High dimensional statistical inference and random matrices

Iain M. Johnstone*

Abstract. Multivariate statistical analysis is concerned with observations on several variables which are thought to possess some degree of inter-dependence. Driven by problems in genetics and the social sciences, it first flowered in the earlier half of the last century. Subsequently, random matrix theory (RMT) developed, initially within physics, and more recently widely in mathematics. While some of the central objects of study in RMT are identical to those of multivariate statistics, statistical theory was slow to exploit the connection. However, with vast data collection ever more common, data sets now often have as many or more variables than the number of individuals observed. In such contexts, the techniques and results of RMT have much to offer multivariate statistics. The paper reviews some of the progress to date.

Mathematics Subject Classification (2000). Primary 62H10, 62H25, 62H20; Secondary 15A52.

Keywords. Canonical correlations, eigenvector estimation, largest eigenvalue, principal components analysis, random matrix theory, Wishart distribution, Tracy–Widom distribution.

1. Introduction

Much current research in statistics, both in statistical theory, and in many areas of application, such as genomics, climatology or astronomy, focuses on the problems and opportunities posed by availability of large amounts of data. (More detail may be found, for example, in the paper by Fan and Li [40] in these proceedings.) There might be many variables and/or many observations on each variable. Loosely one can think of each variable as an additional dimension, and so many variables corresponds to data sitting in a high dimensional space. Among several mathematical themes one could follow – Banach space theory, convex geometry, even topology – this paper focuses on random matrix theory, and some of its interactions with important areas of what in statistics is called "multivariate analysis."

Multivariate analysis deals with observations on more than one variable when there is or may be some dependence between the variables. The most basic phenomenon

*The author is grateful to Persi Diaconis, Noureddine El Karoui, Peter Forrester, Matthew Harding, Plamen Koev, Debashis Paul, Donald Richards and Craig Tracy for advice and comments during the writing of this paper, to the Australian National University for hospitality, and to NSF DMS 0505303 and NIH R01 EB001988 for financial support.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

is that of correlation – the tendency of quantities to vary together: tall parents tend to have tall children. From the beginning, there has also been a focus on summarizing and interpreting data by reducing dimension, for example by methods such as principal components analysis (PCA). While there are many methods and corresponding problems of mathematical interest, this paper concentrates largely on PCA as a leading example, together with a few remarks on related problems. Other overviews with substantial statistical content include [5], [30] and [36].

In an effort to define terms and give an example, the earlier sections cover introductory material, to set the stage. The more recent work, in the later sections, concentrates on results and phenomena which appear in an asymptotic regime in which p, the number of variables increases to infinity, in proportion to sample size n.

2. Background

2.1. Principal components analysis. Principal components analysis (PCA) is a standard technique of multivariate statistics, going back to Karl Pearson in 1901 [75] and Harold Hotelling in 1933 [51]. There is a huge literature [63] and interesting modern variants continue to appear [87], [80]. A brief description of the classical method, an example and references are included here for convenience.

PCA is usually described first for abstract random variables, and then later as an algorithm for observed data. So first suppose we have p variables $\mathsf{X}_1, \ldots, \mathsf{X}_p$. We think of these as random variables though, initially, little more is assumed than the existence of a *covariance matrix* $\Sigma = (\sigma_{kk'})$, composed of the mean-corrected second moments

$$\sigma_{kk'} = \mathrm{Cov}(\mathsf{X}_k, \mathsf{X}_{k'}) = E(\mathsf{X}_k - \mu_k)(\mathsf{X}_{k'} - \mu_{k'}).$$

The goal is to reduce dimensionality by constructing a smaller number of "derived" variables $W = \sum_k v_k \mathsf{X}_k$, having variance

$$\mathrm{Var}(W) = \sum_{k,k'} v_k \sigma_{kk'} v_{k'} = \boldsymbol{v}^T \Sigma \boldsymbol{v}.$$

To concentrate the variation in as few derived variables as possible, one looks for vectors that maximize $\mathrm{Var}(W)$. Successive linear combinations are sought that are orthogonal to those previously chosen. The *principal component eigenvalues* ℓ_j and *principal component eigenvectors* $\boldsymbol{v}_j$ are thus obtained from

$$\ell_j = \max \left\{ \boldsymbol{v}^T \Sigma \boldsymbol{v} : \boldsymbol{v}^T \boldsymbol{v}_{j'} = 0;\ j' < j,\ |\boldsymbol{v}| = 1 \right\}. \tag{1}$$

In statistics, it is common to assume a stochastic model in terms of random variables whose distributions contain unknown parameters, which in the present case would be the covariance matrix and its resulting principal components. To *estimate* the unknown parameters of this model we have observed data, assumed to be n observations on each of the p variables. The observed data on variable X_k is viewed as

a vector $\boldsymbol{x}_k \in \mathbb{R}^n$. The vectors of observations on each variable are collected as rows into a $p \times n$ data matrix

$$X = (x_{ki}) = [\boldsymbol{x}_1 \dots \boldsymbol{x}_p]^T.$$

A standard pre-processing step is to center each variable by subtracting the sample mean $\bar{x}_k = n^{-1} \sum_i x_{ki}$, so that $x_{ki} \leftarrow x_{ki} - \bar{x}_k$. After this centering, define the $p \times p$ *sample covariance matrix* $S = (s_{kk'})$ by

$$S = (s_{kk'}) = n^{-1} X X^T, \quad s_{kk'} = n^{-1} \sum_i x_{ki} x_{k'i}.$$

The derived variables in the sample, $\boldsymbol{w} = X\boldsymbol{v} = \sum_k v_k \boldsymbol{x}_k$, have sample variance $\widehat{\mathrm{Var}}(\boldsymbol{w}) = \boldsymbol{v}^T S \boldsymbol{v}$. Maximising this quadratic form leads to successive sample principal components $\hat{\ell}_j$ and $\hat{\boldsymbol{v}}_j$ from the sample analog of (1):

$$\hat{\ell}_j = \max \left\{ \boldsymbol{v}^T S \boldsymbol{v} : \boldsymbol{v}^T \hat{\boldsymbol{v}}_{j'} = 0,\ j' < j,\ |\boldsymbol{v}| = 1 \right\}$$

Equivalently, we obtain for $j = 1, \dots, p$,

$$S\hat{\boldsymbol{v}}_j = \hat{\ell}_j \hat{\boldsymbol{v}}_j, \quad \hat{w}_j = X\hat{\boldsymbol{v}}_j.$$

Note the statistical convention: estimators derived from samples are shown with hats. Figure 1 shows a conventional picture illustrating PCA.

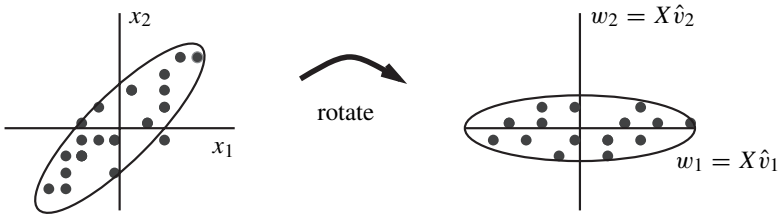


Figure 1. The n data observations are viewed as n points in p dimensional space, the p dimensions corresponding to the variables. The sample PC eigenvectors $\hat{\boldsymbol{v}}_j$ create a rotation of the variables into the new derived variables, with most of the variation on the low dimension numbers. In this two dimensional picture, we might keep the first dimension and discard the second.

Observed data are typically noisy, variable, and limited in quantity, so we are interested in the estimation errors

$$\hat{\ell}_j(X) - \ell_j, \quad \hat{\boldsymbol{v}}_j(X) - \boldsymbol{v}_j.$$

An additional key question in practice is: how many dimensions are "significant", or should be retained? One standard approach is to look at the percent of total variance explained by each of the principal components:

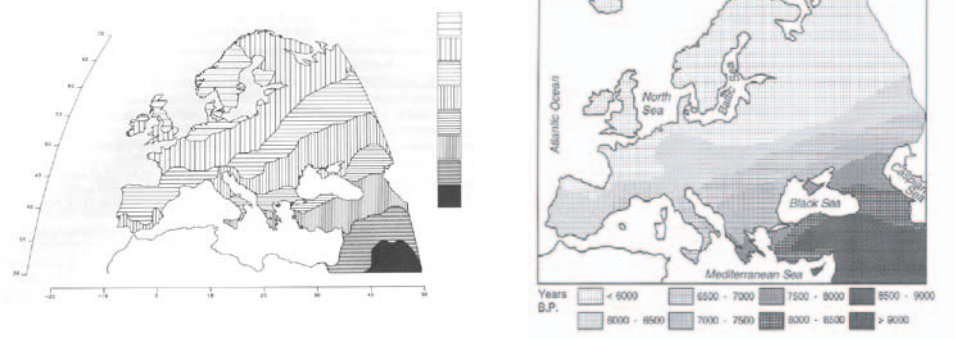
$$p_j = \hat{\ell}_j / \sum \hat{\ell}_{j'} = \hat{\ell}_j / \operatorname{tr} S.$$

An example. Menozzi, Piazza, and Cavalli-Sforza [70] is a celebrated example of the use of PCA in human genetics and anthropology. It was known from archaeological excavations that farming spread gradually from Near East across Europe 9000–5000 yrs ago (map below right). A motivating question was whether this represented spreading of the farmers themselves (and hence their genes) or transfer of technology to pre-existing populations (without a transfer of genes).

