Hypothesis Testing and Random Matrix Theory

Tung Dang

Laboratory of Biometrics and Bioinformatics,
University of Tokyo

Corresponding author:
dangthanhtung91@vn-bml.com

Outline

- **Foundational ideas**
  - The distribution of principal components
  - Universality of distribution of the Largest Eigenvalues
  - Phase transition of the largest eigenvalue
- **Application to real-world problems**
  - Non-Parametric hypothesis tests
  - The Generalized Likelihood Ratio Test (GLRT)
  - Population genetics
  - Genetic Association

On the distribution of the largest eigenvalue in principal components analysis

- **Basic notation and phenomena**
- The eigenvalue–eigenvector decomposition of the sample covariance matrix

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

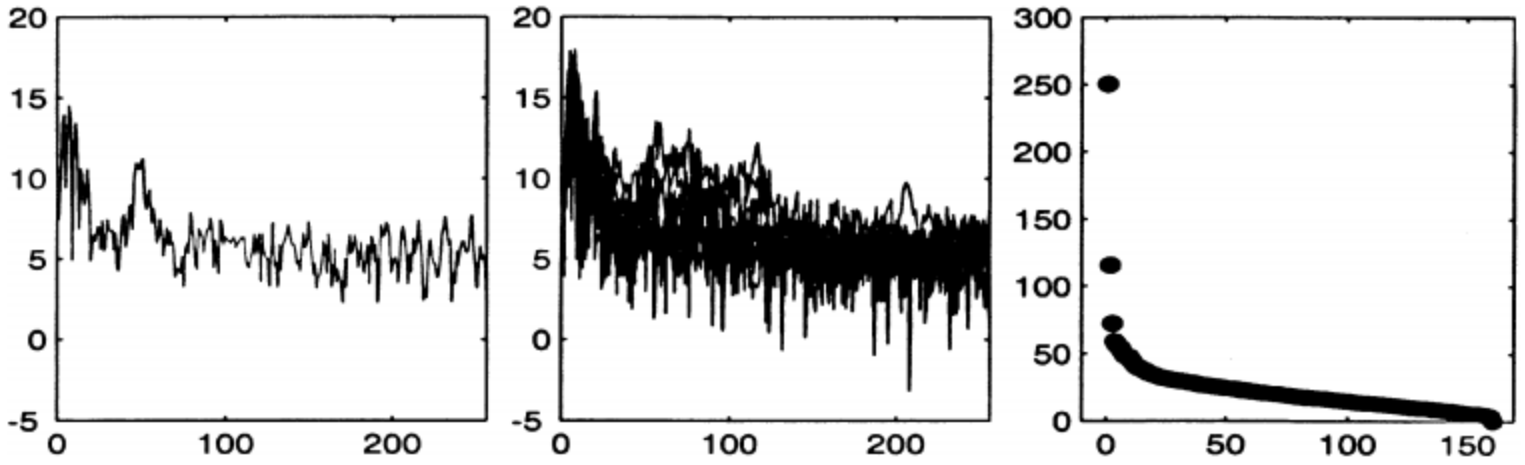


FIG. 1. (a) *a single instance of a periodogram from the phoneme dataset*; (b) *ten instances, to indicate variability*; (c) *screeplot of eigenvalues in phoneme example.*

Iain M. Johnstone. (2001). The Annals of Statistics, Vol. 29, No. 2, 295–327

On the distribution of the largest eigenvalue in principal components analysis

- **Eigenvalue distributions**
- **Bulk spectrum:**
  - which refers to the properties of the full set $l_1 > l_2 \cdots > l_p$
  - Suppose that both n and p tend to ∞, in some ratio n/p → γ ≥ 1.
  - The Marcenko–Pastur result is the empirical distribution of the eigenvalues converges almost surely,

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

and the limiting distribution has a density

$$g(t) = \frac{\gamma}{2\pi t}\sqrt{(b-t)(t-a)}, \quad a \le t \le b,$$

$$a = (1-\gamma^{-1/2})^2 \qquad b = (1+\gamma^{-1/2})^2.$$

Iain M. Johnstone. (2001). The Annals of Statistics, Vol. 29, No. 2, 295–327

On the distribution of the largest eigenvalue in principal components analysis

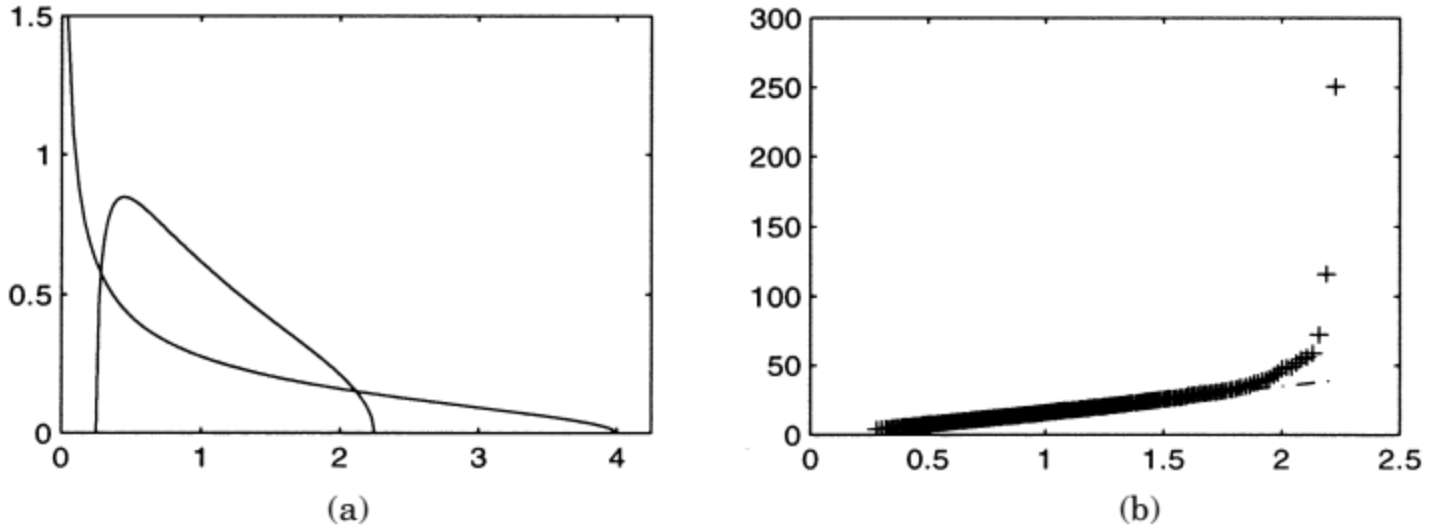


FIG. 2. *Panel* (a) *limiting densities* (1.2) *corresponding to* $n = 4p, \gamma = 4$ *and* $n = p, \gamma = 1$ (*monotone line*); (b) *Wachter plot of the empirical singular values of the phoneme data* (*vertical axis*) *versus quantiles*.

- The smaller n/p, the more spread the eigenvalues; even asymptotically, the spread of the empirical eigenvalues does not disappear.
- For n = p, the largest normalized eigenvalue approaches 4 and the smallest approaches 0

Iain M. Johnstone. (2001). The Annals of Statistics, Vol. 29, No. 2, 295–327

On the distribution of the largest eigenvalue in principal components analysis

- **Eigenvalue distributions**
- **Extremes:**
  - which addresses the (first few) largest and smallest eigenvalues.
  - $X = (X_{jk})_{n \times p}$ has entries which are i.i.d $X_{jk} \sim N(0,1)$. The sample eigenvalues of the Wishart matrix by $l_1 > l_2 \cdots > l_p$

Center and scaling constants

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

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

***Tracy–Widom* law** of order 1

$$F_1(s) = \exp\left\{ -\tfrac{1}{2} \int_s^\infty q(x) + (x-s) q^2(x)\, dx \right\}$$

The (nonlinear) Painleve II differential equation

$$q''(x) = x q(x) + 2q^3(x),$$

$$q(x) \sim \mathrm{Ai}(x) \quad \text{as } x \to +\infty$$

  - This distribution was the limiting law of the largest eigenvalue of an n by n Gaussian symmetric matrix.

Iain M. Johnstone. (2001). The Annals of Statistics, Vol. 29, No. 2, 295–327

On the distribution of the largest eigenvalue in principal components analysis

- **Main result**
- 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* $n/p \to \gamma \geq 1$,

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

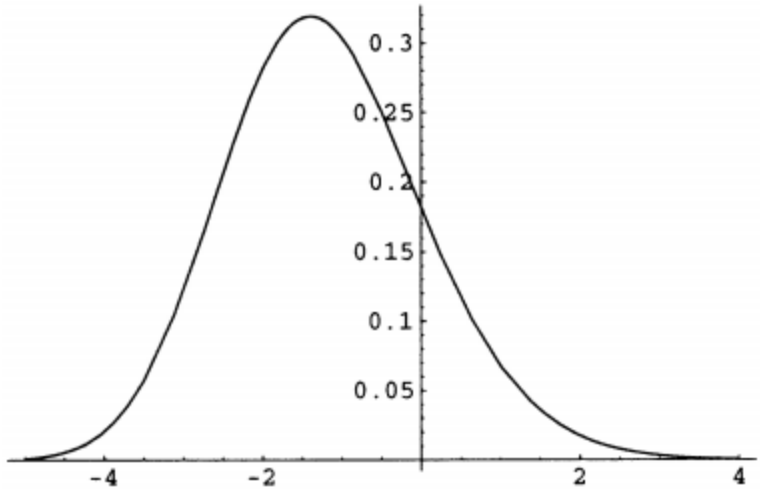


FIG. 3. *Density of the Tracy–Widom distribution* F_1.

