High Dimensional Data Matrices and Random Matrices

Craig A. Tracy
UC Davis

YITP @ 40

May 3, 2007

Data Matrices and Principal Component Analysis (PCA)

Suppose in some measurement (experiment) we have p variables

$$X_1, X_2, \ldots, X_p$$

Theoretically, the X_k are random variables. The observed data on X_k can be viewed as a vector $x_k \in \mathbb{R}^n$. We form a $p \times n$ data matrix

$$X = \begin{pmatrix} \longleftarrow & x_1 & \longrightarrow \\ & \vdots & \\ \longleftarrow & x_p & \longrightarrow \end{pmatrix}$$

- Astrophysics example: "In Orion A we have mapped 32 molecular transitions in the 3 mm wavelength band . . . ." Each "map" contains 360 pixels. Here $p = 32$ and $n = 360$.

- Human Gene Structure: Properties of $p = 38$ genes at $n = 400$ locations in Europe.
- Finance: Twenty years of stock returns in the S&P 500. Here $p = 500$ and n is the number of data points on an individual stock.

The idea of PCA (Hotelling, 1933; see Johnstone [6])

- Reduce dimensionality: $W = \sum_k v_k X_k$ by requiring

$$\mathrm{var}(W) = \sum_{k,k'} v_k \mathrm{cov}(X_k, X_{k'}) v_{k'}$$

have *maximum variance*. Vector v is the 1st principal component vector. Then choose successive linear combinations that are orthogonal to previously chosen and maximize variance.

$$\ell_j = \max \left\{ v^T \Sigma v : v^T v_{j'} = 0, j' < j, |v| = 1 \right\}$$

Sample Covariance Matrix

Use data matrix X to get $p \times p$ sample covariance matrix

$$S = \frac{1}{n} X^T X$$

and look for *sample principle components*:

$$S\hat{v}_j = \hat{\ell}_j \hat{v}_j$$

Spreading of sample eigenvalues

Take $n = p = 10$ for Wishart distribution: $X_1, \ldots, X_p$ follow a p-variate Gaussian distribution with $\Sigma = 1$:

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

On the basis of this data, might (erroneously!) conclude population eigenvalues are quite different from each other (they all equal 1).

This spreading of the eigenvalues is the statistics version of

Wigner semicircle

or as its called here

Marčenko-Pastur limit density

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

where $p/n \to \gamma$ as $n, p \to \infty$. When $n = p$ the density is supported on the interval $[0, 4]$.

Question: Suppose one sees a largest sample eigenvalue of 4.25. Is this consistent with an identity covariance matrix? (It lies outside the M-P support.)

- $H_0 : \Sigma = I$. The *null hypothesis*.
- $H_A : \Sigma \neq I$. The *alternative hypothesis*.

Want

$$\mathbb{P}\left(\hat{\ell}_1 > t | H_0 = W_p(n, I)\right)$$

Theorem (Johnstone [5]):

$$\mathbb{P}\left(\hat{\ell}_1 \leq \mu_{np} + \sigma_{np} s | H_0 = W_p(n, I)\right) \longrightarrow F_1(s)$$

where

$$\begin{aligned} \mu_{np} &= \left(\sqrt{n - 1/2} + \sqrt{p - 1/2}\right)^2 \\ \sigma_{np} &= \left(\sqrt{n - 1/2} + \sqrt{p - 1/2}\right)\left(\frac{1}{\sqrt{n - 1/2}} + \frac{1}{\sqrt{p - 1/2}}\right)^{1/3} \end{aligned}$$

What is F_1 ?

F_1 is one of three distributions first discovered by **Harold Widom** and **C.T.** [13, 14] in the context of the distribution of the largest eigenvalue in

GOE, GUE, and GSE

$$
\begin{aligned}
F_2(s) &= \exp\left(-\int_s^\infty (x-s)q(x)^2\,dx\right) \\
F_1(s)^2 &= F_2(s)\,\exp\left(-\int_s^\infty q(x)\,dx\right) \\
q'' &= sq + 2q^3,\ \text{Painlevé II equation} \\
q(s) &\sim \mathrm{Ai}(s)\ \text{as}\ s \to \infty
\end{aligned}
$$

Appearance of F_1 also in **double Wishart** : Two independent Wishart matrices $A \sim W_p(n_1, I)$ and $B \sim W_p(n_2, I)$. Appears in

- Canonical correlation analysis

William Chen, of the IRS, computed tables of the exact distribution in the double Wishart

$$m_c = (n_1 - p - 1)/2, \;\; n_c = (n_2 - p - 1)/2$$

Johnstone compared this with the TW approximation which is a limit theorem

$$n, p \to \infty, \quad \frac{p}{n} \to \gamma < \infty$$

The agreement is good. (See next slide.)