Menozzi et al. [70] brought genetic data to bear on the issue. Simplifying considerably, the data matrix X consisted of observations on the frequencies of alleles of $p = 38$ genes in human populations at $n = 400$ locations in Europe. The authors sought to combine information from the 38 genes to arrive at a low dimensional summary.

A special feature of the genetics data is that the observations i have associated locations loc$[i]$, so that it is possible to create a map from each of the principal components w_j, by making a contour plot of the values of the derived variable $w_j[i]$ at each of the sampling locations loc$[i]$. For example the first principal component (map below left) shows a clear trend from south-east to north-west, from Asia Minor to Britain and Scandinavia. The remarkable similarity of the PC map, derived from the gene frequencies, with the farming map, derived from archaeology, has been taken as strong support for the spread of the farmers themselves.

For the genetics data, the first component (out of 38) explains $p_1 = 27\%$ of the variance, the second $p_2 = 18\%$, and the third $p_3 = 11\%$. Thus, and this is typical, more than half the variation is captured in the first three PCs. The second and third, and even subsequent PCs also show patterns with important linguistic and migratory interpretations. For more detail, we refer to books of Cavalli-Sforza [23], [22], from which the maps below are reproduced with the kind permission of the author.



2.2. Gaussian and Wishart distributions. For quantitative analysis, we need more specific assumptions about the process generating the data. The simplest and most conventional model assumes that the p random variables $\mathsf{X}_1, \ldots, \mathsf{X}_p$ follow a p-variate Gaussian distribution $N_p(\mu, \Sigma)$, with mean μ and covariance matrix Σ, and

with probability density function for $\mathsf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_p)$ given by

$$f(\mathsf{X}) = \left|\sqrt{2\pi}\,\Sigma\right|^{-1/2} \exp\big\{ -\tfrac{1}{2}(\mathsf{X}-\mu)^T \Sigma^{-1}(\mathsf{X}-\mu)\big\}.$$

The observed sample is assumed to consist of n independent draws $X_1, \ldots, X_n$ from $\mathsf{X} \sim N_p(\mu, \Sigma)$, collected into a $p \times n$ data matrix $X = [X_1 \ldots X_n]$. When focusing on covariances, it is a slight simplification to assume that $\mu = 0$, as we shall here. In practice, this idealized model of independent draws from a Gaussian is generally at best approximately true – but we may find some reassurance in the dictum "All models are wrong, some are useful." [16]

The (un-normalized) cross product matrix $A = XX^T$ is said to have a p-variate *Wishart* distribution on n degrees of freedom. The distribution is named for John Wishart who in 1928 [97] derived the density function

$$f(A) = c_{n,p}|\Sigma|^{-n/2}|A|^{(n-p-1)/2} \exp\big\{ -\tfrac{1}{2}\operatorname{tr}(\Sigma^{-1}A)\big\},$$

which is supported on the cone of non-negative definite matrices. Here $c_{n,p}$ is a normalizing constant, and it is assumed that Σ is positive definite and that $n \geq p$.

The eigendecomposition of the Wishart matrix connects directly with principal components analysis. Start with a Gaussian data matrix, form the covariance S, yielding a Wishart density for $A = nS$. The eigenvalues and vectors of A, given by

$$Au_i = l_i u_i, \quad l_1 \geq \cdots \geq l_p \geq 0, \tag{2}$$

are related to the principal component eigenvalues and vectors *via*

$$l_i = n\hat{\ell}_i, \quad u_i = \hat{\boldsymbol{v}}_i.$$

Canonical correlations. We digress briefly from the PCA theme to mention one additional multivariate technique, also due to Hotelling [52], since it will help indicate the scope of the results. Given two sets of variables $\mathsf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_p)$ and $\mathsf{Y} = (\mathsf{Y}_1, \ldots, \mathsf{Y}_q)$, with a joint $p+q$-variate Gaussian distribution, we may ask for that linear combination of X that is most correlated with some linear combination of Y, seeking the canonical correlations

$$r_i^2 = \max_{u_i, v_i} \operatorname{Corr}(u_i^T \mathsf{X}, v_i^T \mathsf{Y}),$$

and the maximization is subject to $|u_i| = |v_i| = 1$.

To take an example from climatology [8]: the X variables might be sea surface temperatures at various ocean locations, and the Y variables might be land temperatures at various North American sites. The goal may be to find the combination of sea temperatures that is most tightly correlated with some combination of land temperatures. For a recent example in functional magnetic resonance imaging, see [44].

If we have n draws (X_i, Y_i), $i = 1, \ldots, n$ from the joint distribution, the sample version of this problem may be written as a generalized eigenequation that involves

two independent matrices A and B, each following p-variate Wishart distributions – on q and $n-q$ degrees of freedom respectively:

$$Av_j = r_j^2(A+B)v_j, \quad r_1^2 \geq \cdots \geq r_p^2.$$

The parameters of the Wishart distribution depend on those of the parent Gaussian distribution of the data – if X and Y are independent, then they both reduce to Wishart matrices with identity covariance matrix: $A \sim W_p(q, I)$ and $B \sim W_p(n-q, I)$.

The double Wishart setting. Suppose we have two independent Wishart matrices $A \sim W_p(n_1, I)$ and $B \sim W_p(n_2, I)$, with the degrees of freedom parameters $n_1, n_2 \geq p$. We call this the double Wishart setting. Two remarks: By writing Wishart distributions with *identity* matrices, we emphasize, for now, the "null hypothesis" situation in which there is no assumed structure (compare Section 4). Second, by taking a limit with $n_2 \to \infty$, one recovers the single Wishart setting.

Of central interest are the roots $x_i, i = 1, \ldots, p$ of the generalized eigenproblem constructed from A and B:

$$\det[x(A+B) - A] = 0. \tag{3}$$

The canonical correlations problem is a leading example. In addition, essentially all of the classical multivariate techniques involve an eigendecomposition that reduces to some form of this equation. Indeed, we may collect almost all the chapter titles in any classical multivariate statistics textbook (e.g. [3], [72], [68], [58]) into a table:

Double Wishart	Single Wishart
Canonical correlation analysis	Principal component analysis
Multivariate analysis of variance	Factor analysis
Multivariate regression analysis	Multidimensional scaling
Discriminant analysis	
Tests of equality of covariance matrices	

This table emphasizes the importance of finding the distribution of the roots of (3), which are basic to the use of these methods in applications.

Joint density of the eigenvalues. The joint null hypothesis distribution of the eigenvalues for canonical correlations and principal components was found in 1939. The results were more or less simultaneously obtained by five distinguished statisticians in three continents [41], [45], [54], [71], [81]:

$$f(x_1, \ldots, x_p) = c \prod_i w^{1/2}(x_i) \prod_{i<j} (x_i - x_j), \quad x_1 \geq \cdots \geq x_p, \tag{4}$$

with

$$w(x) = \begin{cases} x^{n-p-1}e^{-x} & \text{single Wishart,} \\ x^{n_1-p-1}(1-x)^{n_2-p-1} & \text{double Wishart.} \end{cases}$$

The normalizing constant c is given, using the multivariate Gamma function $\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma(a - (i-1)/2)$, by

$$c = \begin{cases} \frac{2^{-pn/2} \pi^{p^2/2}}{\Gamma_p(p/2)\Gamma_p(n/2)} & \text{single Wishart,} \\ \frac{\pi^{p^2/2}\Gamma_p((n_1+n_2)/2)}{\Gamma_p(p/2)\Gamma_p(n_1/2)\Gamma_p(n_2/2)} & \text{double Wishart.} \end{cases}$$

Thus, the density has a product term involving each of the roots one at a time, through a weight function w which one recognizes as the weight function for two of the classical families of orthogonal polynomials, Laguerre and Jacobi respectively.

The second product is the so-called "Jacobian" term, which arises in the change of variables to eigenvalue and eigenvector co-ordinates. It is this pairwise interaction term, also recognizable as a Vandermonde determinant (see (13) below), that causes difficulty in the distribution theory.

This result was the beginning of a rich era of multivariate distribution theory in India, Britain, the U.S., and Australia, summarized, for example, in [3], [72], [68]. While some of this theory became so complicated that it lost much influence on statistical practice, with new computational tools and theoretical perspectives the situation may change.

