\kappa^{2/3}\zeta) \le c_0 e^{-\frac{2}{3}\kappa\zeta^{3/2}}$, and a (by no means sharp) argument shows that $(2/3)\kappa\zeta^{3/2} \ge s$ for s large enough.

Large N asymptotics of ϕ_τ and ψ_τ, needed for Section 3, are now straightforward to derive from (5.24). Returning to (3.4) and (3.5), write

$$\phi_\tau = \phi_{I,N} + \phi_{II,N}, \qquad \psi_\tau = \psi_{I,N} + \psi_{II,N}, \tag{5.25}$$

where, for example,

$$(-1)^N \phi_{I,N}(x) = \sigma_N[a_N(N+\alpha_N)]\phi_N(x)/\sqrt{2}x = N^{-1/6}x^{1/2}\phi_N(x)\cdot d_N(x) \tag{5.26}$$

where

$$d_N(x) = d_N(x/\mu_N)^{-3/2}, \qquad d_N = \sigma_N N^{1/6+1/4}(N+\alpha_N)^{3/4}/\sqrt{2}\mu_N^{3/2}.$$

As $N \to \infty$, it is easily seen that $x/\mu = 1 + s\sigma_N/\mu_N \to 1$ and

$$\sqrt{\rho}d_N \sim \frac{\rho^{3/2}\gamma^{3/4}}{\sqrt{2}(1+\sqrt{\gamma})^3} = \frac{\sqrt{\gamma}}{\sqrt{2}(1+\sqrt{\gamma})} =: 2^{-1/2}b_\gamma.$$

Using, (5.24) this establishes the first part of

$$\phi_{I,N}(\mu_N + \sigma_N s) \begin{cases} \to 2^{-1/2}b_\gamma \mathrm{Ai}(s) \\ \le C(\gamma)e^{-s}. \end{cases} \tag{5.27}$$

The second part follows from (5.26), the second part of (5.1) and the simple bound

$$(x/\mu_N)^{-3/2} = (1 + s\sigma_N/\mu_N)^{-3/2} \le e^{-(3\sigma_N/2\mu_N)s}.$$

It is easily verified that all the other terms (5.25) can be written in terms of $\phi_{I,N}$. Indeed, setting $u_N = (\sigma_N/\sigma_{N-1})\sqrt{a_N/a_{N-1}}$, then $n_N = N + \alpha_N$ and $v_N = (N/n_N)^{1/2}$, we have

$$\begin{aligned} \psi_{I,N} &= v_N\ \phi_{I,N}, \\ \phi_{II,N} &= u_N v_{N-1}\ \phi_{I,N-1}, \\ \psi_{II,N} &= u_N v_N^{-1} v_{N-1}\ \phi_{I,N-1}. \end{aligned} \tag{5.28}$$

Let us show how (3.7) and (3.8) follow from these and (5.28). First, note that writing $\mu_N + \sigma_N s = \mu_{N-1} + \sigma_{N-1}s'$ yields

$$s' = s + \frac{\mu_N - \mu_{N-1}}{\sigma_{N-1}} + \frac{\sigma_N - \sigma_{N-1}}{\sigma_{N-1}}s = s + O(N^{-1/3}). \tag{5.29}$$

From this it follows that $\lim_{N\to\infty}\phi_{I,N-1}(\mu_N + \sigma_N s)$ is the same as (5.27). Since $u_N \to 1$ and $v_N^{-2} = n_N/N \to \gamma$ as $N \to \infty$, we obtain immediately from (5.28)

$$\lim_{N\to\infty}\phi_{II,N}(\mu_N + \sigma_N s) = 2^{-1/2}(1 - b_\gamma)\mathrm{Ai}(s). \tag{5.30}$$

Adding (5.27) and (5.30) yields (3.7) for ϕ_τ. The corresponding result for ψ_τ follows from (5.28) in a similar way. Turning now to the uniform bounds (3.8), note first that since μ_N and σ_N are increasing with N, (5.29) implies that $s' \ge s$, and so, for example

$$\phi_{I,N-1}(\mu_N + \sigma_N s) = \phi_{I,N-1}(\mu_{N-1} + \sigma_{N-1}s') \le Ce^{-s'} \le Ce^{-s}.$$

Now (3.8) follows directly from this, (5.25) and (5.28).