In the earlier example, the TW approximation yields a 6% chance of seeing a value more extreme that 4.25 even if "no structure" is present.

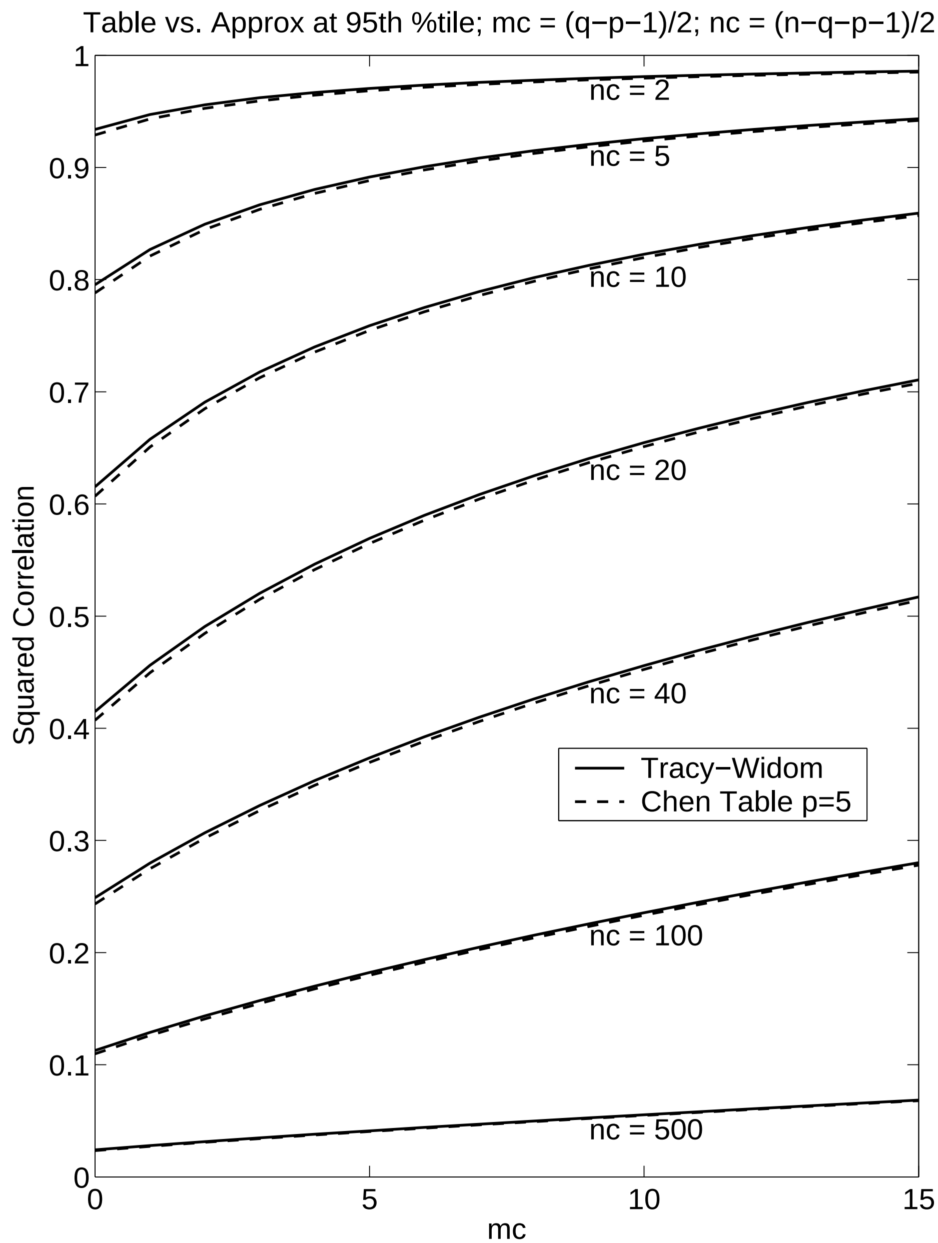
Table vs. Approx at 95th %tile; mc = (q−p−1)/2; nc = (n−q−p−1)/2
Squared Correlation
mc
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
nc = 2
nc = 5
nc = 10
nc = 20
nc = 40
nc = 100
nc = 500
Tracy−Widom
Chen Table p=5