2.3. Random matrices. We detour around this theory and digress a moment to introduce the role of random matrix theory. Beginning in the 1950s, physicists began to use random matrix models to study quantum phenomena. In quantum mechanics the energy levels of a system, such as the nucleus of a complex atom, are described by the eigenvalues of a Hermitian operator H, the Hamiltonian: $H\psi_i = E_i\psi_i$, with $E_0 \leq E_1 \leq \cdots$. The low-lying energy levels can be understood by theoretical work, but at higher energy levels, for example in the millions, the analysis becomes too complicated.

Wigner proposed taking the opposite approach, and sought a purely statistical description of an "ensemble" of energy levels – that could yield properties such as their empirical distribution and the distribution of spacings. He further made the hypothesis that the local statistical behavior of energy levels (or eigenvalues) is well modeled by that of the eigenvalues of a random matrix. Thus the approximation is to replace the Hermitian operator H by a large finite *random* $N \times N$ matrix H_N.

One example of a statistical description that we will return to later is the celebrated SemiCircle Law [95], [96]. This refers to the eigenvalues of a so-called Wigner matrix H_N, with independent and identically distributed entries of mean 0 and a finite variance σ^2. With no further conditions on the distribution of the matrix entries, the empirical distribution $F_N(t) = \#\{i : x_i \leq t\}/N$ of the eigenvalues converges to a limit with density given by a semicircle:

$$dF_N(x\sigma\sqrt{N}) \to \frac{1}{4\pi}\sqrt{4 - x^2}dx.$$

Ensembles and orthogonal polynomials. Quite early on, there was interest in eigenvalue distributions whose densities could be described by more general families of weight functions than the Gaussian. For example, Fox and Kahn [43] used weight functions from the families of classical orthogonal polynomials. Analogies with statistical mechanics made it natural to introduce an additional (inverse temperature) parameter β, so that the eigenvalue density takes the form

$$f(x_1, \ldots, x_N) = c \prod_1^N w(x_i)^{\beta/2} \prod_{i<j} |x_i - x_j|^\beta. \tag{5}$$

At this time, it was only partially realized that in the case $\beta = 1$, these densities were already known in statistics. But the table shows that in fact, the three classical orthogonal polynomial weight functions correspond to the three most basic null eigenvalue distributions in multivariate statistics:

Table 1. The orthogonal polynomials are taken in the standard forms given in Szegö [86].

$w(x) =$				
	$e^{-x^2/2}$	Hermite	H_k	*Gaussian*
	$x^a e^{-x}$	Laguerre	L_k^a	*Wishart*
	$(1-x)^a(1+x)^b$	Jacobi	$P_k^{a,b}$	*Double Wishart*

Dyson [34] showed that physically reasonable symmetry assumptions restricted β to one of three values:

	Symmetry type	Matrix entries
$\beta = 1$	orthogonal	real
$\beta = 2$	unitary	complex
$\beta = 4$	symplectic	quaternion

Mathematically, the complex-valued case is always the easiest to deal with, but of course it is the real case that is of primary statistical (and physical) interest; though cases with complex data do occur in applications, notably in communications.

To summarize, the classical "null hypothesis" distributions in multivariate statistics correspond to the *italicized* eigenvalue densities in the

$$\left\{\begin{matrix}\text{Gaussian}\\ \textit{Laguerre}\\ \textit{Jacobi}\end{matrix}\right\} \left\{\begin{matrix}\textit{orthogonal}\\ \textit{unitary}\\ \text{symplectic}\end{matrix}\right\} \text{ ensemble.}$$

These are often abbreviated to LOE, JUE, etc. We have not italicized the Symplectic case for lack (so far) of motivating statistical applications (though see [4]).

Some uses of RMT in statistics. This table organizes some of the classical topics within RMT, and some of their uses in statistics and allied fields. This paper will

focus selectively (topics in italics), and in particular on largest eigenvalue results and their use for an important class of hypothesis tests, where RMT brings something quite new in the approximations.

Bulk	Graphical methods [92], [93] [finance [15], [77], communications [91]]
Linear Statistics	Hypothesis tests, distribution theory
Extremes	*Hypothesis tests, distribution theory,* role in proofs [21], [33]
Spacings	[[10], otherwise few so far]
General	Computational tools [65], role in proofs

Types of asymptotics. The coincidence of ensembles between RMT and statistical theory is striking, but what can it be *used* for? The complexity of finite sample size distributions makes the use of asymptotic approximations appealing, and here an interesting dichotomy emerges. Traditional statistical approximations kept the number of variables p fixed while letting the sample size $n \to \infty$. This was in keeping with the needs of the times when the number of variables was usually small to moderate.

On the other hand, the nuclear physics models were developed precisely for settings of high energy levels, and so the number of variables in the matrix models were large, as seen in the Wigner semi-circle limit. Interestingly, the many-variables limit of RMT is just what is needed for modern statistical theories with many variables.

	Stat: $\mathbb{C}$Wishart	RMT: Laguerre UE
Density	$\prod_{j=1}^{p} x_j^{n-p} e^{-x_j} \Delta(x)$	$\prod_{j=1}^{N} x_j^{\alpha} e^{-x_j} \Delta(x)$
# variables:	p	N
Sample size:	$n - p$	α

Comparison of the parameters in the statistics and RMT versions of the Wishart density in the table above leads to an additional important remark: in statistics, there is no necessary relationship between sample size n and number of variables p. We will consider below limits in which $p/n \to \gamma \in (0, \infty)$, so that γ could take any positive value. In contrast, the most natural asymptotics in the RMT model would take N large and α fixed. Thus, from the perspective of orthogonal polynomial theory, the statistics models lead to somewhat less usual Plancherel–Rotach asymptotics in which both parameters N *and* α of the Laguerre polynomials are large.

Spreading of sample eigenvalues. To make matters more concrete, we first describe this phenomenon by example. Consider $n = 10$ observations on a $p = 10$ variable Gaussian distribution with identity covariance. The sample covariance matrix follows a Wishart density with $n = p = 10$, and the *population* eigenvalues $\ell_j(I)$ are all equal to 1.

Nevertheless, there is an extreme spread in the *sample* eigenvalues $\hat{\ell}_j = \hat{\ell}_j(S)$, indeed in a typical sample

$$(\hat{\ell}_j) = (.\mathbf{003}, .036, .095, .16, .30, .51, .78, 1.12, 1.40, \mathbf{3.07})$$

and the variation is over three orders of magnitude! Without some supporting theory, one might be tempted to (erroneously) conclude from the sample that the population eigenvalues are quite different from one another.

This spread of sample eigenvalues has long been known, indeed it is an example of the repulsion of eigenvalues induced by the Vandermonde term in (4). It also complicates the estimation of population covariance matrices – also a long standing problem, discussed for example in [85], [47], [98], [27], [66].

The quarter circle law. Marčenko and Pastur [69] gave a systematic description of the spreading phenomenon: it is the version of the semi-circle law that applies to sample covariance matrices. We consider only the special case in which $A \sim W_p(n, I)$. The *empirical distribution function* (or empirical *spectrum*) counts how many sample eigenvalues fall below a given value t:

$$G_p(t) = p^{-1}\#\{\hat{\ell}_j \le t\}.$$

The empirical distribution has a limiting density g^{MP} if sample size n and number of variables p grow together: $p/n \to \gamma$:

$$g^{\mathrm{MP}}(t) = \frac{\sqrt{(b_+ - t)(t - b_-)}}{2\pi\gamma t}, \quad b_\pm = (1 \pm \sqrt{\gamma})^2.$$

The larger p is relative to n, the more spread out is the limiting density. In particular, with $p = n/4$, one gets the curve supported in $\left[\frac{1}{4}, \frac{9}{4}\right]$. For $p = n$, the extreme situation discussed above, the curve covers the full range from 0 to 4, which corresponds to the huge condition numbers seen in the sample.

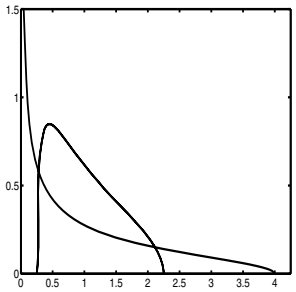


Figure 2. Marčenko–Pastur limit density for $\gamma = \frac{1}{4}$ and $\gamma = 1$.

3. Largest eigenvalue laws

Hypothesis test for largest eigenvalue. Suppose that in a sample of $n = 10$ observations from a $p = 10$ variate Gaussian distribution $N_{10}(0, \Sigma)$, we see a largest sample eigenvalue of 4.25. Is the observed value consistent with an identity covariance matrix (with all population eigenvalues = 1), even though 4.25 lies outside the support interval [0, 4] in the quarter-circle law?

In statistical terms, we are testing a *null hypothesis* of identity covariance matrix, $H_0 : \Sigma = I$ against an *alternative hypothesis* $H_A : \Sigma \neq I$ that Σ has some more general value. Normally, of course, one prefers the simpler model as a description of the data, unless forced by evidence to conclude otherwise.