Remark. It can be shown if n^2/N is increasing in N, then so also are both μ_N and σ_N.

6 Appendix

A1. Tails of the Tracy-Widom density of order 1. Write $f(s) \approx g(s)$ when $f(s) \sim r(s)g(s)$ and $r(s)$ involves at most rational powers of s. The density of F_1 is

$$f_1(s) = \tfrac{1}{2}F_1(s)\{q(s) + \textstyle\int_s^\infty q^2(x)dx\}.$$

For large positive s, from (1.5), $q(s) \sim \mathrm{Ai}(s) \sim 2^{-1}\pi^{-1/2}s^{-1/4}e^{-\frac{2}{3}s^{3/2}}$. Since $\int_s^\infty q^2(x)dx \sim (8\pi s)^{-1}e^{-\frac{4}{3}s^{3/2}} = o(q(s))$, we have $f_1(s) \approx q(s) \approx e^{-\frac{2}{3}s^{3/2}}$ as $s \to \infty$. For large negative s, Hastings and McLeod (1980) show that $q(s) \sim \sqrt{-s/2}$, so it follows that $\int_s^\infty q \sim (\sqrt{2}/3)|s|^{3/2}$ and $\int_s^\infty (x-s)q^2(x)dx \sim |s|^3/12$. Hence, $f_1(s) \approx e^{-|s|^3/24}$ for large negative s.

A2. Proof of Proposition 1.2. For a square symmetric $p \times p$ matrix A, write $l_1(A) \geq l_2(A) \geq \ldots \geq l_p(A)$ for the ordered eigenvalues. Let r be an integer with $1 \leq r \leq p$, and let A_{p-r} be a $(p-r) \times (p-r)$ submatrix obtained by deleting r rows and the corresponding columns from A. Then the "inclusion principle" (e.g. Horn and Johnson (1985, Thm. 4.3.15)) states that for each integer k such that $1 \leq k \leq p-r$,

$$l_{r+k}(A) \leq l_k(A_{p-r}) \leq l_k(A). \tag{6.1}$$

Now let Y be an $n \times p$ data matrix with rows independently drawn from $N(0, \Sigma_\tau)$, where $\Sigma_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_r^2, 1, \ldots, 1)$. Partition $Y = [Y_1 : Y_2]$ with the $n \times r$ matrix Y_1 containing the first r columns and Y_2 the remaining $n-r$ ones. Now simply apply (6.1) with $k = 1$ to $A = Y'Y$ and $A_{p-r} = Y_2'Y_2 \sim W_{p-r}(n, I_{p-r})$. The first inequality is precisely the assertion of the Proposition.

A3. Details on $\mathcal{V}(\zeta)$. Since $\psi(\zeta)$ is continuous, finiteness of $\mathcal{V}(\zeta)$ follows from the continuity of $\psi(\zeta)$ in $(\zeta(2), \infty)$ ([O,Lemma 11.3.1]) together with observation that $\psi(\zeta) \sim -1/(4\zeta^2)$ as $\zeta \to \infty$. [The latter comes from calculation with the formula for ψ preceding (5.9) after observing that $\hat{f}(\zeta) \sim 1/(4\zeta)$ for large ζ and that $\zeta^{3/2} \sim 3\xi/4$ for large ξ.]

A4. Evaluation of c_κ. Since $w_N(x) = x^{(\alpha+1)/2}e^{-x/2}L_N^\alpha(x)$ and $L_N^\alpha(x) \sim (-1)^N x^N/N!$ for large x, we have

$$w_N(\kappa\xi) \sim e^{-\kappa\xi/2}(\kappa\xi)^{(\alpha+1)/2}\frac{(-1)^N}{N!}(\kappa\xi)^N, \qquad \xi \to \infty.$$

From (5.12) $\mathrm{Ai}(x) \sim 2^{-1}\pi^{-1/2}x^{-1/4}e^{-\frac{2}{3}x^{3/2}}$ for large x, and so

$$w_2(\kappa, \xi) \sim \frac{1}{2\sqrt{\pi}\kappa^{1/6}f^{1/4}(\xi)}\exp\{-\tfrac{2}{3}\kappa\zeta^{3/2}\}, \qquad \xi \to \infty.$$