Iain M. Johnstone. (2001). The Annals of Statistics, Vol. 29, No. 2, 295–327

On the distribution of the largest eigenvalue in principal components analysis

- **The practical applicability of Theorem 1.1**

TABLE 1

Simulations for finite $n \times p$ versus Tracy–Widom Limit. The first column shows the probabilities of the F_1 limit distribution corresponding to fractions in second column. The next three columns show estimated cumulative probabilities for l_1, centered and scaled as in (1.3) and (1.4), 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. The final column gives approximate standard errors based on binomial sampling. The bold font highlights some conventional significance levels. The Tracy–Widom distribution F_1 was evaluated on a grid of 121 points $-6(0.1)6$ using the `Mathematica` *package* `p2Num` *written by Craig Tracy. Remaining computations were done in* `MATLAB`*, with percentiles obtained by inverse interpolation and using* `randn()` *for normal variates and* `norm()` *to evaluate largest singular values*

Percentile	TW	5×5	10×10	100×100	5×20	10×40	100×400	2 * SE
−3.90	0.01	0.000	0.001	0.007	0.002	0.003	0.010	(0.002)
−3.18	0.05	0.003	0.015	0.042	0.029	0.039	0.049	(0.004)
−2.78	0.10	0.019	0.049	0.089	0.075	0.089	0.102	(0.006)
−1.91	0.30	0.211	0.251	0.299	0.304	0.307	0.303	(0.009)
−1.27	0.50	0.458	0.480	0.500	0.539	0.524	0.508	(0.010)
−0.59	0.70	0.697	0.707	0.703	0.739	0.733	0.714	(0.009)
0.45	**0.90**	**0.901**	**0.907**	**0.903**	**0.919**	**0.918**	**0.908**	(0.006)
0.98	**0.95**	**0.948**	**0.954**	**0.950**	**0.960**	**0.961**	**0.957**	(0.004)
2.02	**0.99**	**0.988**	**0.991**	**0.991**	**0.992**	**0.993**	**0.992**	(0.002)

Iain M. Johnstone. (2001). The Annals of Statistics, Vol. 29, No. 2, 295–327

On the distribution of the largest eigenvalue in principal components analysis

- **The practical applicability of Theorem 1.1**

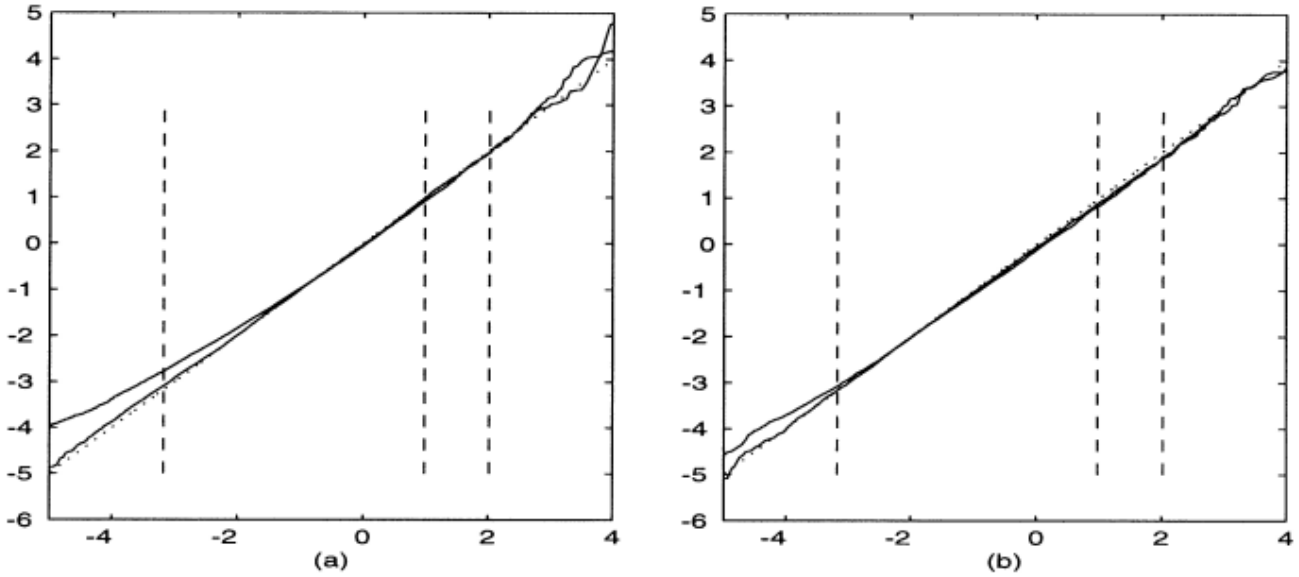


FIG. 4. *Panel* (a): *Probability plots of* $R = 10{,}000$ *observed replications of* l_1 *drawn from* $W_p(n, I)$ *for* $n = p = 10$ *and* 100. *That is, the* $10{,}000$ *ordered observed values of* l_1 *are plotted against* $F_1^{-1}((i - 0.5)/R)$, $i = 1, \ldots, R$. *The line for* $n = p = 10$ *is the one elevated in the left tail. The vertical dashed lines show* 5*th*, 95*th and* 99*th percentiles. The dotted line is the* 45 *degree line of perfect agreement of empirical law with asymptotic limit. Panel* (b): *Same plots for* $n = 40$, $p = 10$ *and* $n = 400$, $p = 100$.

Iain M. Johnstone. (2001). The Annals of Statistics, Vol. 29, No. 2, 295–327

On the distribution of the largest eigenvalue in principal components analysis

- **Nonnull cases: empirical results.**
- 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 a "spiked" covariance model, with a fixed number r, eigenvalues greater than 1

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

- $\mathscr{L}(l_k | n, p, \Sigma_\tau)$ is the distribution of the kth largest sample eigenvalue of the sample covariance matrix X'X where the n by p matrix X is derived from n independent draws from $N_p(0, \Sigma_\tau)$

PROPOSITION 1.2. *Assume* $r < p$. *Then* $\mathscr{L}(l_{r+1} | n, p, \Sigma_\tau) \overset{\text{st}}{<} \mathscr{L}(l_1 | n, p - r, I_{p-r})$.

- $\mathscr{L}(l_1 | n, p - r, I_{p-r})$ provides a conservative p-value for testing H_0: $\tau_{r+1}^2 = 1$

Iain M. Johnstone. (2001). The Annals of Statistics, Vol. 29, No. 2, 295–327

On the distribution of the largest eigenvalue in principal components analysis

- **Empirical evidence**

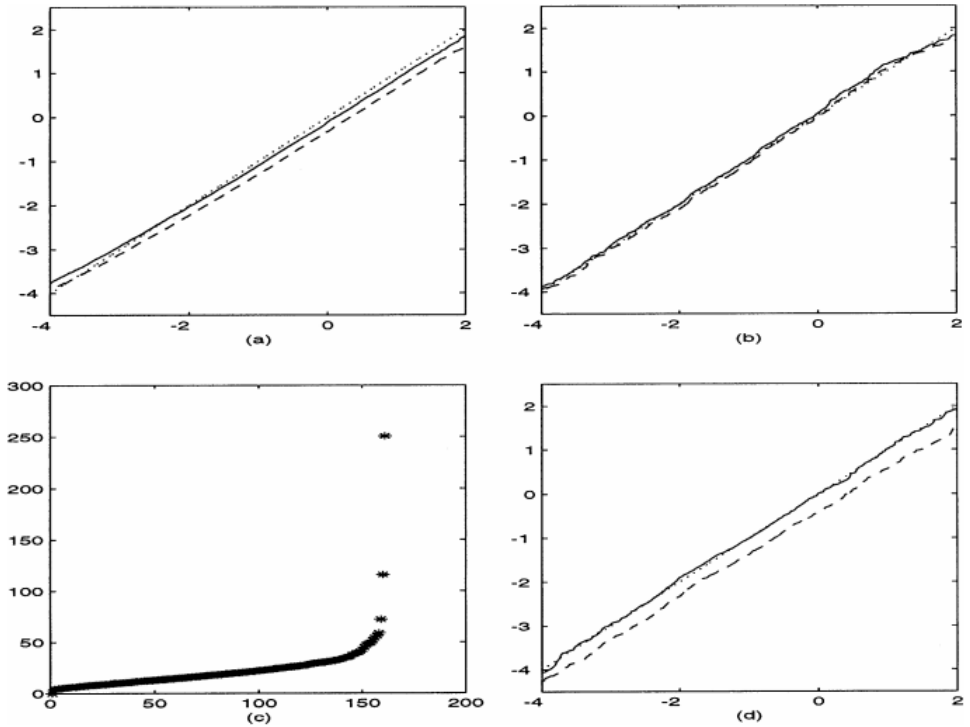


FIG. 5. (a) 10 *unit roots and one with* $\tau = 10$ *in model* (1.6). $p = 10, n = 40$, *with* $N = 10{,}000$ *replications. The dashed line is a qq plot of the second largest value against the TW distribution. For comparison, the solid line is the simulated null distribution for* 10×40 *white Wishart 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. The 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) *The dashed line is qq plot of fourth largest eigenvalue from spiked covariance model with top three values set at the observed values in the phoneme data. Solid line is qq plot of largest eigenvalue from a null Wishart with* $p = 158$ *and* $n = 256$. $N = 1{,}000$ *replications.*

Iain M. Johnstone. (2001). The Annals of Statistics, Vol. 29, No. 2, 295–327

Universality of the Distribution of the Largest Eigenvalues in Certain Sample Covariance Matrices

- **Real Sample Covariance Matrices**
- The ensemble consists of p-dimensional random matrices $A_p=X'X$, where X is an $n\times p$ matrix with independent real random entries x_{ij}.
- **Assumptions:**
  - i. With $\mathbf{E}x_{ij}=0,\ \mathbf{E}(x_{ij})^2=1,\ \ 1\leqslant i\leqslant n,\qquad 1\leqslant j\leqslant p.$
  - ii. The random variables x_{ij} have symmetric laws of distribution.
  - iii. All moments of these random variables are finite; in particular (ii) implies that all odd moments vanish.
  - iv. The distributions of x_{ij}, decay at infinity at least as fast as a Gaussian distribution, namely

$$\mathbf{E}(x_{ij})^{2m}\leqslant(\text{const}\ m)^m.$$

Alexander Soshnikov. (2002). Journal of Statistical Physics, Vol. 108, Nos. 5/6

Universality of the Distribution of the Largest Eigenvalues in Certain Sample Covariance Matrices

- **Complex Sample Covariance Matrices**
- The ensemble consists of *p*-dimensional random matrices *Ap=X*X* (*X** denotes a complex conjugate matrix), where *X* is an *n×p* matrix with independent complex random entries *xij*
- **Assumptions**

(i′) With $\mathbf{E}x_{ij} = 0,\ \mathbf{E}(x_{ij})^2 = 0,\ \mathbf{E}\,|x_{ij}|^2 = 1,\ 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p$

(ii′) The random variables Re x_{ij}, Im x_{ij} have symmetric laws of distribution.