One might compare 4.25 to random samples of the largest eigenvalue from the null hypothesis distribution (three examples yielding 2.91, 3.40 and 3.50); but what is actually needed is an approximation to the null hypothesis distribution of the largest sample eigenvalue:

$$P\{\hat{\ell}_1 > t : H_0 = W_p(n, I)\}.$$

Tracy–Widom limits. Random matrix theory leads to the approximate distribution we need. In the single Wishart case, assume that $A \sim W_p(n, I)$, either real or complex, that $p/n \to \gamma \in (0, \infty)$ and that $\hat{\ell}_1$ is the largest eigenvalue in equation (2). For the double Wishart case, assume that $A \sim W_p(n_1, I)$ is independent of $B \sim W_p(n_2, I)$, either real or complex together, and that $(p/n_1, p/n_2) \to (\gamma_1, \gamma_2) \in (0, 1)^2$, and that $\hat{\ell}_1$ is the largest generalized eigenvalue in equation (3). With appropriate centering μ_{np} and scaling σ_{np} detailed below, the distribution of the largest eigenvalue approaches one of the Tracy–Widom F_β laws:

$$P\{n\hat{\ell}_1 \leq \mu_{np} + \sigma_{np} s | H_0\} \to F_\beta(s). \tag{6}$$

These laws were first found by Craig Tracy and Harold Widom [88], [89] in the setting of the Gaussian unitary and orthogonal ensembles, *i.e.* (Hermitian) symmetric Gaussian matrices with i.i.d. entries. There are elegant formulas for the distribution functions

$$F_2(s) = \exp\left(-\int_s^\infty (x - s)^2 q(x)\, dx\right), \quad F_1(s)^2 = F_2(s) \exp\left(-\int_s^\infty q(x)\, dx\right).$$

in terms of the solution q to classical (Painlevé II) non-linear second-order differential equation

$$q'' = sq + 2q^3, \quad q(s) \sim \mathrm{Ai}(s) \text{ as } s \to \infty.$$

While q and F_β are somewhat tricky to compute numerically[1], from the point of view of applied data analysis with a software package, it is a special function just like the normal curve.

[1] At time of writing, for available software in MATLAB see http://math.arizona.edu/ momar/research.htm and [31] in S-PLUS see http://www.vitrum.md/andrew/MScWrwck/codes.txt and [9]. Both are based on ideas of [76] [see also [35]]

As will be seen from the explicit formulas (8)- (12) below, the scale of fluctuation σ_{np}/μ_{np} of the largest eigenvalue is $O(n^{-2/3})$ rather than the $O(n^{-1/2})$ seen in the Gaussian domain of attraction. This reflects the constraining effect of eigenvalue repulsion due to the Vandermonde term in (4).

The fact that the same limit arises in the single and double Wishart settings (Laguerre, Jacobi ensembles) is an instance of the universality discussed in P. Deift's paper [29] in this volume. In a different direction, one can modify the assumption that the i.i.d. entries in the $p \times n$ data matrix X are Gaussian. Soshnikov [82] shows that if $n - p = O(p^{1/3})$ and the matrix entries X_{ij} have sufficiently light (subGaussian) tails, then the largest eigenvalue continues to have a limiting Tracy–Widom distribution. The behavior of the largest eigenvalues changes radically with heavy tailed X_{ij} – for Cauchy distributed entries, after scaling by n^2p^2, [83], [84] shows a weak form of convergence to a Poisson process. If the density of the matrix entries behaves like $|x|^{-\mu}$, then [13] give physical arguments to support a phase transition from Tracy–Widom to Poisson at $\mu = 4$.

Second-order accuracy. To demonstrate the relevance of this limiting result for statistical application, it is important to investigate its accuracy when the parameters p and n are not so large. The generic rate of convergence of the left side of (6) to $F_\beta(s)$ is $O(p^{-1/3})$. However, small modifications in the centering and scaling constants μ and σ, detailed in the four specific cases below, lead to $O(p^{-2/3})$ errors, which one might call "second-order accuracy". With this improvement, (6) takes the form

$$\left|P\{n\hat{\ell}_1 \le \mu_{np} + \sigma_{np}s|H_0\} - F_\beta(s)\right| \le Ce^{-cs}p^{-2/3}. \tag{7}$$

This higher-order accuracy is reminiscent of that of the central limit, or normal, approximation to the t-test of elementary statistics for the testing of hypotheses about means, which occurs when the underlying data has a Gaussian distribution.

Single Wishart, complex data. Convergence in the form (6) was first established by Johansson [57] as a byproduct of a remarkable analysis of a random growth model, with

$$\mu^o_{np} = (\sqrt{n} + \sqrt{p})^2, \quad \sigma^o_{np} = (\sqrt{n} + \sqrt{p})\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}}\right)^{1/3}. \tag{8}$$

The second-order result (7) is due to El Karoui [38], If μ'_{np} and σ'_{np} denote the quantities in (8) with n and p replaced by $n + 1/2$ and $p + 1/2$, then the centering μ_{np} is a weighted combination of $\mu'_{n-1,p}$ and $\mu'_{n,p-1}$ and the scaling σ_{np} a similar combination of $\sigma'_{n-1,p}$ and $\sigma'_{n,p-1}$.

Single Wishart, real data. Convergence without rates in the form (6) to $F_1(s)$ with centering and scaling as in (8) is given in [60]. The assumption that $p/n \to \gamma \in (0, \infty)$ can be weakened to $\min\{n, p\} \to \infty$, as shown by El Karoui [37] – this extension is of considerable statistical importance since in many settings $p \gg n$ (see for example [40] in these proceedings).

Analysis along the lines of [61] suggests that the second order result (7) will hold with

$$\mu_{np} = \left(\sqrt{n-\tfrac{1}{2}} + \sqrt{p-\tfrac{1}{2}}\right)^2, \tag{9}$$

$$\sigma_{np} = \left(\sqrt{n-\tfrac{1}{2}} + \sqrt{p-\tfrac{1}{2}}\right)\left(\frac{1}{\sqrt{n-\tfrac{1}{2}}} + \frac{1}{\sqrt{p-\tfrac{1}{2}}}\right)^{1/3}. \tag{10}$$

Double Wishart, complex data. Set $\kappa = n_1 + n_2 + 1$ and define

$$\sin^2\left(\frac{\phi}{2}\right) = \frac{n_1 + \frac{1}{2}}{\kappa}, \quad \sin^2\left(\frac{\gamma}{2}\right) = \frac{p + \frac{1}{2}}{\kappa}. \tag{11}$$

Then

$$\mu_p^o = \sin^2\left(\frac{\phi+\gamma}{2}\right), \quad (\sigma_p^o)^3 = \frac{\sin^4(\phi+\gamma)}{4\kappa^2 \sin\phi \sin\gamma}. \tag{12}$$

The second-order result (7) (currently without the exponential bound, i.e., with $c = 0$) is established in [61] with μ_{np} a weighted combination of μ_p^o and μ_{p-1}^o and the scaling σ_{np} a similar combination of σ_p^o and σ_{p-1}^o.

Double Wishart, real data. Bound (7) is shown in [61] (again still for $c = 0$) with μ_{np} and σ_{np} given by (12) with $\kappa = n_1 + n_2 - 1$.

Approximation vs. tables for $p = 5$. With second-order correction, Tracy–Widom approximation turns out to be surprisingly accurate. William Chen [24], [25], [26] has computed tables of the exact distribution in the double Wishart, real data, case that cover a wide range of the three parameters p, n_1 and n_2, and allow a comparison with the asymptotic approximation. Even for $p = 5$ variables, the TW approximation is quite good, Figure 3, across the entire range of n_1 and n_2.

A different domain of attraction. The Tracy–Widom laws are quite different from other distributions in the standard statistical library. A full probabilistic understanding of their origin is still awaited (but see [78] for a recent characterization in terms of the low lying eigenvalues of a random operator of stochastic diffusion type). Instead, we offer some incomplete remarks as prelude to the original papers [88], [89].

Since one is looking at the largest of many eigenvalues, one might be reminded of extreme value theory, which studies the behavior of the largest of a collection of variables, which in the simplest case are independent. However, extreme value theory exploits the independence to study the maximum via products: $\{\max_{1\le i\le p} l_i \le t\} = \prod_{i=1}^p I\{l_i \le t\}$ For eigenvalues, however, the Jacobian term, or Vandermonde determinant,

$$V(l) = \prod_{i<j}(l_j - l_i) = \det[l_i^{k-1}]_{1\le i,k\le p}, \tag{13}$$

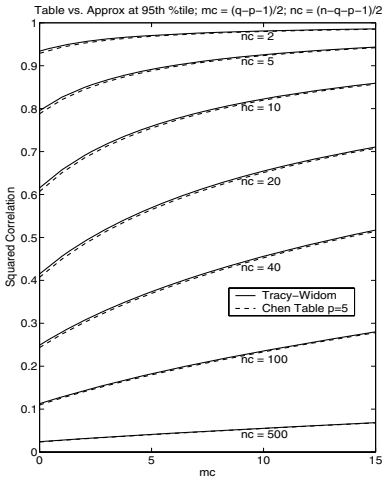


Figure 3. A comparison of the 95th percentile, relevant for hypothesis tests, from Chen's table (dashed line) and the Tracy–Widom approximation (solid line). Chen's parameters m_c, n_c are related to our double Wishart parameters n_1, n_2 by $m_c = (n_1 - p - 1)/2, n_c = (n_2 - p - 1)/2$.

changes everything. The theory uses the inclusion-exclusion relation:

$$\prod_{i=1}^{p} I\{l_i \le t\} = \sum_{k=0}^{p} (-1)^k \binom{p}{k} \prod_{i=1}^{k} I\{l_i > t\}.$$

The product structure of the left side, central to extreme value theory, is discarded in favor of the right side, which leads to an expression for $P\{\max_{1\le i\le p} l_i \le t\}$ in terms of so-called Fredholm determinants.