Since $\kappa = \frac{\alpha+1}{2} + N$,

$$(-1)^N c_\kappa = \frac{2\sqrt{\pi}\kappa^{1/6}}{N!}\lim_{\xi\to\infty} f^{1/4}(\xi)(\kappa\xi)^\kappa \exp\{\kappa(\tfrac{2}{3}\zeta^{3/2} - \xi/2)\} \tag{6.2}$$

To evaluate this limit, we need the large ξ asymptotics of $\zeta(\xi)$. Set $R(\xi) = (\xi - \xi_1)^{1/2}(\xi - \xi_2)^{1/2}$. According to Dunster (1989, formula (4.6)), (5.7) is given by

$$\tfrac{2}{3}\zeta^{3/2} = \tfrac{1}{2}R(\xi) + \tfrac{\omega}{2}\ln\tfrac{\xi}{\xi_2}\{\tfrac{2\xi_2 - \omega^2}{2\xi - \omega^2 - \omega R(\xi)}\} + \ln\{\tfrac{\xi_2 - 2}{\xi - 2 + R(\xi)}\}.$$

Noting that $R(\xi) = \xi - (\xi_1 + \xi_2)/2 + O(\xi^{-1})$, and that $\xi_1 + \xi_2 = 4$, we arrive at

$$\tfrac{2}{3}\zeta^{3/2} = \frac{\xi}{2} - \ln \xi + c_0 + O(\xi^{-1}),$$

where

$$c_0 = -1 + \frac{\omega}{2} \log \frac{2\xi_2 - \omega^2}{\xi_2(2-\omega)} + \log(\frac{\xi_2}{2} - 1). \tag{6.3}$$

[The value of c_0 corrects Dunster (1989, formula (4.7)), which omits the -1.] Since $f(\xi) \sim 1/4$ as $\xi \to \infty$, the previous display and (6.2) yield

$$(-1)^N c_\kappa = \frac{\sqrt{2\pi}\kappa^{1/6}}{N!}\kappa^\kappa \lim_{\xi\to\infty} \exp[\kappa c_0 + O(\xi^{-1})]$$

which reduces to (5.16).

A5. Proof of (5.18)*:* We first rewrite the constant c_0, using (5.5) along with $\omega^2 = \xi_1\xi_2$, (5.6) and (5.2) and $n = N + \alpha$ to obtain

$$\frac{2\xi_2 - \omega^2}{\xi_2(2-\omega)} = \frac{2-\xi_1}{2-\omega} = \Big(\frac{2+\omega}{2-\omega}\Big)^{1/2} = \Big(\frac{2\kappa+\alpha}{2\kappa-\alpha}\Big)^{1/2} = \Big(\frac{n+1/2}{N+1/2}\Big)^{1/2},$$

and

$$\frac{\xi_2}{2} - 1 = \frac{(2+\omega)^{1/2}(2-\omega)^{1/2}}{2} = \frac{(2\kappa+\alpha)^{1/2}(2\kappa-\alpha)^{1/2}}{2\kappa} = \frac{(n+1/2)^{1/2}(N+1/2)^{1/2}}{\kappa}.$$

Since $\omega\kappa/2 = \alpha/2$, it follows that

$$\kappa^{2\kappa} e^{2c_0\kappa} = e^{-2\kappa}(n+1/2)^{n+1/2}(N+1/2)^{N+1/2}. \tag{6.4}$$

By Stirling's formula

$$N!(N+\alpha)! \sim 2\pi\, e^{-2N-\alpha}\, n^{n+1/2}\, N^{N+1/2}, \tag{6.5}$$

and so as $N \to \infty$,

$$r_N^2 = \frac{2\pi\kappa^{2\kappa}e^{2\kappa c_0}}{N!(N+\alpha)!} \sim e^{-1}\Big(1+\frac{1}{2n}\Big)^{n+1/2}\Big(1+\frac{1}{2N}\Big)^{N+1/2} \sim 1.$$

A6. Evaluation of $\dot\zeta, \mu_N$ *and* σ_N We first derive formulas for fixed N and then evaluate the large N limits. From (5.9), and using L'Hôpital's rule as $\xi \to \xi_2$,

$$\dot\zeta(\xi) = \frac{1}{2\xi}\Big[\frac{(\xi-\xi_1)(\xi-\xi_2)}{\zeta(\xi)}\Big]^{1/2} \to \frac{1}{2\xi_2}\Big(\frac{\xi_2-\xi_1}{\dot\zeta}\Big)^{1/2}.$$