Further Developments for Wishart Distribution

- Johnstone [5], El Karoui [7, 8], Choup [3]: **Second-order accuracy**

$$\left| P\left(n\hat{\ell}_1 \leq \mu_{np} + \sigma_{np} s | H_0\right) - F_\beta(s)\right| \leq Cp^{-2/3}$$

- El Karoui in null case for the largest eigenvalue, proves the limit law for $0 \leq \gamma \leq \infty$. This requires additional estimates to allow $\gamma = \infty$.

- Soshnikov [12] **removes Gaussian assumption** on the distribution of the matrix elements of X and only requires odd moments are zero and even moments satisfy a Gaussian type bound. Then for $\Sigma = I$ with under the restriction that as $n, p \to \infty$ that

$$n - p = \mathrm{O}(p^{1/3})$$

we get the same limit law described by F_1.

- **Note Added:** Péché [11] has removed this restriction on $\gamma := \lim p/n$.

- **Beyond the Null Hypothesis**: If $A \sim W_p(n, \Sigma)$, then joint eigenvalue density

$$c_{p,n,\Sigma} \prod_{j=1}^{p} l_j^{(n-p-1)/2} \prod_{j<k} |l_j - l_k| \times$$
$$\int_{\mathcal{O}(p)} \mathrm{e}^{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1} Q L Q^T)}\, dQ,$$

where $L = \mathrm{diag}(l_1, \ldots, l_p)$ and dQ is normalized Haar measure. Difficulty is the integral

$$\int_{\mathcal{O}(p)} \mathrm{e}^{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1} Q L Q^T)}\, dQ$$

For **complex data** above integral is replaced by

$$\int_{\mathcal{U}(p)} \mathrm{e}^{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1} U L U^*)}\, dU$$

This integral can be evaluated in terms of determinants: Harish Chandra–Itzykson–Zuber integral.

- **Spiked population data:**

$$\text{Eigenvalues of } \Sigma : \lambda_1 > 1 = \lambda_2 = \cdots = \lambda_p$$

When can we detect λ_1 from the data? It depends! Baik, Ben Arous, Péché [1, 2] describe a phase transition for population covariance matrices of above form (special case of their theorem). For real data, see conjectures of Patterson, Price and Reich [10].

C.N. Yang Institute for Theoretical Physics

Happy 40th Birthday!

References

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

[2] J. Baik, Painlevé formulas of the limiting distributions for nonnull complex sample covariance matrices, *Duke Math. J.* **133** (2006), 205–235.

[3] L. Choup, Edgeworth expansion of the largest eigenvalue distribution function of GUE and LUE, *Int. Math. Res. Not.* (2006), Art. ID 61049, 32 pp.

[4] M. Dieng and C. A. Tracy, Application of random matrix theory to multivariate statistics, preprint, arXiv:math.PR/0603543.

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

[6] I. M. Johnstone, High dimensional statistical inference and

random matrices, arXiv: math.ST/0611589.

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

[8] N. El Karoui, Tracy-Widom limit for the largest eigenvalue of a large class of complex Wishart matrices, arXiv: math.PR/0503109.

[9] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, Inc., 1982.

[10] N. Patterson, A. L. Price and D. Reich, Population structure and eigenanalysis, PLoS Genetics **2** (2006), 2074–2093.

[11] S. Péché, Universality at the soft edge for some white sample covariance matrices ensembles, preprint, arXiv:0705.1701.

[12] A. Soshnikov, A note on universality of the distribution of the largest eigenvalue in certain sample covariance matrices, *J. Statistical Physics* **108** (2002), 1033–1056.

[13] C. A. Tracy and H. Widom, Level-spacing distributions and the Airy kernel, *Phys. Letts. B* **305** (1993), 115–118; *Commun. Math. Phys.* **159** (1994), 151–174.

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