For example, it is shown by Tracy and Widom [90] that for complex data

$$P\{\max l_i \le t\} = \det(I - K_p \chi_{(t,\infty)}),$$

where χ_I is the indicator function for interval I, and $K_p \colon L_2 \to L_2$ is an operator whose kernel is the two-point correlation function

$$K_p(x, y) = \sum_{k=1}^{p} \phi_k(x)\phi_k(y),$$

written in terms of weighted orthonormal polynomials $\phi_k = h_k^{-1/2} w^{1/2} p_k$, where the polynomials p_k and weight functions w are given in Table 1 for the Gaussian, Wishart and double Wishart settings respectively.

For real data, Tracy and Widom [90] show that

$$P\{\max l_i \le t\} = \sqrt{\det(I - \mathcal{K}_p \chi_{(t,\infty)})},$$

where $\mathcal{K}_p$ is now a 2×2 matrix-valued operator on $L_2 \otimes L_2$. The corresponding kernel has form

$$\mathcal{K}_p(x, y) = \begin{pmatrix} \tilde{K}_p & -D_2\tilde{K}_p \\ \varepsilon_1\tilde{K}_p & \tilde{K}_p^T \end{pmatrix},$$

where $\tilde{K}_p = K_p + r_1$ and r_1 is a rank one kernel described in the three cases in more detail in [1], [42], [61]. Here D_2 and ε_1 denote partial differentiation and integration with respect to second and first variables respectively.

The expressions are thus somewhat more complicated in the real data case of primary interest in statistics. However they are amenable to analysis and approximation using orthogonal polynomial asymptotics near the largest zero, and to analysis based on the error terms to get the higher order approximation.

Back to the example. We asked if an observed largest eigenvalue of 4.25 was consistent with $H_0 : \Sigma = I$ when $n = p = 10$. The Tracy–Widom approximation using moments (9)–(10) yields a 6% chance of seeing a value more extreme than 4.25 even if "no structure" is present. Against the traditional 5% benchmark, this would not be strong enough evidence to discount the null hypothesis.

This immediately raises a question about the *power* of the largest root test, namely evaluation of

$$P\{\hat{\ell}_1 > t : W_p(n, \Sigma)\}$$

when $\Sigma \neq I$. How different from 1 does $\lambda_{\max}(\Sigma)$ need to be before H_0 is likely to be rejected? To this we now turn.

4. Beyond the null hypothesis

From the perspective of multivariate distribution theory, we have, in a sense, barely scratched the surface with the classical RMT ensembles, since they correspond to symmetric situations with no structure in the population eigenvalues or covariance matrix. Basic statistical quantities like power of tests and confidence intervals, as well as common applications in signal processing, genetics or finance, call for distributions under structured, asymmetric values for the covariance matrix Σ.

Statistical theory (pioneered by Alan James [56, e.g.], and summarized in the classic book by Robb Muirhead [72]) gives expressions for the classical multivariate eigenvalue distributions in more general settings, typically in terms of hypergeometric functions of matrix argument. For example, if $L = \operatorname{diag}(l_i)$ are the eigenvalues of $A \sim W_p(n, \Sigma)$, then the joint eigenvalue density

$$\frac{f_\Sigma(l_1, \ldots, l_p)}{f_I(l_1, \ldots, l_p)} = |\Sigma|^{-n/2} \exp\left\{\tfrac{1}{2} \operatorname{tr} L\right\} {}_0F_0\big(-\tfrac{1}{2}\Sigma^{-1}, L\big),$$

with

$$ {}_0F_0(S,T) = \int_{O(p)} \exp\{\mathrm{tr}(SUTU^T)\}\, dU, \tag{14} $$

and dU normalized Haar measure, but many other versions occur in the general theory. Despite recent major advances in computation by Alan Edelman and Plamen Koev [65], [64], and considerable work on the use of Laplace approximations (see e.g. [19], [20]), statistical theory would benefit from further serviceable approximations to these typically rather intractable objects.

Persistence of the Tracy–Widom limit. One basic question asks, in the setting of principal components analysis, for what conditions on the covariance Σ does the Tracy–Widom approximation continue to hold,

$$ P\{\hat{\ell}_1 \le \mu_{np}(\Sigma) + \sigma_{np}(\Sigma)s\} \to F_\beta(s), \tag{15} $$

perhaps with modified values for centering and scaling to reflect the value of Σ?

Fascinating answers are beginning to emerge. For example, El Karoui [39] establishes that (15) holds, along with explicit formulas for $\mu_{np}(\Sigma)$ and $\sigma_{np}(\Sigma)$, if enough eigenvalues accumulate near the largest eigenvalue, or if a small number of eigenvalues are not too isolated, as we describe below in a specific setting below.

Some of the results are currently restricted to complex data, because they build in a crucial way on the determinantal representation of the unitary matrix integral (the complex analog of (14))

$$ \int_{U(p)} \exp\{\mathrm{tr}\,\Sigma^{-1}ULU^*\}\, dU = c\, \frac{\det(e^{\pi_j l_k})}{V(\pi)V(l)} \tag{16} $$

known as the Harish-Chandra–Itzykson–Zuber formula [50], [55], see also [46]. Here the eigenvalues of Σ^{-1} are given by $\mathrm{diag}(\pi_j)$ and $V(l)$ is the Vandermonde determinant (13). While it is thought unlikely that there are direct analogs of (16), we very much need extensions of the distributional results to real data: there are some results in the physics literature [17], but any statistical consequences are still unclear.

Finite rank perturbations. We focus on a simple concrete model, and describe a phase transition phenomenon. Assume that

$$ \Sigma = \mathrm{diag}(\ell_1, \ldots, \ell_M, \sigma_e^2, \ldots, \sigma_e^2), \tag{17} $$

so that a fixed number M of population eigenvalues are greater than the base level σ_e^2, while both dimensions p and n increase in constant ratio $p/n \to \gamma \in (0, \infty)$.

First some heuristics: if all population eigenvalues are equal, then the largest sample eigenvalue $\hat{\ell}_1$ has $n^{-2/3}$ fluctuations around the upper limit of the support of the Marčenko–Pastur quarter circle law, the fluctuations being described by the Tracy–Widom law. For simplicity, consider $M = 1$ and $\sigma_e^2 = 1$. If ℓ_1 is large

and so very clearly separated from the bulk distribution, then one expects Gaussian fluctuations of order $n^{-1/2}$, and this is confirmed by standard perturbation analysis.

Baik et al. [7] describe, for *complex* data, a 'phase transition' that occurs between these two extremes. If $\ell_1 \le 1 + \sqrt{\gamma}$, then

$$n^{2/3}(\hat{\ell}_1 - \mu)/\sigma \Rightarrow \begin{cases} F_2 & \ell_1 < 1 + \sqrt{\gamma}, \\ \tilde{F}_2 & \ell_1 = 1 + \sqrt{\gamma} \end{cases}$$

where, from (8), we may set

$$\mu = (1 + \sqrt{\gamma})^2, \quad \sigma = (1 + \sqrt{\gamma})\left(1 + \sqrt{\gamma^{-1}}\right)^{1/3},$$

and $\tilde{F}_2$ is related to F_2 as described in Baik et al. [7]. On the other hand, if $\ell_1 > 1+\sqrt{\gamma}$,

$$n^{1/2}(\hat{\ell}_1 - \mu(\ell_1))/\sigma(\ell_1) \Rightarrow N(0, 1),$$

with

$$\mu(\ell_1) = \ell_1\left(1 + \frac{\gamma}{\ell_1 - 1}\right), \quad \sigma^2(\ell_1) = \ell_1^2\left(1 - \frac{\gamma}{(\ell_1 - 1)^2}\right). \tag{18}$$

Thus, below the phase transition the distribution of $\hat{\ell}_1$ is unchanged, Tracy–Widom, regardless of the value of ℓ_1. As ℓ_1 increases through $1 + \sqrt{\gamma}$, the law of $\hat{\ell}_1$ jumps to Gaussian and the mean increases with ℓ_1, but is biased low, $\mu(\ell_1) < \ell_1$, while the variance $\sigma^2(\ell_1)$ is lower than its value, ℓ_1^2, in the limit with p fixed.