Solving for $\dot\zeta$ yields

$$\dot\zeta = \Big[\frac{\xi_2-\xi_1}{4\xi_2^2}\Big]^{1/3}. \tag{6.6}$$

We use the abbreviations $N_+ = N + 1/2, n_+ = n + 1/2$. It follows from (5.2) that

$$\kappa = \frac{N_+ + n_+}{2}, \qquad \kappa^2 - \lambda^2 = N_+ n_+.$$

Hence, since $\kappa\omega = 2\lambda$,

$$\mu_N = \kappa\xi_2 = 2\kappa + 2\sqrt{\kappa^2 - \lambda^2} = (\sqrt{N_+} + \sqrt{n_+})^2. \tag{6.7}$$

Note that

$$\kappa(\xi_2 - \xi_1) = 4\sqrt{\kappa^2 - \lambda^2} = 4\sqrt{N_+ n_+}.$$

Using (5.21) and (6.6),

$$\sigma_N^3 = \frac{\kappa}{\dot\zeta^3} = \frac{4(\kappa\xi_2)^2}{\kappa(\xi_2 - \xi_1)} = \frac{(\sqrt{N_+} + \sqrt{n_+})^4}{\sqrt{N_+ n_+}},$$

so that

$$\sigma_N = (\sqrt{N_+} + \sqrt{n_+})\Big(\frac{1}{\sqrt{N_+}} + \frac{1}{\sqrt{n_+}}\Big)^{1/3}. \tag{6.8}$$

We now turn to large N approximations. Since $n = \gamma N$, we immediately find from (6.7) and (6.8),

$$\mu_N \sim (\sqrt{N} + \sqrt{n})^2 \sim (1 + \sqrt{\gamma})^2 N \tag{6.9}$$

$$\sigma_N \sim \rho N^{1/3} \qquad \rho = \gamma^{-1/6}(1 + \sqrt{\gamma})^{4/3}. \tag{6.10}$$

In addition, (5.9) again shows that $\hat f^{-1/4}(\xi) = \dot\zeta^{-1/2}(\xi)$ and so, as $N \to \infty$,

$$\kappa^{1/6}\hat f^{-1/4}(\xi) \sim (\kappa^{1/3}/\dot\zeta)^{1/2} \sim \sqrt{\rho} N^{1/6}. \tag{6.11}$$

A7. Limiting value of c_ϕ. Throughout this argument, $\alpha = \alpha_N = (\gamma - 1)N$, and we set $\beta = (\gamma - 1)/2$. We will assume $\gamma > 1$. [For $\gamma = 1$, (essentially α fixed) the same result can be established by a more direct argument.] From (4.6) and (3.4), we have

$$\sqrt{2} c_\phi = \tfrac{1}{2}\sqrt{a_N}[\sqrt{N + \alpha}\int \xi_N - \sqrt{N}\int \xi_{N-1}].$$

Since $\int \psi = 0$, (3.5) yields $\sqrt{N + \alpha}\int \xi_{N-1} = \sqrt{N}\int \xi_N$, and hence

$$\sqrt{2} c_\phi = \Big(\frac{N}{N + \alpha}\Big)^{1/4}\frac{\alpha}{2}\int \xi_N = \frac{\alpha}{2}\Big(\frac{N}{N + \alpha}\Big)^{1/4}\Big(\frac{N!}{(N + \alpha)!}\Big)^{1/2} c_N. \tag{6.12}$$

The large N behavior of

$$c_N = \int_0^\infty x^{\alpha/2 - 1} e^{-x/2} L_N^\alpha(x)\, dx$$

follows via generating functions and the saddlepoint method. By Szegö (1967, eq. (5.1.9)),

$$\sum_{N=0}^\infty L_N^\alpha(x) t^N = (1 - t)^{\alpha - 1}\exp\{-xt/(1 - t)\},$$

and so, after evaluating a gamma integral,

$$h(t) := \sum_{N=0}^{\infty} c_N t^N = 2^{\alpha/2}\Gamma(\alpha/2)(1-t)^{-1}(1-t^2)^{-\alpha/2}.$$