(iii′) All moments of these random variables are finite; in particular (ii′) implies that all odd moments vanish.

(iv′) The distributions of Re x_{ij}, Im x_{ij} decay at infinity at least as fast as a Gaussian distribution, namely

$$\mathbf{E}\,|x_{ij}|^{2m} \leqslant (\text{const}\ m)^m.$$

Alexander Soshnikov. (2002). Journal of Statistical Physics, Vol. 108, Nos. 5/6

Universality of the Distribution of the Largest Eigenvalues in Certain Sample Covariance Matrices

Theorem. Suppose that a matrix $A_p = X^t X$ ($A_p = X^* X$) has a real (complex) Wishart distribution (defined in Remark 1 above) and $n/p \to \gamma > 0$. Then

$$\frac{\lambda_{\max}(A_p) - \mu_{n,p}}{\sigma_{n,p}}$$

where

$$\mu_{n,p} = (n^{1/2} + p^{1/2})^2, \tag{1.12}$$

$$\sigma_{n,p} = (n^{1/2} + p^{1/2})(n^{-1/2} + p^{-1/2})^{1/3} \tag{1.13}$$

converges in distribution to the Tracy–Widom law (F_1 in the real case, F_2 in the complex case).

$$F_1(x) = \exp\left\{ -\tfrac{1}{2} \int_x^\infty q(t) + (x-t)\, q^2(t)\, dt \right\},$$

$$F_2(x) = \exp\left\{ -\int_x^\infty (x-t)\, q^2(t)\, dt \right\},$$

$$d^2q(x)/dx^2 = xq(x) + 2q^3(x)$$

$$q(x) \sim \mathrm{Ai}(x) \qquad \text{as} \quad x \to +\infty$$

Alexander Soshnikov. (2002). Journal of Statistical Physics, Vol. 108, Nos. 5/6

Universality of the Distribution of the Largest Eigenvalues in Certain Sample Covariance Matrices

Theorem 1. The joint distribution of the first, second, third, etc. largest eigevalues (rescaled as in (1.12), (1.13)) of a real (complex) Wishart matrix converges to the distribution given by the Tracy–Widom law (i.e., the limiting distribution of the first, second, etc. rescaled eigenvalues for GOE ($\beta = 1$, real case) or GUE ($\beta = 2$, complex case) correspondingly).

Theorem 2. Let a real (complex) sample covariance matrix satisfy the conditions (i−iv)((i′−iv′)) and $n - p = O(p^{1/3})$. Then the joint distribution of the first, second, third, etc. largest eigenvalues (rescaled as in (1.12), (1.13)) converge to the Tracy–Widom law with $\beta = 1(2)$.

Theorem 3. Let a real (complex) sample covariance matrix satisfy (i)–(iv) ((i′)–(iv′)) and $n/p \to \gamma > 0$. Then

$$\text{(a)} \quad \mathbf{E}\,\mathrm{Trace}\, A_p^m = \frac{(\sqrt{\gamma}+1)\,\gamma^{1/4}}{2\sqrt{\pi}} \frac{p\mu_{n,p}^m}{m^{3/2}} (1+o(1)) \qquad \text{if} \quad m = o(\sqrt{p}).$$

$$\text{(b)} \quad \mathbf{E}\,\mathrm{Trace}\, A_p^m = O\left(\frac{p\mu_{n,p}^m}{m^{3/2}}\right) \qquad \text{if} \quad m = O(\sqrt{p}).$$

Alexander Soshnikov. (2002). Journal of Statistical Physics, Vol. 108, Nos. 5/6

Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices

THEOREM 1.1. *Let λ_1 be the largest eigenvalue of the sample covariance matrix constructed from M independent, identically distributed complex Gaussian sample vectors of N variables. Let $\ell_1, \ldots, \ell_N$ denote the eigenvalues of the covariance matrix of the samples. Suppose that for a fixed integer $r \geq 0$,*

$$\ell_{r+1} = \ell_{r+2} = \cdots = \ell_N = 1. \tag{35}$$

As $M, N \to \infty$ while $M/N = \gamma^2$ is in a compact subset of $[1, \infty)$, the following hold for any real x in a compact set.

JINHO BAIK et al. (2005). The Annals of Probability, Vol. 33, No. 5, 1643–1697.

Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices

(a) *When for some* $0 \le k \le r$,

$$\ell_1 = \cdots = \ell_k = 1 + \gamma^{-1} \tag{36}$$

and $\ell_{k+1}, \ldots, \ell_r$ *are in a compact subset of* $(0, 1+\gamma^{-1})$,

$$\mathbb{P}\left((\lambda_1 - (1+\gamma^{-1})^2) \cdot \frac{\gamma}{(1+\gamma)^{4/3}} M^{2/3} \le x \right) \to F_k(x), \tag{37}$$

where $F_k(x)$ *is defined in* (17).

DEFINITION 1.1. For $k = 1, 2, \ldots$, define for real x,

$$F_k(x) = \det(1 - \mathbf{A}_x) \cdot \det\left(\delta_{mn} - \left\langle \frac{1}{1-\mathbf{A}_x} s^{(m)}, t^{(n)} \right\rangle \right)_{1 \le m,n \le k}, \tag{17}$$

where $\langle \cdot, \cdot \rangle$ denotes the (real) inner product of functions in $L^2((x, \infty))$. Let $F_0(x) = \det(1 - \mathbf{A}_x)$.

JINHO BAIK et al. (2005). The Annals of Probability, Vol. 33, No. 5, 1643–1697.

Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices

$$F_0(x) = \det(1 - \mathbf{A}_x) = F_{\mathrm{GUE}}(x) \qquad F_1(x) = \det(1 - \mathbf{A}_x) \cdot \left(1 - \left\langle \frac{1}{1 - \mathbf{A}_x} s^{(1)}, t^{(1)} \right\rangle\right) = (F_{\mathrm{GOE}}(x))^2.$$

$$u(x) \sim -\mathrm{Ai}(x), \qquad x \to +\infty.$$

$$u(x) = -\frac{e^{-(2/3)x^{3/2}}}{2\sqrt{\pi}x^{1/4}} + O\left(\frac{e^{-(4/3)x^{3/2}}}{x^{1/4}}\right) \qquad \text{as } x \to +\infty,$$

$$u(x) = -\sqrt{\frac{-x}{2}}(1 + O(x^{-2})) \qquad \text{as } x \to -\infty.$$

$$F_0(x) = \det(1 - \mathbf{A}_x^{(0)}) = \exp\left(-\int_x^\infty (y - x)u^2(y)\,dy\right). \qquad F_1(x) = F_0(x) \exp\left(\int_x^\infty u(y)\,dy\right).$$

JINHO BAIK et al. (2005). The Annals of Probability, Vol. 33, No. 5, 1643–1697.

Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices

(b) *When for some* $1 \le k \le r$,

(38) $$\ell_1 = \cdots = \ell_k \text{ is in a compact set of } (1+\gamma^{-1}, \infty)$$

and $\ell_{k+1}, \ldots, \ell_r$ *are in a compact subset of* $(0, \ell_1)$,

(39) $$\mathbb{P}\left(\left(\lambda_1 - \left(\ell_1 + \frac{\ell_1\gamma^{-2}}{\ell_1 - 1}\right)\right) \cdot \frac{\sqrt{M}}{\sqrt{\ell_1^2 - \ell_1^2\gamma^{-2}/(\ell_1-1)^2}} \le x\right) \to G_k(x),$$

where $G_k(x)$ *is defined in* (28).

DEFINITION 1.2. For $k = 1, 2, 3, \ldots$, define the distribution $G_k(x)$ by

(28) $$G_k(x) = \frac{1}{Z_k} \int_{-\infty}^{x} \cdots \int_{-\infty}^{x} \prod_{1 \le i < j \le k} |\xi_i - \xi_j|^2 \cdot \prod_{i=1}^{k} e^{-(1/2)\xi_i^2}\, d\xi_1 \cdots d\xi_k.$$

In other words, G_k is the distribution of the *largest eigenvalue* of $k \times k$ GUE. When $k = 1$, this is the Gaussian distribution,

(29) $$G_1(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(1/2)\xi_1^2}\, d\xi_1 = erf(x).$$

JINHO BAIK et al. (2005). The Annals of Probability, Vol. 33, No. 5, 1643–1697.

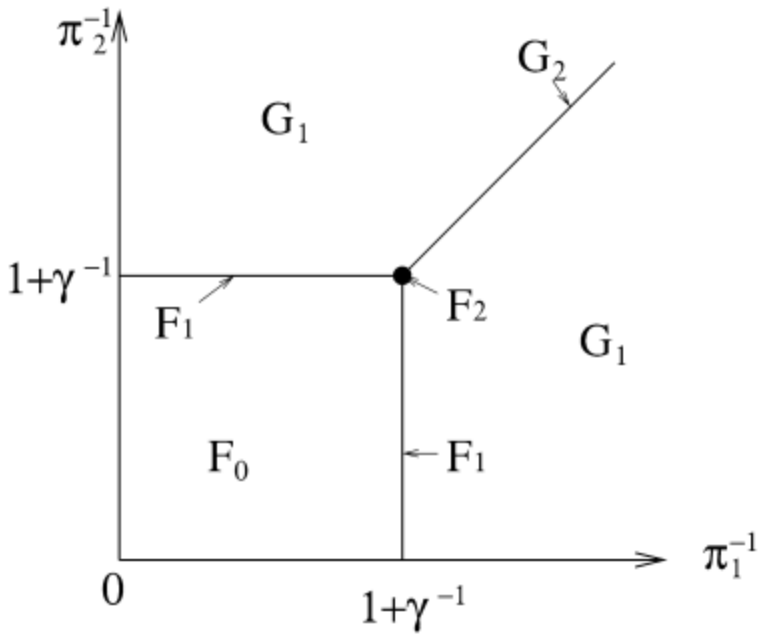
Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices

$$\mathbb{P}\left((\lambda_1 - (1+\gamma^{-1})^2) \cdot \frac{\gamma}{(1+\gamma)^{4/3}} M^{2/3} \le x \right)$$

$$\to \begin{cases} F_0(x), & 0 < \ell_1, \ell_2 < 1+\gamma^{-1}, \\ F_1(x), & 0 < \ell_2 < 1+\gamma^{-1} = \ell_1, \\ F_2(x), & \ell_1 = \ell_2 = 1+\gamma^{-1}, \end{cases}$$

$$\mathbb{P}\left(\left(\lambda_1 - \left(\ell_1 + \frac{\ell_1 \gamma^{-2}}{\ell_1 - 1} \right) \right) \cdot \frac{\sqrt{M}}{\sqrt{\ell_1^2 - \ell_1^2 \gamma^{-2}/(\ell_1 - 1)^2}} \le x \right)$$

$$\to \begin{cases} G_1(x), & \ell_1 > 1+\gamma^{-1}, \ell_1 > \ell_2, \\ G_2(x), & \ell_1 = \ell_2 > 1+\gamma^{-1}, \end{cases}$$



JINHO BAIK et al. (2005). The Annals of Probability, Vol. 33, No. 5, 1643–1697.

Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices

COROLLARY 1.1. *Under the same assumption of Theorem* 1.1, *the following hold.*

(a) *When for some* $0 \le k \le r$,

$$\ell_1 = \cdots = \ell_k = 1 + \gamma^{-1}, \tag{43}$$

and $\ell_{k+1}, \ldots, \ell_r$ *are in a compact subset of* $(0, 1+\gamma^{-1})$,

$$\lambda_1 \to (1+\gamma^{-1})^2 \quad \text{in probability.} \tag{44}$$

(b) *When for some* $1 \le k \le r$,

$$\ell_1 = \cdots = \ell_k > 1 + \gamma^{-1} \tag{45}$$

and $\ell_{k+1}, \ldots, \ell_r$ *are in a compact subset of* $(0, \ell_1)$,

$$\lambda_1 \to \ell_1\left(1 + \frac{\gamma^{-2}}{\ell_1 - 1}\right) \quad \text{in probability.} \tag{46}$$

JINHO BAIK et al. (2005). The Annals of Probability, Vol. 33, No. 5, 1643–1697.

Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices

- **Around the transition point; interpolating distributions**

DEFINITION 1.3. For $k = 1, 2, \ldots$, define for real x and $w_1, \ldots, w_k$,

$$F_k(x; w_1, \ldots, w_k)$$

$$= \det(1 - \mathbf{A}_x) \cdot \det\left(1 - \left\langle \frac{1}{1 - \mathbf{A}_x} s^{(m)}(w_1, \ldots, w_m), t^{(n)}(w_1, \ldots, w_{n-1}) \right\rangle\right)_{1 \le m,n \le k}. \tag{54}$$

$$s^{(m)}(u; w_1, \ldots, w_m) = \frac{1}{2\pi} \int e^{iua + i(1/3)a^3} \prod_{j=1}^{m} \frac{1}{w_j + ia}\, da,$$

$$t^{(m)}(v; w_1, \ldots, w_{m-1}) = \frac{1}{2\pi} \int e^{ivb + i(1/3)b^3} \prod_{j=1}^{m-1} (w_j - ib)\, db,$$

JINHO BAIK et al. (2005). The Annals of Probability, Vol. 33, No. 5, 1643–1697.

Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices

- **Around the transition point; interpolating distributions**

THEOREM 1.2. *Suppose that for a fixed r, $\ell_{r+1} = \ell_{r+2} = \cdots = \ell_N = 1$. Set for some $1 \le k \le r$,*

$$\ell_j = 1 + \gamma^{-1} - \frac{(1+\gamma)^{2/3} w_j}{\gamma M^{1/3}}, \qquad j = 1, 2, \ldots, k. \tag{55}$$

When w_j, $1 \le j \le k$, is in a compact subset of $\mathbb{R}$, and ℓ_j, $k+1 \le j \le r$, is in a compact subset of $(0, 1+\gamma^{-1})$, as $M, N \to \infty$ such that $M/N = \gamma^2$ is in a compact subset of $[1, \infty)$,

$$\mathbb{P}\left((\lambda_1 - (1+\gamma^{-1})^2) \cdot \frac{\gamma}{(1+\gamma)^{4/3}} M^{2/3} \le x \right) \to F_k(x; w_1, \ldots, w_k) \tag{56}$$

for any x in a compact subset of $\mathbb{R}$.

JINHO BAIK et al. (2005). The Annals of Probability, Vol. 33, No. 5, 1643–1697.

Non-Parametric Detection of the Number of Signals

the steering matrix p x K of linearly independent dimensional vectors.

The unknow noise level

a px1 additive Gaussian noise vector, $\mathrm{N}\left(0,\mathrm{I}_p\right)$ and independent of s(t).

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \sigma\mathbf{n}(t)$$

$\{\mathbf{x}_i = \mathbf{x}(t_i)\}_{i=1}^{n}$ denote p dimensional i.i.d. observations sampled at distinct times t_i

$\mathbf{s}(t) = [s_1(t), \ldots, s_K(t)]^T$ represents the random signals, assumed zero mean and stationary with full rank covariance matrix.

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Non-Parametric Detection of the Number of Signals

- The population covariance matrix of x(t) has a diagonal form,

$$\mathbf{W}^H \boldsymbol{\Sigma} \mathbf{W} = \sigma^2 \mathbf{I}_p + \mathrm{diag}(\lambda_1, \ldots, \lambda_K, 0, \ldots, 0) \qquad \text{where } \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_K > 0.$$

- Sn the sample covariance matrix of the n observations from xi

$$\mathbf{S}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^H \qquad \ell_1 \geqslant \ell_2 \geqslant \ldots \geqslant \ell_p \text{ its eigenvalues.}$$

- **Problem:**
  - Estimate the unknown number of sources given the observations under the *nonparametric* setting, where no prior information is assumed about the matrix A beyond it being of rank K.
  - Consider methods to infer the number of signals that use the eigenvalues ℓ_j of the sample covariance matrix, rather than the original observations.

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Non-Parametric Detection of the Number of Signals

- The key principle in nonparametric estimation of the number of sources is that for sufficiently large n, in the presence of K sources, the first K largest sample eigenvalues correspond to signals, whereas the remaining eigenvalues correspond to noise.

Theorem 1: Let $\mathbf{S}_n$ denote the sample covariance matrix of n pure noise vectors distributed $\mathcal{N}(0, \sigma^2\mathbf{I}_p)$. In the joint limit $p, n \to \infty$, with $p/n \to c \geqslant 0$, the distribution of the largest eigenvalue of $\mathbf{S}_n$ converges to a Tracy–Widom distribution

$$\Pr\left\{\frac{\ell_1/\sigma^2 - \mu_{n,p}}{\xi_{n,p}} < s\right\} \to F_\beta(s) \qquad (3)$$

$$\mu_{n,p} = \frac{1}{n}(\sqrt{n-1/2} + \sqrt{p-1/2})^2, \qquad (4)$$

$$\xi_{n,p} = \sqrt{\frac{\mu_{n,p}}{n}}\left(\frac{1}{\sqrt{n-1/2}} + \frac{1}{\sqrt{p-1/2}}\right)^{1/3}. \qquad (5)$$

$\beta = 1$ for real valued noise
$\beta = 2$ for complex valued noise.

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Non-Parametric Detection of the Number of Signals

- **A false alarm (type I error)** with asymptotic probability α as, $p, n \to \infty$ the threshold $s(\alpha)$ should satisfy

$$F_\beta(s(\alpha)) = 1 - \alpha.$$

$$F_1(x) = 1 - \frac{e^{-2/3x^{3/2}}}{4\sqrt{\pi}x^{3/2}}(1 + O(x^{-3/2}))$$

$$F_2(x) = 1 - \frac{e^{-4/3x^{3/2}}}{16\pi x^{3/2}}(1 + O(x^{-3/2})).$$

$$s(\alpha) \approx \begin{cases} (-3/2 \log 4\sqrt{\pi}\alpha)^{2/3} & \beta = 1 \\ (-3/4 \log 16\pi\alpha)^{2/3} & \beta = 2. \end{cases}$$

There is no explicit closed form expression for the TW distribution, this inversion can be done numerically

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Non-Parametric Detection of the Number of Signals

Theorem 2: Let ℓ_1 be the largest eigenvalue as in Theorem 1, then

$$\Pr\left\{\frac{\ell_1}{\sigma^2} > \left(1+\sqrt{\frac{p}{n}}\right)^2 + \varepsilon\right\} \leqslant \exp\{-nJ_{\mathrm{LAG}}(\varepsilon)\} \qquad (8)$$

where

$$J_{\mathrm{LAG}} = \int_1^x (x-y)\frac{(1+c)y+2\sqrt{c}}{(y+B)^2}\frac{dy}{\sqrt{y^2-1}} \qquad (9)$$

with $c=(p/n), x=1+(\varepsilon)/(2\sqrt{c})$ and $B=(1+c)/(2\sqrt{c})$.

$n \to \infty, S_n \xrightarrow{a.s.} \Sigma$ and so the sample eigenvalues converge w.p.1 to the corresponding population eigenvalues.

- Large values of p and n, **signal strengths** follows that the largest eigenvalue due to noise is approximately $\sigma^2(1+\sqrt{p/n})^2$

- **A weak signal** cannot be detected by the largest sample eigenvalue, since the variance in its direction will be smaller than the largest variance in a random direction due to noise.