A key feature is that the phase transition point $1 + \sqrt{\gamma}$, located at the zero of $\sigma(\ell_1)$, is buried deep inside the bulk, whose upper limit is $(1 + \sqrt{\gamma})^2$. A good heuristic explanation for this location is still lacking, though see El Karoui [39].

Further results on almost sure and Gaussian limits for both real and complex data, and under weaker distributional assumptions have been obtained by Paul [74] and Baik and Silverstein [6].

A recent example. Harding [49] illustrates simply this phase transition phenomenon in a setting from economics and finance. In a way this is a negative example for PCA; but statistical theory is as concerned with describing the limits of techniques as their successes.

Factor analysis models, of recently renewed interest in economics, attempt to "explain" the prices or returns of a portfolio of securities in terms of a small number of common "factors" combined with security-specific noise terms. It has been further postulated that one could estimate the number and sizes of these factors using PCA. In a 1989 paper that is widely cited and taught in economics and finance, Brown [18] gave a realistic simulation example that challenged this view, in a way that remained incompletely understood until recently.

Brown's example postulated four independent factors, with the parameters of the model calibrated to historical data from the New York Stock Exchange. The return in period t of security k is assumed to be given by

$$R_{kt} = \Sigma_{\nu=1}^{4} b_{k\nu} f_{\nu t} + e_{kt}, \quad k = 1, \ldots, p, \ t = 1, \ldots, T, \tag{19}$$

where it is assumed that $b_{k\nu} \sim N(\beta, \sigma_b^2)$, $f_{\nu t} \sim N(0, \sigma_f^2)$ and $e_{\nu t} \sim N(0, \sigma_e^2)$, all independently of one another. The population covariance matrix has the form (17) with $M = 4$ and

$$\ell_j = p\sigma_f^2(\sigma_b^2 + 4\beta\delta_{j1}) + \sigma_e^2, \quad j = 1, \ldots, 4. \tag{20}$$

Here δ_{j1} is the Kronecker delta, equal to 1 for $j = 1$ and 0 otherwise. Figure 4 (a) plots the population eigenvalues ℓ_1 (the dominant 'market' factor), the common value $\ell_2 = \ell_3 = \ell_4$ and the base value $\ell_5 = \sigma_e^2$ against p, the number of securities in the portfolio. One might expect to be able to recover an estimate of ℓ_2 from empirical data, but this turns out to be impossible for $p \in [50, 200]$ when $T = 80$ as shown in Figure 4 (b). First, the range of observed values of the top or market eigenvalue is biased upward from the true top eigenvalue. In addition, there are many sample eigenvalues above the anticipated value for ℓ_2.

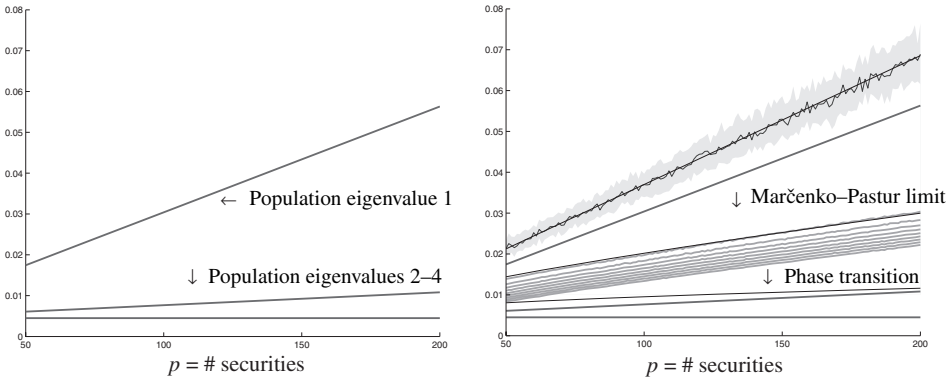


Figure 4. Population and sample eigenvalues for a four factor model (19) with $\beta = 0.6, \sigma_b = .4, \sigma_f = .01257, \sigma_e = .0671$. [Brown and Harding use $\beta = 1, \sigma_b = .1$; the values are modified here for legibility of the plot.] (a) Left panel: Population eigenvalues according to (20) (b) Right panel: The top sample eigenvalue in replications spreads about a sample average line which tracks the solid line given by (18), in particular overestimating the population value ℓ_1. The next nine sample eigenvalues fall at or below the Marčenko–Pastur upper limit, swamping the next three population eigenvalues.

Harding shows that one can directly apply the (real version) of the phase transition results previously discussed to fully explain Brown's results. Indeed, the inability to identify factors is because they fall on the wrong side of the phase transition $\sigma_e^2(1 + \sqrt{p/T})$, and so we can not expect the observed eigenvalue estimates to

exceed the Marčenko–Pastur upper bound $\sigma_e^2(1+\sqrt{p/T})^2$. Finally, the bias between the observed and true values of the top eigenvalue is also accurately predicted by the random matrix formulas (18).

5. Estimating eigenvectors

Most of the literature at the intersection of random matrix theory and statistics is focused on eigenvalues. We close with a few remarks on the estimation of eigen*vectors*. Of course, the question is only meaningful in non-symmetric settings when the covariance matrix Σ is not proportional to I. We again assume that $S \sim W_p(n, \Sigma)$ and now focus attention on covariance models which are a finite-rank perturbation of the identity[2]:

$$\Sigma = \sigma^2 I + \sum_{\nu=1}^{M} \lambda_\nu \theta_\nu \theta_\nu^T, \tag{21}$$

with $\lambda_1 \geq \cdots \geq \lambda_M > 0$ and $\{\theta_\nu\}$ orthonormal. We ask how well can the population eigenvectors θ_ν be estimated when both p and n are large.

First some remarks on how model (21) can arise from an *orthogonal factor* or *variance components* model for the data. Assume that the p-dimensional observations $X_i, i = 1, \ldots, n$ have the form

$$X_i = \mu + \sum_{\nu=1}^{M} \sqrt{\lambda_\nu} v_{\nu i} \theta_\nu + \sigma Z_i,$$

where $\{v_{\nu i} : 1 \leq \nu \leq M\}$ are i.i.d. $N(0,1)$, independently of $Z_i \sim N_p(0, I_p)$, for all i. If we further assume, for convenience, that $\mu = 0$, then with the sample covariance S defined as in Section 2, then $S \sim W_p(n, \Sigma)$. If we express X_i, θ_ν and Z_i in (22) in terms of co-ordinates in a suitable basis $\{e_k, k = 1, \ldots, p\}$ and write $f_{\nu i} = \sqrt{\lambda_\nu} v_{\nu i}$ we obtain

$$X_{ki} = \sum_{\nu=1}^{M} \theta_{k\nu} f_{\nu i} + \sigma Z_{ki},$$

in which $\theta_{k\nu}$ is viewed as the factor loading of the kth variable on the νth factor, and $f_{\nu i}$ is the factor score of the νth factor for the ith individual. As we have seen in (19) in the previous section, in economics X_{ki} may represent the return on the kth security in time period i.

[2]The situation is different in *functional* principal components analysis, where smoothness of the observed data (functions) leads to covariance matrices with smoothly decaying eigenvalues. For entries into this literature, see for example [28], [14], [48].

Assume that $\lambda_1 > \cdots > \lambda_M > 0$. Let $\hat{\theta}_\nu$ denote the normalized sample eigenvectors of S (denoted $\hat{\boldsymbol{v}}_\nu$ in Section 2.1) associated with the M largest sample eigenvalues. In classical asymptotics, with n large and p fixed, there is a well understood Gaussian limit theory:

$$\sqrt{n}(\hat{\theta}_\nu - \theta_\nu) \to N_p(0, \Gamma_\nu) \tag{22}$$

where Γ_ν is given, for example, in [2], [3].

The situation is radically different when $p/n \to \gamma > 0$ – indeed, ordinary PCA is necessarily inconsistent:

$$\langle \hat{\theta}_\nu, \theta_\nu \rangle \to \begin{cases} 0 & \lambda_\nu \in [0, \sqrt{\gamma}], \\ \frac{1-\gamma/\lambda_\nu^2}{1+\gamma/\lambda_\nu} & \lambda_\nu > \sqrt{\gamma}. \end{cases}$$

For signal strengths λ below the phase transition just discussed, nothing can be estimated – the estimate is asymptotically orthogonal to the truth. The angle decreases as λ_ν grows, but is never exactly consistent.

This result has emerged in several literatures, starting in the learning theory/statistical physics community, with non-rigorous arguments based on the replica method [79], [53], where this phenomenon has been termed "retarded learning" [11], [94]. More recently, rigorous results have been obtained [62], [74], [73].