By Cauchy's theorem, for a suitable contour $\mathcal{C}$ encircling 0,

$$c_N = \frac{1}{2\pi i}\int_{\mathcal{C}} \frac{h(t)}{t^{N+1}}\,dt = \frac{2^{\alpha/2}\Gamma(\alpha/2)}{2\pi i} I_N, \tag{6.13}$$

where, on setting $p(t) = \log\{t(1-t^2)^{\beta}\}$ and $q(t) = t^{-1}(1-t)^{-1}$,

$$I_N = \int_{\mathcal{C}} e^{-Np(t)} q(t)\,dt.$$

The saddlepoints, being the (simple) zeros of $p'(t)$, are given by $t_{\pm} = \pm\gamma^{-1/2}$. On the circular contour $t = t(u) = \gamma^{-1/2}e^{iu}$, traversed counterclockwise, $2\mathrm{Re}\,p(t) = -\log\gamma + \beta\log\{1+\gamma^{-2}-2\gamma^{-1}\cos 2u\}$ has equal global minima at $t_{\pm}$. Consequently, the saddlepoint approximation [e.g. Olver (1974, Thm 4.7.1)] yields, as $N\to\infty$,

$$I_N \sim (\frac{2\pi}{N})^{1/2} \sum \frac{q(t_{\pm})}{[p''(t_{\pm})]^{1/2}} e^{-Np(t_{\pm})}, \tag{6.14}$$

where the sum is over the two cases t_+ and t_-. In forming $[p''(t_{\pm})]^{1/2}$, the branches $\omega_{0,\pm}$ of $\arg\{p''(t_{\pm})\}$ are chosen to satisfy $|\omega_{0,\pm} + 2\omega_{\pm}| \le \pi/2$, where $\omega_{\pm} = \pm\pi/2$ are limiting values of $\arg(t-t_{\pm})$ as $t\to t_{\pm}$ *clockwise* along the contour. Thus, since $p''(t_{\pm}) = -2\gamma^2/(\gamma-1)$, we have $\omega_{0,\pm} = \mp\pi$ and so $[p''(t_{\pm})]^{-1/2} = \pm i\sqrt{\gamma-1}/\gamma\sqrt{2}$. Since $e^{2p(t_{\pm})} = (\gamma-1)^{\gamma-1}\gamma^{-\gamma}$ and $q(t_{\pm}) = \pm\gamma(\sqrt{\gamma}\pm 1)^{-1}$, we have, for large even N, after combining the two terms in (6.14),

$$\Big(\frac{I_N}{2\pi i}\Big)^2 \sim \frac{1}{\pi N}\frac{\gamma}{\gamma-1}\Big[\frac{\gamma^{\gamma}}{(\gamma-1)^{\gamma-1}}\Big]^N.$$

Returning at last to c_{ϕ}, collecting (6.12) and (6.13) gives $\sqrt{2}c_{\phi} = b_N I_N/(2\pi i)$, where

$$\begin{aligned} b_N^2 &= \Big(\frac{N}{N+\alpha}\Big)^{1/2}\frac{N!}{(N+\alpha)!}\Big[(\frac{\alpha}{2})!\Big]^2 2^{\alpha} \\ &\sim \frac{\pi}{2}\frac{N^{N+1}\alpha^{\alpha+1}}{(N+\alpha)^{N+\alpha+1}} = \pi N\frac{\gamma-1}{\gamma}\,[\frac{(\gamma-1)^{\gamma-1}}{\gamma^{\gamma}}]^N, \end{aligned}$$

after using Stirling's formula. Since b_N and $I_N/(2\pi i) > 0$, it follows that $c_{\phi}\to 1/\sqrt{2}$.

A8. Uniform bound for $T_N(s)$. In the following, the constant $c = c(\gamma)$ is not necessarily the same at each appearance.

1. We first construct s_1 such that if N is large and $s \ge 2s_1$, then $E^{-1}(\kappa^{2/3}\zeta) \le Ce^{-s}$. Indeed, writing $\xi = \xi_2 + s\sigma_N/\kappa$ and using the definition of f, we find

$$\frac{\sigma_N^2}{4}f(\xi) = \frac{s}{4}\,\frac{\sigma_N^3}{\kappa}\,\frac{\xi_2-\xi_1+s\sigma_N/\kappa}{[\xi_2+s\sigma_N/\kappa]^2} \sim \frac{s}{c(\gamma)} \qquad \text{as } N\to\infty.$$