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Non-Parametric Detection of the Number of Signals

- **A phase transition**, where signals can be detected by the largest eigenvalues if and only if they are above a certain deterministic threshold

Theorem 3: Let $\mathbf{S}_n$ denote the sample covariance matrix of n observations from (1) with a single signal of strength λ. Then, in the joint limit $p, n \to \infty$, with $p/n \to c > 0$, the largest eigenvalue of $\mathbf{S}_n$ converges w.p.1 to

$$\lambda_{\max}(\mathbf{S}_n) \xrightarrow{\text{a.s.}} \begin{cases} \sigma^2(1+\sqrt{p/n})^2 & \lambda \leqslant \sigma^2\sqrt{p/n} \\ (\lambda+\sigma^2)\left(1+\frac{p}{n}\frac{\sigma^2}{\lambda}\right) & \lambda > \sigma^2\sqrt{p/n}. \end{cases} \quad (10)$$

- **The threshold** as the non-parametric *asymptotic limit of detection*

$$\lambda_{\text{DET}} = \sigma^2\sqrt{\frac{p}{n}}.$$

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Non-Parametric Detection of the Number of Signals

- **Likelihood Ratio Asymptotics**

$$\mathrm{LRT} = \frac{p(\ell_1,\ldots,\ell_p|\mathcal{H}_1)}{p(\ell_1,\ldots,\ell_p|\mathcal{H}_0)} \lessgtr C_\alpha$$

$$\begin{aligned} &p(\ell_1,\ldots,\ell_p|\boldsymbol{\Sigma}) \\ &\quad = C_{n,p}\prod_i \mu_i^{-n/2}\prod_i \ell_i^{(n-p-1)/2} \\ &\qquad \times \prod_{i<j}(\ell_i-\ell_j)\,{}_0F_0^p\left(-\frac{1}{2}nL, A\right) \end{aligned}$$

$$\log\frac{p(\ell_1,\ldots,\ell_p|\mathcal{H}_1)}{p(\ell_1,\ldots,\ell_p|\mathcal{H}_0)} = \frac{n}{2}\left[\ell_1\frac{\lambda}{\lambda+1} - \log(1+\lambda)\right](1+o(1))$$

Hence, asymptotically in sample size, we accept $\mathcal{H}_1$ if

$$\ell_1 > \left(1+\frac{1}{\lambda}\right)\left[\log(1+\lambda)+\frac{2}{n}\log C_\alpha\right].$$

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \mathbf{I}_p \quad \text{vs.}$$
$$\mathcal{H}_1 : \mathbf{W}^{\mathbf{H}}\boldsymbol{\Sigma}\mathbf{W} = \mathbf{I}_p + \mathrm{diag}(\lambda,0,\ldots,0)$$

Under $\mathcal{H}_0$, no signals are present ($\boldsymbol{\Sigma} = \mathbf{I}_p$), and

$${}_0F_0^p\left(-\frac{n}{2}L, \mathbf{I}_p\right) = \exp\left\{-\frac{n}{2}\sum_{j=1}^p \ell_j\right\}.$$

Under $\mathcal{H}_1$, a single signal of strength λ is present.

$$\begin{aligned} &{}_0F_0^p\left(-\frac{n}{2}L, A\right) \\ &= C_p \exp\left\{-\frac{n}{2}\left[\frac{\ell_1}{\lambda+1}+\sum_{j=2}^p \ell_j\right]\right\} \\ &\quad \times \prod_{j>1}\left[\frac{2\pi}{n}\frac{1}{\ell_1-\ell_j}\frac{1}{1-1/(\lambda+1)}\right]^{1/2} \\ &\quad \times \left[1+O\left(\frac{1}{n}\right)\right]. \end{aligned}$$

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Non-Parametric Detection of the Number of Signals

- **An RMT Based Estimation Algorithm**

The algorithm works as follows: For $k = 1, \ldots, \min(p, n) - 1$, we test

$$\mathcal{H}_0: \text{at most } k-1 \text{ signals} \quad \text{vs.} \quad \mathcal{H}_1: \text{at least } k \text{ signals.}$$

Under the null hypothesis, ℓ_k arises from noise. Thus, we reject $\mathcal{H}_0$ if ℓ_k is too large,

$$\ell_k > \hat{\sigma}^2(k) C_{n,p,k}(\alpha)$$

$$C_{n,p,k}(\alpha) = \mu_{n,p-k} + s(\alpha)\xi_{n,p-k} \qquad \hat{\sigma}^2 = \frac{1}{p-K}\sum_{K+1}^{p} \ell_j.$$

$$\hat{K}_{\mathrm{RMT}} = \arg\min_k \left\{ \ell_k < \hat{\sigma}^2(k)(\mu_{n,p-k} + s(\alpha)\xi_{n,p-k}) \right\} - 1.$$

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Non-Parametric Detection of the Number of Signals

- **Performance Analysis compare with Rao-Edelman and Schott algorithms**
- The condition for RMT algorithm to report at least the correct number of signals is

$$\ell_K > \mu_{n,p-K} + s(\alpha)\xi_{n,p-K}.$$

$$\lambda_{\mathrm{RMT}} \approx \sqrt{\frac{p-K}{n}}\left(1+\frac{\sqrt{s(\alpha)(1+\sqrt{(p-K)/n})}}{(p-K)^{1/3}}\right)$$

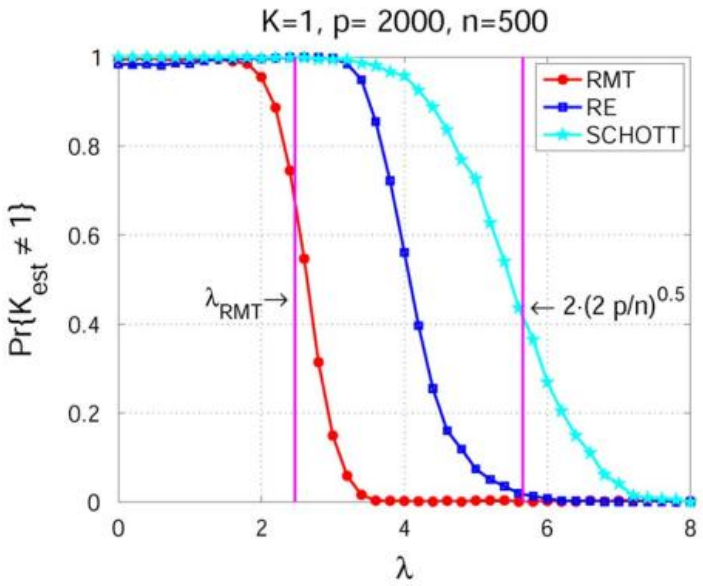


Fig. 7. Misdetection (error) probability as a function of signal strength, for $p = 2000, n = 500$ $(p > n)$.

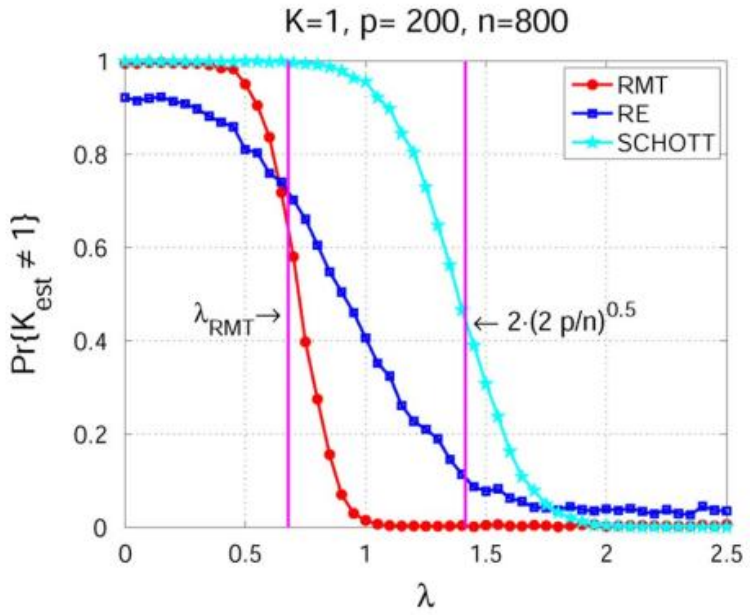


Fig. 8. Misdetection (error) probability as a function of signal strength, for $p = 200, n = 800$ $(p < n)$.

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Non-Parametric Detection of the Number of Signals

- **Simulations**

The presence of two signals (K=2) with strengths $(\lambda_1, \lambda_2) = (1, 0.4)$ with p = 10 sensors

p = 25 sensors.
- Superior detection performance of the RMT algorithm.
- AIC estimator is asymptotically inconsistent, having a non-negligible probability to overestimate the number of signals when is large.

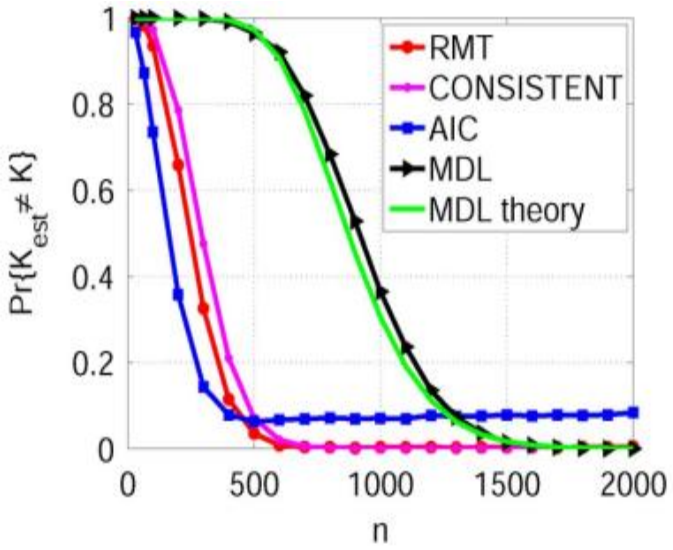


Fig. 9. Misdetection (error) probability as a function of sample size n.

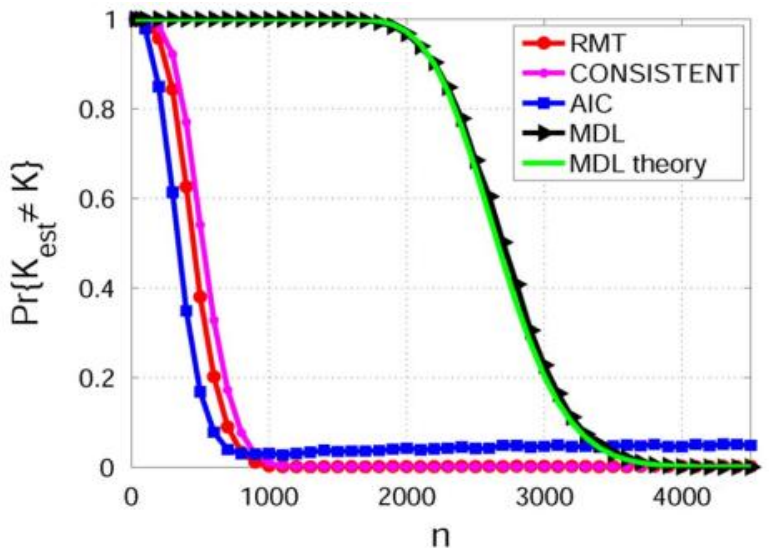


Fig. 10. Misdetection (error) probability as a function of sample size n.