To obtain consistent estimates, further assumptions are needed. One plausible situation is that in which there exists a basis $\{e_k\}_{k=1:p}$ in which it is believed that the vectors θ_ν have a sparse representation. In microarray genetics, for example X_{ki} might be the expression of gene k in the ith patient, and it may be believed that (in the standard basis) each factor ν is related to only a small number of genes [67]. In EEG studies of the heart, the beat-to-beat cycle might be expressed in a wavelet basis, in which the components of variation θ_ν may well be sparsely represented [62].

We briefly describe results in the sparse setting of work in progress by D. Paul, and by Paul and the author. For simplicity only, we specialize to $M = 1$. The error of estimation, or loss, of $\hat{\theta}$ is measured on unit vectors by

$$L(\hat{\theta}, \theta) = \|\hat{\theta} - \mathrm{sign}(\langle \hat{\theta}, \theta \rangle)\theta\|^2 = 4 \sin^2 \tfrac{1}{2} \angle(\hat{\theta}, \theta).$$

If $\hat{\theta}$ is now the ordinary PCA estimate of θ, and if $p/n \to \gamma > 0$, then to first order,

$$EL(\hat{\theta}, \theta) = \frac{p}{nh(\lambda)}(1 + o(1)), \quad h(\lambda) = \frac{\lambda^2}{1+\lambda},$$

from which it is natural to define the "per-variable" noise level $\tau_n = 1/\sqrt{nh(\lambda)}$.

As is common in non-parametric estimation theory, we use ℓ_q norm, $q < 2$, as a measure of sparsity: with $\|\theta\|_q^q = \sum_k |\theta_k|^2$, define $\Theta_q(C) = \{\theta \in S^{p-1} : \|\theta\|_q \le C\}$. Paul proposes a two-step procedure for selecting a reduced subset of variables on which to perform PCA, resulting in an estimator $\hat{\theta}^P$ for which

$$\sup_{\theta \in \Theta_q(C)} EL(\hat{\theta}^P, \theta) \le K(C) \log p \cdot m_n \tau_n^2. \tag{23}$$

Here m_n is an effective dimension parameter, equal to $(C^2/(\tau^2 \log p))^{q/2}$ in the "sparse" case when this is smaller than $c_1 p$, and equal to p in the contrary "dense" case. Lower bounds are obtained that show that this estimation error is optimal, in a minimax sense, up to factors that are at most logarithmic in p.

Bounds such as (23) are reminiscent of those for estimation of sparse *mean* sequences in white Gaussian noise [32], [12], [59]. An observation due to Paul provides a link between eigenvector estimation and the estimation of means. Again with $M = 1$ for simplicity, let $\hat{\theta}$ be the ordinary PCA estimate of θ. Write $\hat{C} = \langle \hat{\theta}, \theta \rangle$ and $\hat{\theta}^\perp = \hat{\theta} - C\theta$. Then, with $\hat{S}^2 = 1 - \hat{C}^2$, in the decomposition

$$\hat{\theta} = \hat{C}\theta + \hat{S}U, \quad U = \hat{\theta}^\perp / \|\hat{\theta}^\perp\|$$

it happens that U is uniformly distributed on a copy of S^{p-2}, independently of $\hat{S}$.

It is a classical remark that a high-dimensional isotropic Gaussian vector is essentially concentrated uniformly on a sphere. We may reverse this remark by starting with a uniform distribution on a sphere, and introducing an ultimately inconsequential randomization with $R^2 \sim \chi^2_{p-1}/p$ and $z_1 \sim N(0, 1/p)$ with the result that $z = RU + z_1\theta$ has an $N_p(0, I)$ distribution. This leads to a signal-in-Gaussian-noise representation

$$Y = \hat{C}\theta + \tau^2 z, \quad \tau^2 = 1/(2nh(\hat{\lambda})),$$

Work is in progress to use this connection to improve the extant estimation results for eigenvectors.

6. Coda

One may expect a continuing fruitful influence of developments in random matrix theory on high dimensional statistical theory, and perhaps even some flow of ideas in the opposite direction. A snapshot of current trends may be obtained from http://www.samsi.info/workshops/2006ranmat-opening200609.shtml, being the presentations from the opening workshop of a semester devoted to high dimensional inference and random matrices at the NSF Statistics and Applied Mathematics Institute in Fall 2006.

References

[1] Adler, M., Forrester, P. J., Nagao, T., and van Moerbeke, P., Classical skew orthogonal polynomials and random matrices. *J. Statist. Physics* **99** (1–2) (2000), 141–170.

[2] Anderson, T. W., Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34** (1963), 122–148.

[3] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*. 2nd ed., Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., John Wiley & Sons, Inc., New York 1984.

[4] Andersson, S. A., Brøns, H. K., and Jensen, S. T., Distribution of eigenvalues in multivariate statistical analysis. *Ann. Statist.* **11** (2) (1983), 392–415.

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

[6] Baik, J., and Silverstein, J. W., Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** (2006), 1382–1408.

[7] Baik, J., Ben Arous, G., and Péché, S., Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** (5) (2005), 1643–1697.

[8] Barnett, T. P., and Preisendorfer, R., Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review* **115** (1987), 1825–1850.

[9] Bejan, A., Largest eigenvalues and sample covariance matrices. Tracy-Widom and Painlevé II: computational aspects and realization in S-Plus with applications. http://www.vitrum.md/andrew/TWinSplus.pdf, 2005.

[10] Ben Arous, G., and Péché, S., Universality of local eigenvalue statistics for some sample covariance matrices. *Comm. Pure Appl. Math.* **58** (10) (2005), 1316–1357.

[11] Biehl, M., and Mietzner, A., Statistical mechanics of unsupervised structure recognition. *J. Phys. A* **27** (6) (1994), 1885–1897.

[12] Birgé. L., and Massart, P., Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** (2001), 203–268.

[13] Biroli, G., Bouchaud, J.-P., and Potters, M., On the top eigenvalue of heavy-tailed random matrices. Preprint, 2006; arXiv:cond-mat/0609070.

[14] Bosq, D., *Linear processes in function spaces*. Lecture Notes in Statist. 149, Springer-Verlag, New York 2000.

[15] Bouchaud, J.-P., and Potters, M., *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press, Cambridge 2003.

[16] Box, G. E. P., Robustness in the strategy of scientific model building. In R. L. Launer and G. N. Wilkinson, editors, *Robustness in Statistics* (R. L. Launer and G. N. Wilkinson, eds.), Academic Press, New York 1979, 201–236.

[17] Brézin, E., and Hikami, S., New correlation functions for random matrices and integrals over supergroups. *J. Phys. A* **36** (3) (2003), 711–751.

[18] Brown, S. J., The number of factors in security returns. *J. Finance* **XLIV** (5) (1989), 1247–1261.

[19] Butler, R. W., and Wood, A. T. A., Laplace approximations for hypergeometric functions with matrix argument. *Ann. Statist.* **30** (4) (2002), 1155–1177.

[20] Butler, R. W., and Wood, A. T. A., Laplace approximations to hypergeometric functions of two matrix arguments. *J. Multivariate Anal.* **94** (1) (2005), 1–18.

[21] Candès, E., and Tao, T., Near Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? *IEEE Trans. Inform. Theory* **52** (12) (2006), 5406–5425.

[22] Cavalli-Sforza, L. L., *Genes, peoples, and languages*. North Point Press, New York 2000.

[23] Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A., *The history and geography of human genes*. Princeton University Press, Princeton, N.J., 1994.

[24] Chen, W. R., Some new tables of the largest root of a matrix in multivariate analysis: A computer approach from 2 to 6, 2002. Presented at the 2002 American Statistical Association.

[25] Chen, W. R., Table for upper percentage points of the largest root of a determinantal equation with five roots. *InterStat* (5), February 2003. http://interstat.statjournals.net/YEAR/2003/abstracts/0302005.php.

[26] Chen, W. R., The new table for upper percentage points of the largest root of a determinantal equation with seven roots. *InterStat* (1), September 2004. http://interstat.statjournals.net/YEAR/2004/abstracts/0409001.php.

[27] Daniels, M. J., and Kass, R. E., Shrinkage estimators for covariance matrices. *Biometrics* **57** (4) (2001), 1173–1184.

[28] Dauxois, J., Pousse, A., and Romain, Y., Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.* **12** (1) (1982), 136–154.

[29] Deift, P., Universality for mathematical and physical systems. *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume I, EMS Publishing House, Zürich 2007, 125–152.

[30] Diaconis, P., Patterns in eigenvalues: the 70th Josiah Willard Gibbs lecture. *Bull. Amer. Math. Soc.* (N.S.) **40** (2) (2003), 155–178.

[31] Dieng, M., Distribution Functions for Edge Eigenvalues in Orthogonal and Symplectic Ensembles: Painlevé Representations II. arXiv:math.PR/0506586.

[32] Donoho, D. L., and Johnstone, I. M., Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81** (1994), 425–455.

[33] DDonoho, D. L., For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.* **59** (6) (2006), 797–829.

[34] Dyson, F. J., The threefold way. Algebraic structure of symmetry groups and ensembles in quantum mechanics. *J. Math. Phys.* **3** (6) (1962), 1199–1215.