For a fixed $\delta > 0$, set $s_1 = c(\gamma)(1+\delta)$. For large N and $s \geq s_1$ we then have $\sqrt{f(\xi)} \geq 2/\sigma_N$ and hence

$$\tfrac{2}{3}\kappa\zeta^{3/2} = \kappa \int_{\xi_2}^{\xi} \sqrt{f} \geq \kappa \cdot \frac{2}{\sigma_N} \cdot (s - s_1)\frac{\sigma_N}{\kappa} = 2(s - s_1) \geq s,$$

if $s \geq 2s_1$. Since $E(x) \geq c_0 e^{\frac{2}{3}x^{3/2}}$ for $x \geq 0$, we have $E^{-1}(\kappa^{2/3}\zeta) \leq c_0^{-1} e^{-\frac{2}{3}\kappa\zeta^{3/2}} \leq c_0^{-1} e^{-s}$ as claimed.

2. For $s \geq 2s_1$, $f(\xi) \geq f(\xi_2 + 2s_1\sigma_N/\kappa) \geq c(\sigma_N/\kappa)f'(\xi_2)$. Since $f'(\xi_2) = \xi_2^{-2}(\xi_2 - \xi_1) \to c(\gamma) > 0$ as $N \to \infty$, we have for large N and such s, $\kappa^{2/3} f(\xi) \geq c$. Since $M(x) \leq cx^{-1/4}$ for $x \geq 0$, we have (using the definition of $\hat{f}$)

$$\hat{f}(\xi)^{-1/4} M(\kappa^{2/3}\zeta) \leq c[f(\xi)\kappa^{2/3}]^{-1/4} \leq c. \tag{6.15}$$

Combining with 1. we obtain the bound for $T_N(s)$ for $s \geq 2s_1$.

3. For $s \in [s_0, 2s_1]$, we have $\xi \in [\xi_2 + s_0\sigma_N/\kappa, \xi_2 + 2s_2\sigma_N/\kappa] = \xi_2 + O(N^{-2/3})$. But $\lim_{\xi\to\xi_2} \hat{f}(\xi) = \dot{\zeta}^2$, so for large N, $\hat{f}(\xi)^{-1/4} \leq c\dot{\zeta}^{-1/2}$. Since $M \leq 1$ and $E \geq 1$, we obtain from the definition of T_N that $T_N(\xi) \leq c\dot{\zeta}^{-1/2} \leq C$.

References

Anderson, T. (1963), 'Asymptotic theory for principal component analysis', *Annals of Mathematical Statistics* **34**, 122–148.

Bai, Z. D. (1999), 'Methodologies in spectral analysis of large dimensional random matrices, a review', *Statistica Sinica* **9**, 611–677.

Baik, J., Deift, P. and Johansson, K. (1999), 'On the distribution of the length of the longest increasing subsequence of random permutations', *Journal of the American Mathematical Society* **12**, 1119–1178.

Buja, A., Hastie, T. and Tibshirani, R. (1995), 'Penalized discriminant analysis', *Annals of Statistics* **23**, 73–102.

Constantine, A. (1963), 'Some non-central distribution problems in multivariate analysis', *Annals of Mathematical Statistics* **34**, 1270–1285.

Deift, P. (1999*a*), 'Integrable systems and combinatorial theory', *Notices of the American Mathematical Society* **47**, 631–640.

Deift, P. (1999*b*), *Orthogonal Polynomials and Random Matrices: A Riemann-Hilbert Approach*, Courant Institute of Mathematical Sciences.

Dunster, T. (1989), 'Uniform asymptotic expansions for Whittaker's confluent hypergeometric functions', *SIAM J. Math. Anal.* **20**, 744–760.

Dyson, F. (1970), 'Correlations between eigenvalues of a random matrix', *Communications in Mathematical Physics* **19**, 235–250.

Eaton, M. L. (1983), *Multivariate Statistics, A Vector Space Approach*, Wiley.

Edelman, A. (1988), 'Eigenvalues and condition numbers of random matrices', *SIAM Journal on Matrix Analysis and Applications* **9**, 543–560.

Erdélyi, A. (1960), 'Asymptotic forms for Laguerre polynomials', *J. Indian Math. Soc. (Golden Jubilee Commemoration Volume)* **24**, 235–250.