Shira Kritchman and Boaz Nadler. (2009). IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10.

Statistical Tests for Single-Source Detection Using Random Matrix Theory

- The aim of the multisensor cognitive detection phase is to construct and analyze tests associated with the following hypothesis testing problem

An (i.i.d.) process of vectors with circular complex Gaussian entries with mean zero and covariance matrix $\sigma^2 \mathbf{I}_K$

$$\boldsymbol{y}(n) = \begin{cases} \boldsymbol{w}(n), & \text{under } H_0 \\ \boldsymbol{h}\, s(n) + \boldsymbol{w}(n), & \text{under } H_1 \end{cases} \quad \text{for } n = 0 : N-1 \qquad (1)$$

The observed K x 1 complex time series

A deterministic vector represents the propagation channel between the source and the K sensors.

Signal denotes a standard scalar (i.i.d.) circular complex Gaussian process with respect to the n=0: N-1 samples and stands for the source signal to be detected.

P. Bianchi et al. (2011). IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4.

Statistical Tests for Single-Source Detection Using Random Matrix Theory

- The sampled covariance matrix with $K \times N$ matrix $\mathbf{Y} = [\boldsymbol{y}(0), \ldots, \boldsymbol{y}(N-1)]$

$$\hat{\mathbf{R}} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^H$$

- $p_0(\mathbf{Y}; \sigma^2)$ and $p_1(\mathbf{Y}; \boldsymbol{h}, \sigma^2)$ are the likelihood functions of the observation matrix indexed by the unknown parameters h and σ^2 under hypotheses H_0 and H_1
- Y is a K x N matrix whose columns are i.i.d. Gaussian vectors with

Covariance Matrix

$$\boldsymbol{\Sigma} = \begin{cases} \sigma^2 \mathbf{I}_K, & \text{under } H_0 \\ \boldsymbol{h}\boldsymbol{h}^H + \sigma^2 \mathbf{I}_K, & \text{under } H_1 \end{cases}$$

Likelihood functions

$$p_0(\mathbf{Y}; \sigma^2) = (\pi\sigma^2)^{-NK} \exp\left(-\frac{N}{\sigma^2}\operatorname{tr}\hat{\mathbf{R}}\right)$$

$$p_1(\mathbf{Y}; \boldsymbol{h}, \sigma^2) = (\pi^K \det(\boldsymbol{h}\boldsymbol{h}^H + \sigma^2\mathbf{I}_K))^{-N} \times \exp(-N\operatorname{tr}(\hat{\mathbf{R}}(\boldsymbol{h}\boldsymbol{h}^H + \sigma^2\mathbf{I}_K)^{-1})).$$

P. Bianchi et al. (2011). IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4.

Statistical Tests for Single-Source Detection Using Random Matrix Theory

- In the case where $\boldsymbol{h}$ and σ^2 are unknown, **the classical approach**

$$L_N = \frac{\sup_{\boldsymbol{h},\sigma^2} p_1(\mathbf{Y};\boldsymbol{h},\sigma^2)}{\sup_{\sigma^2} p_0(\mathbf{Y};\sigma^2)}. \quad (5)$$

Reject hypothesis H0 whenever $L_N > \xi_N$ is a certain threshold which is selected in order that the PFA $\mathbb{P}_0(L_N > \xi_N)$ does not exceed a given level

- **Approach RMT**

Proposition 1: Let T_N be defined by

$$T_N = \frac{\lambda_1}{\frac{1}{K}\operatorname{tr}\hat{\mathbf{R}}}$$

then, the GLR [cf. (5)] writes

$$L_N = \frac{C}{(T_N)^N \left(1 - \frac{T_N}{K}\right)^{(K-1)N}}$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_K > 0$.

$$C = (1 - \tfrac{1}{K})^{(1-K)N}.$$

The GLRT reduces to the test which rejects the null hypothesis for large values of

$$\begin{matrix} H_1 \\ T_N \gtrless \gamma_N \\ H_0 \end{matrix} \qquad \gamma_N = \phi_{N,K}^{-1}(\xi_N)$$

where $\phi_{N,K} : x \mapsto Cx^{-N}(1 - \frac{x}{K})^{N(1-K)}$

P. Bianchi et al. (2011). IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4.

Statistical Tests for Single-Source Detection Using Random Matrix Theory

- **Exact Threshold and p-Values**
- The threshold is obtained by

$$\gamma_N = p_N^{-1}(\alpha)$$

where $p_N(t)$ represents the complementary c.d.f. of the statistics T_N under the null hypothesis

$$p_N(t) = \mathbb{P}_0(T_N > t).$$

- When the threshold is fixed, GLRT under the following form:

$$p_N(T_N) \underset{H_1}{\overset{H_0}{\gtrless}} \alpha.$$

P. Bianchi et al. (2011). IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4.

Statistical Tests for Single-Source Detection Using Random Matrix Theory

- **Exact Threshold and p-Values**
- T_N is a function of the eigenvalues where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_K > 0$.

$$p_N(t) = \int_{\Delta_t} p^0_{K,N}(x_1, \ldots, x_K) \mathrm{d}x_{1:K}$$

The domain of integration

$$\Delta_t = \left\{ (x_1, \ldots, x_K) \in \mathbb{R}^K, \frac{K x_1}{x_1 + \cdots + x_K} > t \right\}$$

P.D.F of the ordered eigenvalues

$$p^0_{K,N}(x_{1:K}) = \frac{\mathbf{1}_{(x_1 \geq \cdots \geq x_K \geq 0)}}{Z^0_{K,N}} \times \prod_{1 \leq i < j \leq K} (x_j - x_i)^2 \prod_{j=1}^{K} x_j^{N-K} e^{-N x_j}$$

P. Bianchi et al. (2011). IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4.

Statistical Tests for Single-Source Detection Using Random Matrix Theory

- **BEHAVIOR OF THE GLR STATISTICS - Behavior Under Hypothesis H0**
- With N, K go to infinity, the empirical distribution of the eigenvalues is Marcenko–Pastur distribution with $\lambda^+ = (1+\sqrt{c})^2$ and $\lambda^- = (1-\sqrt{c})^2$.

$$\mathbb{P}_{\check{\mathrm{M}}\mathrm{P}}(dy) = \mathbf{1}_{(\lambda^-,\lambda^+)}(y)\frac{\sqrt{(\lambda^+ - y)(y-\lambda^-)}}{2\pi c y}\,dy$$

- The largest eigenvalue converges a.s. to the **right edge** of the Marcenko–Pastur distribution. Let defined

$$\Lambda_1 = N^{2/3}\left(\frac{\lambda_1 - (1+\sqrt{c_N})^2}{b_N}\right)$$

converges a.s

$$F_{TW}(x) = \exp\left(-\int_x^\infty (u-x)q^2(u)\,du\right) \qquad \forall x \in \mathbb{R}$$

where b_N is defined by

$$b_N := (1+\sqrt{c_N})\left(\frac{1}{\sqrt{c_N}}+1\right)^{1/3}$$

$$q''(x) = xq(x) + 2q^3(x)$$
$$q(x) \sim \mathrm{Ai}(x) \quad \text{as} \quad x \to \infty$$

P. Bianchi et al. (2011). IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4.

Statistical Tests for Single-Source Detection Using Random Matrix Theory

- **BEHAVIOR OF THE GLR STATISTICS - Behavior Under Hypothesis H1**
- The covariance matrix writes $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_K + \mathbf{h}\mathbf{h}^*$
- Since the behavior T_N of is not affected if the entries of are multiplied by a given constant, it convenient to consider the model where $\mathbf{\Sigma} = \mathbf{I}_K + \frac{\mathbf{h}\mathbf{h}^*}{\sigma^2}$.
- The *signal-to-noise* ratio (SNR)

$$\rho_K = \frac{\|\mathbf{h}\|^2}{\sigma^2}$$

- The p.d.f. of the ordered eigenvalues writes

$$p_K^{1,N}(x_{1:K}) = \frac{\mathbf{1}_{(x_1 \geq \cdots \geq x_K \geq 0)}}{Z_{K,N}^1} \prod_{1 \leq i < j \leq K} (x_j - x_i)^2 \times \prod_{j=1}^{K} x_j^{N-K} e^{-N x_j} I_K\left(\frac{N}{K}\mathbf{B}_K, \mathbf{X}_K\right)$$

- X_K is the diagonal matrix with eigenvalues $(x_1, \ldots, x_K)$
- B_K is the KxK diagonal matrix with eigenvalues $(\frac{\rho_K}{1+\rho_K}, 0, \ldots, 0)$,
- The spherical integral with m_K the Haar measure on the unitary group of size K

$$I_K(\mathbf{C}_K, \mathbf{D}_K) = \int e^{K \mathrm{tr}(\mathbf{C}_K \mathbf{Q} \mathbf{D}_K \mathbf{Q}^H)} dm_K(\mathbf{Q})$$

P. Bianchi et al. (2011). IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4.

Statistical Tests for Single-Source Detection Using Random Matrix Theory

- **BEHAVIOR OF THE GLR STATISTICS - Behavior Under Hypothesis H1**
- The limiting behavior of the largest eigenvalue can change if the **signal-to-noise** ratio is large enough

Assumption 1: The following constant $\rho \in \mathbb{R}$ exists

$$\rho = \lim_{K\to\infty} \frac{\|\boldsymbol{h}\|^2}{\sigma^2} \left(= \lim_{K\to\infty} \rho_K\right).$$

We refer to ρ as the limiting SNR. We also introduce

$$\lambda_{\text{spk}}^{\infty} = (1+\rho)\left(1+\frac{c}{\rho}\right).$$

- Under hypothesis H1, the largest eigenvalue has the following asymptotic behavior as N, K go to infinity, the largest eigenvalue converges **outside** the support of Marcenko–Pastur distribution

$$\lambda_1 \xrightarrow[H_1]{a.s.} \begin{cases} \lambda_{\text{spk}}^{\infty}, & \text{if } \rho > \sqrt{c} \\ \lambda^{+}, & \text{otherwise} \end{cases}$$