[35] Edelman, A., and Persson, P.-O., Numerical Methods for Eigenvalue Distributions of Random Matrices. Preprint, 2005; arXiv:math-ph/0501068.

[36] Edelman, A., and Rao, N. R., Random matrix theory. *Acta Numer.* **14** (2005), 233–297.

[37] El Karoui, N., On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. Preprint, 2003; arXiv:math. ST/0309355.

[38] El Karoui, N., A rate of convergence result for the largest eigenvalue of complex white Wishart Matrices. *Ann. Probab.* **34** (2006), 2077–2117.

[39] El Karoui, N., Tracy-Widom limit for the largest eigenvalue of a large class of complex Wishart matrices. *Ann. Probab.* **35** (2007).

[40] Fan J., and Li, R., Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume III, EMS Publishing House, Zürich 2006, 595–622.

[41] Fisher, R. A., The sampling distribution of some statistics obtained from non-linear equations. *Ann. Eugenics* **9** (1939), 238–249.

[42] Forrester, P. J., Log-gases and Random matrices. http://www.ms.unimelb. edu.au/~matpjf/matpjf.html. Book manuscript, 2004.

[43] Fox, D., and Kahn, P. B., Higher order spacing distributions for a class of unitary ensembles. *Phys. Rev.* **134** (5B) (1964), B1151–B1155.

[44] Friman, O., Cedefamn, J., Lundberg, P., Borga, M., and H. Knutsson, H., Detection of neural activity in functional MRI using canonical correlation analysis. *Magnetic Resonance in Medicine* **45** (2001), 323–330.

[45] Girshick, M. A., On the sampling theory of roots of determinantal equations. *Ann. Math. Statist.* **10** (1939), 203–224.

[46] Gross, K. I., and Richards, D. St. P., Total positivity, spherical series, and hypergeometric functions of matrix argument. *J. Approx. Theory* **59** (2) (1989), 224–246.

[47] Haff, L. R., The variational form of certain Bayes estimators. *Ann. Statist.* **19** (1991), 1163–1190.

[48] Hall, P., Müller, H.-G., and Wang, J.-L., Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** (3) (2006), 1493–1517.

[49] Harding, M. C., Explaining the single factor bias of arbitrage pricing models in finite samples. Dept. of Economics, MIT, 2006; http://www.mit.edu/~mharding/.

[50] Harish-Chandra, Differential operators on a semisimple Lie algebra. *Amer. J. Math.* **79** (1) (1957), 87–120.

[51] Hotelling, H., Analysis of a complex of statistical variables into principal components. *J. Educational Psychology* **24** (1933), 417–441, 498–520.

[52] Hotelling, H., Relations between two sets of variates. *Biometrika* **28** (1936), 321–377.

[53] Hoyle, D. C., and Rattray, M., Principal-component-analysis eigenvalue spectra from data with symmetry breaking structure. *Phys. Rev. E* **69** (2004), (026124).

[54] Hsu, P. L., On the distribution of roots of certain determinantal equations. *Ann. Eugenics* **9** (1939), 250–258.

[55] Itzykson, C., and Zuber, J.-B., The planar approximation. II. *J. Math. Phys.* **21** (3) (1980), 411–421.

[56] James, A. T., Distributions of matrix variates and latent roots derived from normal samples. *Ann. Math. Statist.* **35** (1964), 475–501.

[57] Johansson, K., Shape fluctuations and random matrices. *Commun. Math. Phys.* **209** (2000), 437–476.

[58] Johnson, R. A., and Wichern, D. W., *Applied Multivariate Statistical Analysis*. 5th ed., Prentice Hall, Englewood Cliffs, NJ, 2002.

[59] Johnstone, I. M., and Silverman, B. W., Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** (2004), 1594–1649.

[60] Johnstone, I. M., On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** (2001), 295–327.

[61] Johnstone, I. M., Canonical correlation analysis and Jacobi ensembles: Tracy-Widom limits and rates of convergence. Manuscript, 50pp, August 2006.

[62] Johnstone, I. M., and Lu, A. Y., Sparse principal components analysis. Technical report, Stanford University, Dept. of Statistics, 2004; tentatively accepted, *J. Appl. Statist. Sci.*

[63] Jolliffe, I. T., *Principal Component Analysis*. Springer Ser. Statist., Springer, 2nd edition, New York 2002.

[64] Koev, P., Software `mhg`, `mhgi` for hypergeometric function of a matrix argument. 2006; http://www-math.mit.edu/~plamen/.

[65] Koev, P., and Edelman, A., The efficient evaluation of the hypergeometric function of a matrix argument. *Math. Comp.* **75** (254) (2006), 833–846.

[66] Ledoit, O., and Wolf, M., A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Analysis* **88** (2004), 365–411.

[67] Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M., Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics* (K. A. Do, P. Mueller, and M. Vannucci, eds.), Cambridge University Press, Cambridge 2006, 155–176.

[68] Mardia, K. V., Kent, J. T., and Bibby, J. M., *Multivariate Analysis*. Academic Press, London, New York, Toronto, Ont., 1979.

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

[70] Menozzi, P., Piazza, A., and Cavalli-Sforza, L., Synthetic maps of human gene frequencies in Europeans. *Science* **201** (4358) (1978), 786–792.

[71] Mood, A. M., On the distribution of the characteristic roots of normal second-moment matrices. *Ann. Math. Statist.* **22** (1951), 266–273.

[72] Muirhead, R. J., *Aspects of Multivariate Statistical Theory*. Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., John Wiley & Sons, Inc., New York 1982.

[73] Onatski, A., Asymptotic distribution of the principal components estimator of large factor models when factors are relatively weak. Dept. of Economics, Columbia University, 2006; http://www.columbia.edu/~ao2027/papers1.html.

[74] Paul, D., Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Technical report, Department of Statistics, Stanford University, 2004; *Statist. Sinica*, to appear.

[75] Pearson, K., On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2** (6) (1901), 559–572,

[76] Persson, P.-O., Numerical methods for random matrices. 2002; http://www.mit.edu/~persson/numrand_report.pdf.

[77] Potters, M., Bouchaud, J. P., and Laloux, L., Financial applications of random matrix theory: Old laces and new pieces. *Acta Phys. Polon. B* **36** (2005), 2767–2784.

[78] Ramírez, J., Rider, B., and Virág, B., Beta ensembles, stochastic Airy spectrum, and a diffusion. 2006.

[79] Reimann, P., Van den Broeck, C., and Bex, G. J., A Gaussian scenario for unsupervised learning. *J. Phys. A* **29** (13) (1996), 3521–3535.

[80] Roweis, S. T., and Saul, L. K., Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290** (5500) (2000), 2323–2326.

[81] Roy, S. N., p-statistics or some generalizations in analysis of variance appropriate to multivariate problems. *Sankhyā* **4** (1939), 381–396.

[82] Soshnikov, A., A note on universality of the distribution of the largest eigenvalues in certain classes of sample covariance matrices. *J. Statist. Phys.* **108** (2002), 1033–1056.

[83] Soshnikov, A., Poisson statistics for the largest eigenvalues in random matrix ensembles. In *Mathematical physics of quantum mechanics*, Lecture Notes in Phys. 690, Springer-Verlag, Berlin 2006, 351–364.

[84] Soshnikov, A., and Fyodorov, Y. V., On the largest singular values of random matrices with independent Cauchy entries. *J. Math. Phys.* **46** (3) (2005), 033302.

[85] Stein, C., Estimation of a covariance matrix. Unpublished manuscript, Stanford University, ca. 1977.

[86] Szegö, G., *Orthogonal Polynomials.* 3rd ed., Amer. Math. Soc. Colloq. Publ. 23, Amer. Math. Soc., Providence, R.I., 1967.

[87] Tenenbaum, J. B., de Silva, V., and Langford, J. C., A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290** (5500) (2000), 2319–2323.

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

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

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

[91] Tulino, A., and Verdu, S., *Random Matrix Theory and Wireless Communications*. Now Publishers Inc., 2004.

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

[93] Wachter, K. W., The limiting empirical measure of multiple discriminant ratios. *Ann. Statist.* **8** (1980), 937–957.

[94] Watkin, T. L. H., and Nadal, J.-P., Optimal unsupervised learning. *J. Phys. A* **27** (6) (1994), 1899–1915.

[95] Wigner, E. P., Characteristic vectors of bordered matrices of infinite dimensions. *Ann. of Math.* **62** (1955), 548–564.

[96] Wigner, E. P., On the distribution of the roots of certain symmetric matrices. *Ann. of Math.* **67** (1958), 325–328.

[97] Wishart, J., The generalised product moment distribution in samples from a normal multivariate population. *Biometrika* **20A** (1–2) (1928), 32–52.

[98] Yang, R., and Berger, J. O., Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22** (1994), 1195–1211.

Department of Statistics, Sequoia Hall, Stanford University, Stanford CA 94305, U.S.A.
E-mail: imj@stanford.edu