Forrester, P. (1993), 'The spectrum edge of random matrix ensembles', *Nuclear Physics* **B402**, 709–728.

Forrester, P. (2000), Painlevé transcendent evaluation of the scaled distribution of the smallest eigenvalue in the Laguerre orthogonal and symplectic ensembles., Technical report, `xxx.lanl.gov` arXiv:nlin.SI/0005064.

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

Gohberg, I. and Krein, M. (1969), *Introduction to the theory of linear nonselfadjoint operators*, American Mathematical Society, Providence. Translations of Mathematical Monographs, 35.

Hastings, S. and McLeod, J. (1980), 'A boundary value problem associated with the second Painlevé transcendent and the Korteweg-de Vries equation', *Archive for Rational Mechanics and Analysis* **73**, 31–51.

Horn, R. and Johnson, C. (1985), *Matrix Analysis*, Cambridge University Press.

James, A. (1964), 'Distributions of matrix variates and latent roots derived from normal samples', *Annals of Mathematical Statistics* **35**, 475–501.

Johansson, K. (2000), 'Shape fluctuations and random matrices', *Communications in Mathematical Physics* **209**, 437–476.

Mardia, K., Kent, J. and Bibby, J. (1979), *Multivariate Analysis*, Academic Press.

Marčenko, V. A. and Pastur, L. A. (1967), 'Distributions of eigenvalues of some sets of random matrices', *Math. USSR-Sb.* **1**, 507–536.

Mehta, M. L. (1991), *Random Matrices, 2nd. edition*, Academic Press.

Muirhead, R. (1974), 'Powers of the largest latent root test of $\Sigma = I$.', *Communications in Statistics* **3**, 513–524.

Muirhead, R. (1982), *Aspects of Multivariate Statistical Theory*, Wiley.

Olver, F. (1974), *Asymptotics and Special Functions*, Academic Press.

Preisendorfer, R. W. (1988), *Principal component analysis in meteorology and oceanography*, Elsevier ; New York.

Riesz, F. and Sz.-Nagy, B. (1955), *Functional Analysis*, Ungar, New York.

Soshnikov, A. (1999), 'Universality at the edge of the spectrum in Wigner random matrices', *Communications in Mathematical Physics* **207**, 697–733.

Stein, C. (1969), Multivariate analysis, I, Technical report, Stanford University, Department of Statistics. Notes prepared by M.L. Eaton in 1966, see pp. 79-81.

Szegö, G. (1967), *Orthogonal Polynomials, 3rd edition*, American Mathematical Society.

Temme, N. M. (1990), 'Asymptotic estimates for Laguerre polynomials', *Journal of Applied Mathematics and Physics (ZAMP)* **41**, 114–126.

Tracy, C. A. and Widom, H. (1994), 'Level-spacing distributions and the Airy kernel', *Communications in Mathematical Physics* **159**, 151–174.

Tracy, C. A. and Widom, H. (1996), 'On orthogonal and symplectic matrix ensembles', *Communications in Mathematical Physics* **177**, 727–754.

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

Tracy, C. A. and Widom, H. (1999), Airy kernel and Painlevé II, Technical report, `xxx.lanl.gov` solv-int/9901004.

Tracy, C. A. and Widom, H. (2000), The distribution of the largest eigenvalue in the Gaussian ensembles, *in* J. van Diejen and L. Vinet, eds, 'Calogero-Moser-Sutherland Models', CRM Series in Mathematical Physics 4, Springer-Verlag, New York, pp. 461–472.

Wachter, K. W. (1976), Probability plotting points for principal components, *in* Hoaglin and Welsch, eds, 'Ninth Interface Symp. Computer Science and Statist.', Prindle, Weber and Schmidt, Boston, MA., pp. 299–308.

Widom, H. (1999), 'On the relation between orthogonal, symplectic and unitary ensembles', *J. Statistical Physics* **94**, 347–363.

Wigner, E. P. (1955), 'Characteristic vectors of bordered matrices of infinite dimensions', *Annals of Mathematics* **62**, 548–564.

Wigner, E. P. (1958), 'On the distribution of the roots of certain symmetric matrices', *Annals of Mathematics* **67**, 325–328.

Wilks, S. (1962), *Mathematical Statistics*, Wiley, New York.

Iain M. Johnstone
Department of Statistics,
Sequoia Hall,
Stanford University,
CA 94305 USA