P. Bianchi et al. (2011). IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4.

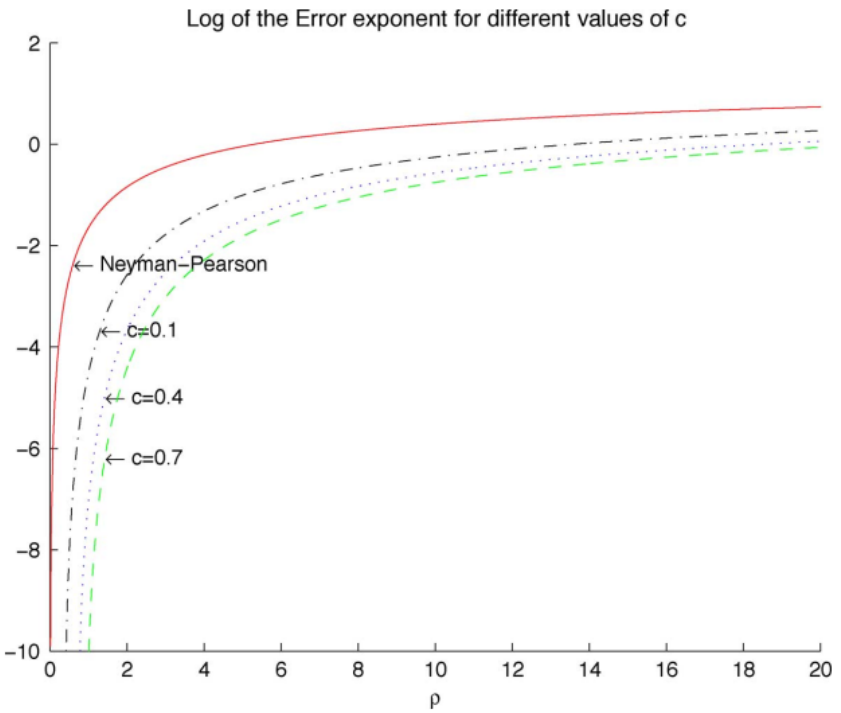
Statistical Tests for Single-Source Detection Using Random Matrix Theory

- **Limiting Behavior of T_N under H_0 and H_1**

Proposition 2: Let Assumption 1 hold true and assume $\rho > \sqrt{c}$, then

$$T_N \xrightarrow[H_0]{a.s.} (1+\sqrt{c})^2 \quad \text{and}$$

$$T_N \xrightarrow[H_1]{a.s.} (1+\rho)\left(1+\frac{c}{\rho}\right) \quad \text{as } N, K \to \infty.$$



Computation of the logarithm of the error exponent associated to the test T_N for different values of and comparison with the optimal result (Neyman-Pearson) obtained in the case where all the parameters are perfectly known.

P. Bianchi et al. (2011). IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4.

Population Structure and Eigenanalysis

- **Motivation:** methods for inferring population structure from genetic data do not provide formal significance tests for population differentiation.
- **Solution:**
  - An approach - principal components analysis - to studying population structure
  - Tracy–Widom Theory – a solid statistical footing, using results from modern statistics to develop formal significance tests.
  - Propose BBP threshold to estimate for data size needed for significant.
- **Approach:** PCA has three major features
  - Runs extremely quickly on large datasets (within a few hours on datasets with hundreds of thousands of markers and thousands of samples)
  - PCA framework provides the first formal tests for the presence of population structure in genetic data.
  - PCA method does not attempt to classify all individuals into discrete populations or linear combinations of populations
  - PCA outputs each individual's coordinates along axes of variation
  - https://github.com/chrchang/eigensoft

Nick Patterson et al. (2006). PLoS Genetics, Volume 2 , Issue 12, e190

Population Structure and Eigenanalysis

- **A Test for Population Structure:** This leads immediately to a formal test for the presence of population structure in a biallelic dataset
  1. Compute the matrix M

$$\mu(j) = \frac{\sum_{i=1}^{m} C(i,j)}{m} \qquad C(i,j) - \mu(j) \qquad M(i,j) = \frac{C(i,j) - \mu(j)}{\sqrt{p(j)(1-p(j))}}$$

  2. Compute X = MM'. X is mxm
  3. Order the eigenvalues of X so that $\lambda_1 > \lambda_2 ... > \lambda_{m'} > 0$ where m' = m-1
  4. Using the eigenvalues $\{\lambda_i\}_{1 \le i \le m}$ and estimate n' from

$$n' = \frac{(m+1)\left(\sum_i \lambda_i\right)^2}{\left((m-1)\sum_i \lambda_i^2\right) - \left(\sum_i \lambda_i\right)^2}$$

  5. The largest eigenvalue of M is λ_1. Set

$$l = \frac{(m')\lambda_1}{\sum_{i=1}^{m'} \lambda_i}$$

Nick Patterson et al. (2006). PLoS Genetics, Volume 2 , Issue 12, e190

Population Structure and Eigenanalysis

- Compute an eigenvector decomposition of X. Eigenvectors corresponding to ‘‘large’’ eigenvalues are exposing nonrandom population structure.

$$X = \frac{1}{n} MM'$$

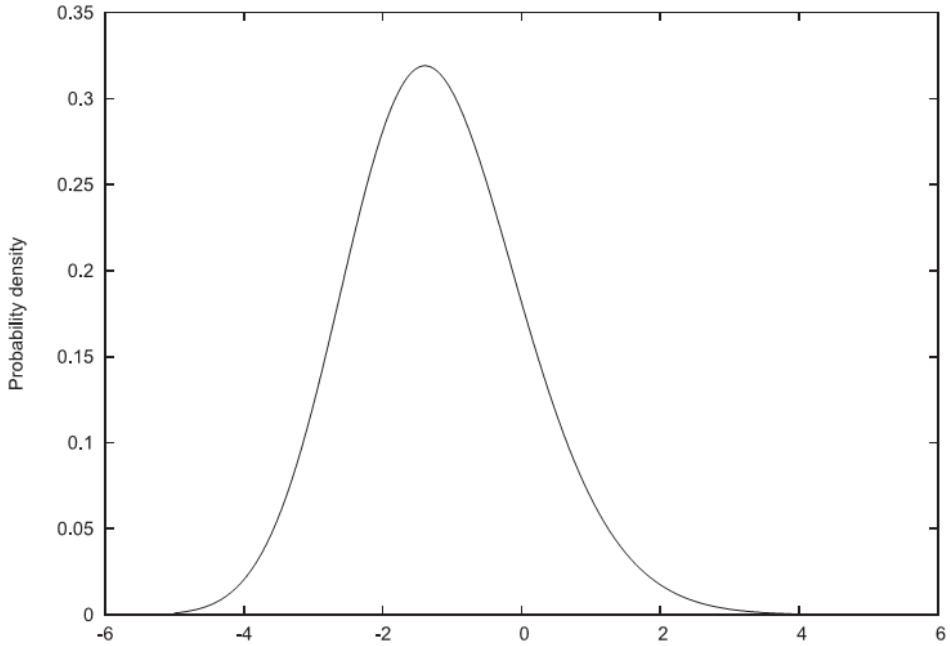
- Tracy–Widom Theory
  - X is a Wishart matrix. $\{\lambda_i\}_{1 \le i \le m}$ be t

$$\sigma(m,n) = \frac{(\sqrt{n-1}+\sqrt{m})}{n}\left(\frac{1}{\sqrt{n-1}}+\frac{1}{\sqrt{m}}\right)^{1/3}$$

$$\mu(m,n) = \frac{(\sqrt{n-1}+\sqrt{m})^2}{n}$$

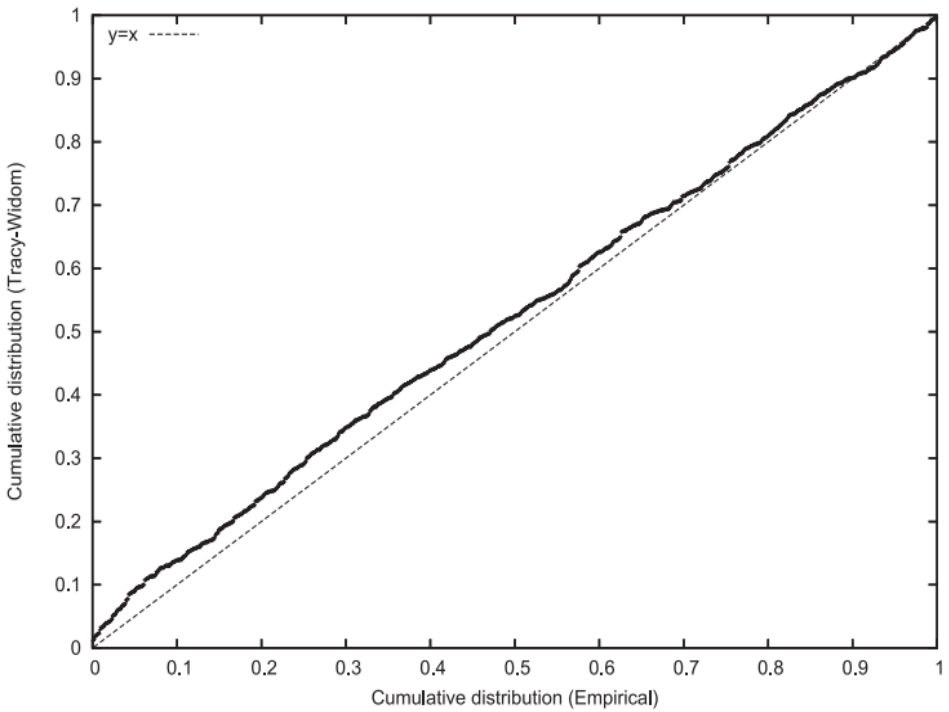
$$x = \frac{\lambda_1 - \mu(m,n)}{\sigma(m,n)}$$

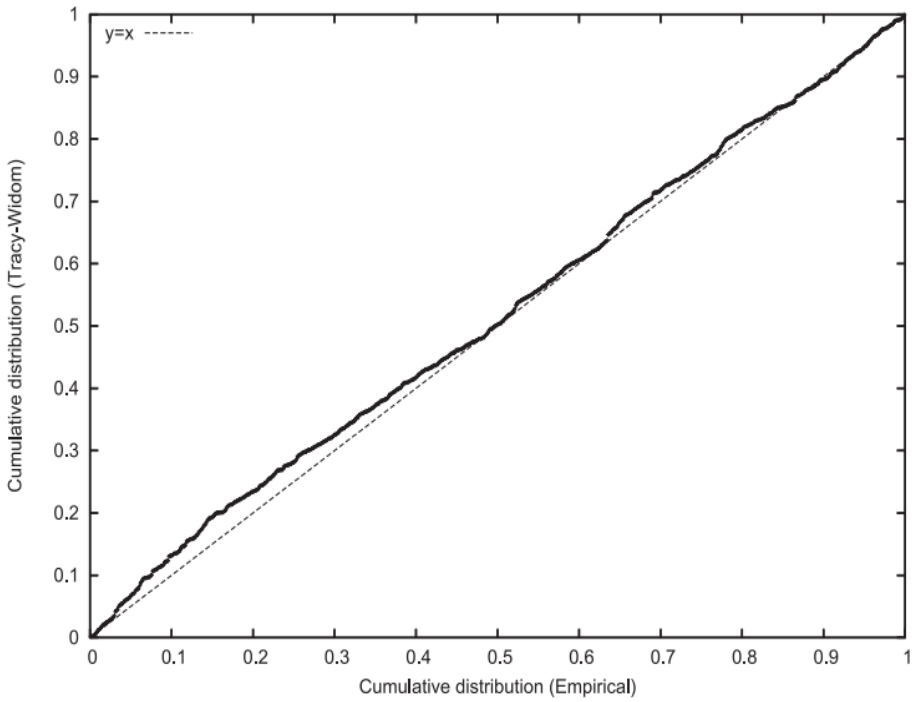
  - Figure 1.The Tracy–Widom Density



Nick Patterson et al. (2006). PLoS Genetics, Volume 2 , Issue 12, e190

Population Structure and Eigenanalysis





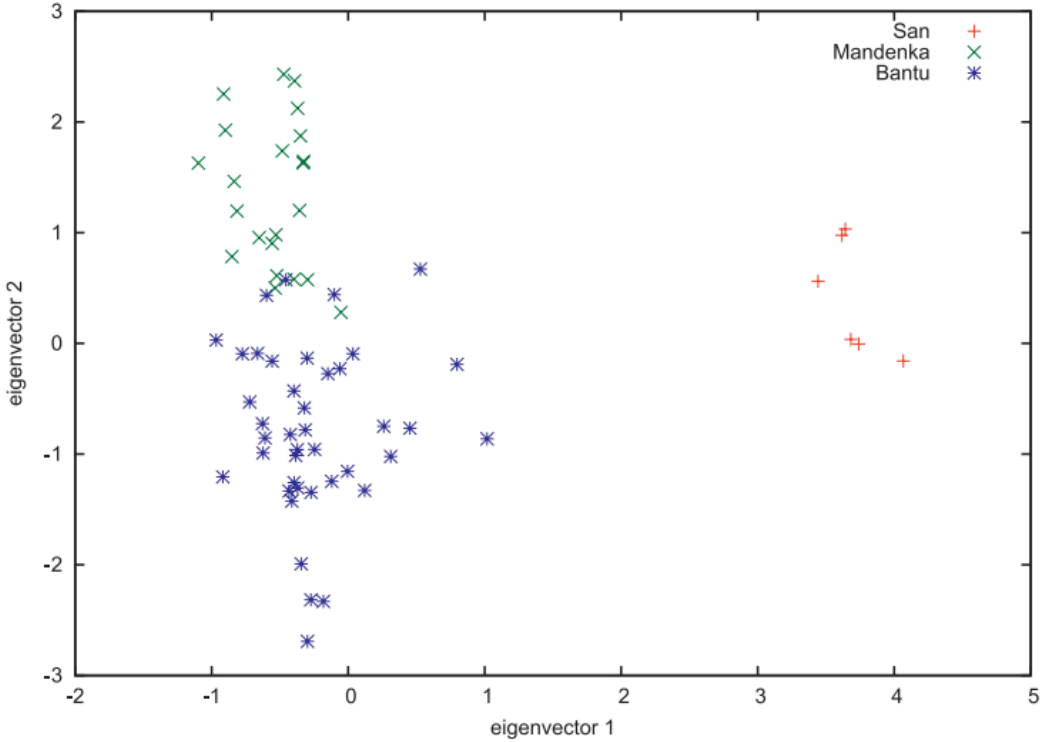
- 1,000 simulations of a panmictic population, a sample size of m =100 and n=5,000 unlinked markers.
- P–P plot of the TW statistic against the theoretical distribution. the fit is good, we expect the plot will lie along the line y =x. Interest is primarily at the top right, corresponding to low p-values

- P–P plot corresponding to a sample size of m=200 and n=50,000 markers.
- The fit is again excellent, demonstrating the appropriateness of the Johnstone normalization

Nick Patterson et al. (2006). PLoS Genetics, Volume 2 , Issue 12, e190

Population Structure and Eigenanalysis

- Figure 4. Three African Populations



- In Table 1, the ANOVA p-value is obtained from the usual F-statistic, and we apply ANOVA to each of the first three eigenvectors

Table 1. Statistics from HGDP African Data

Number	Eigenvalue	TW Statistic	TW *p*-Value	ANOVA *p*-Value
1	2.07	46.2	$<10^{-12}$	$<10^{-12}$
2	1.40	6.717	3.08×10^{-7}	$<10^{-12}$
3	1.31	0.380	.108	.74

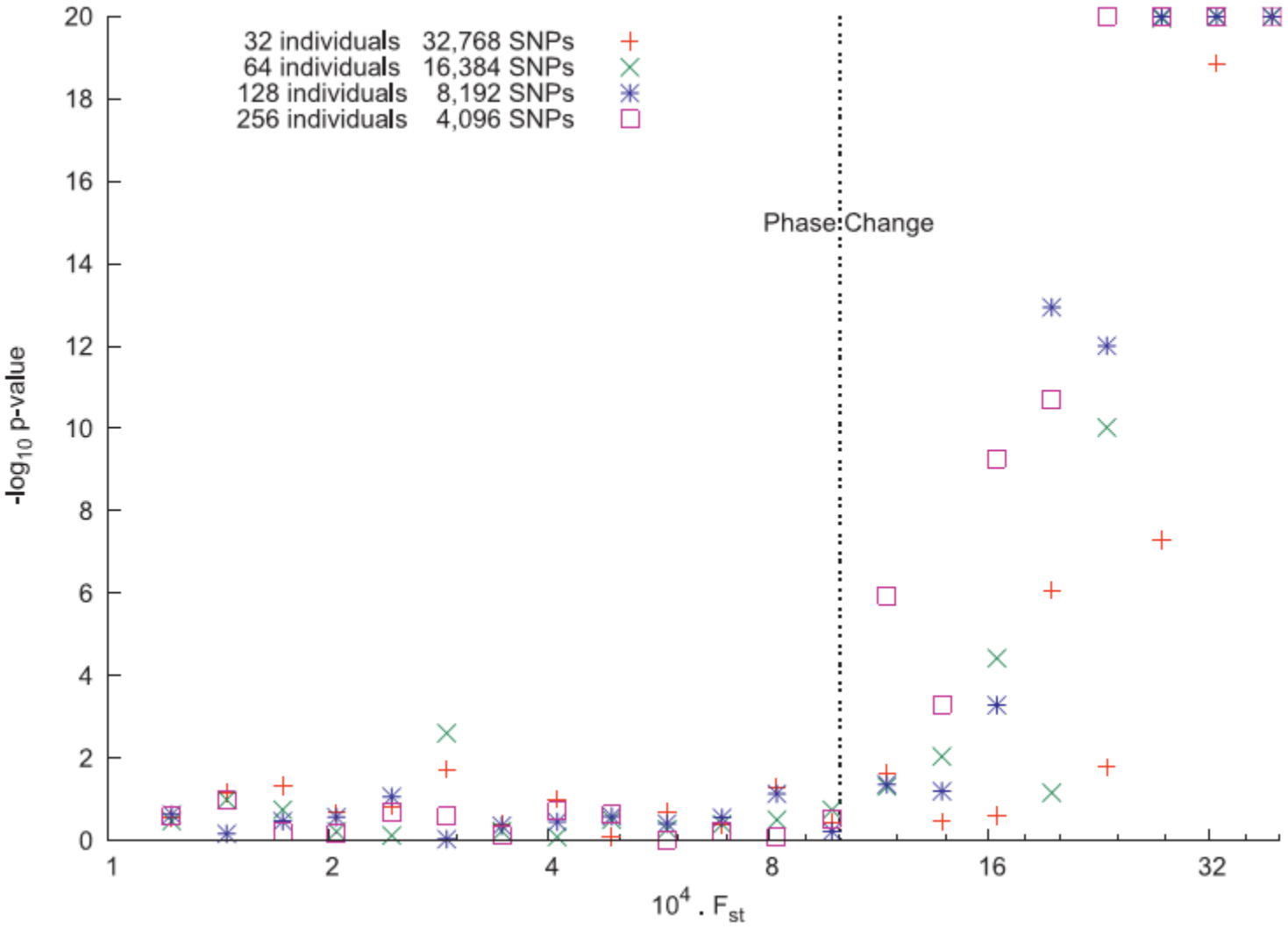
Population Structure and Eigenanalysis

- An Estimate for the Data Size Needed for Significance: a form of the conjecture, which we call the BBP conjecture.
- l_1 be the lead eigenvalue of the theoretical covariance matrix, with the remainder of the eigenvalues 1. Set $\gamma^2 = \frac{n}{m}$
- L_1 be the largest eigenvalue of the sample covariance
- the behavior of L_1 is qualitatively different depending on whether l_1 is greater or less than $1+ 1/\gamma$
- A phase-change phenomenon, as the BBP threshold. $1 + 1/\gamma = \frac{\sqrt{m} + \sqrt{n}}{\sqrt{n}}$
- Conclude: For two equal size subpopulation below which there will be essentially no evidence of population structure. Above the threshold, the evidence accumulates very rapidly, as we increase the divergence or the data size. Above the threshold for fixed data size mn, the evidence is stronger as we increase m, populations, there is a threshold value of F_{ST},
- Most large genetic datasets with human data will show some detectable population structure

Nick Patterson et al. (2006). PLoS Genetics, Volume 2 , Issue 12, e190

Population Structure and Eigenanalysis

- Figure 6. The BBP Phase Change

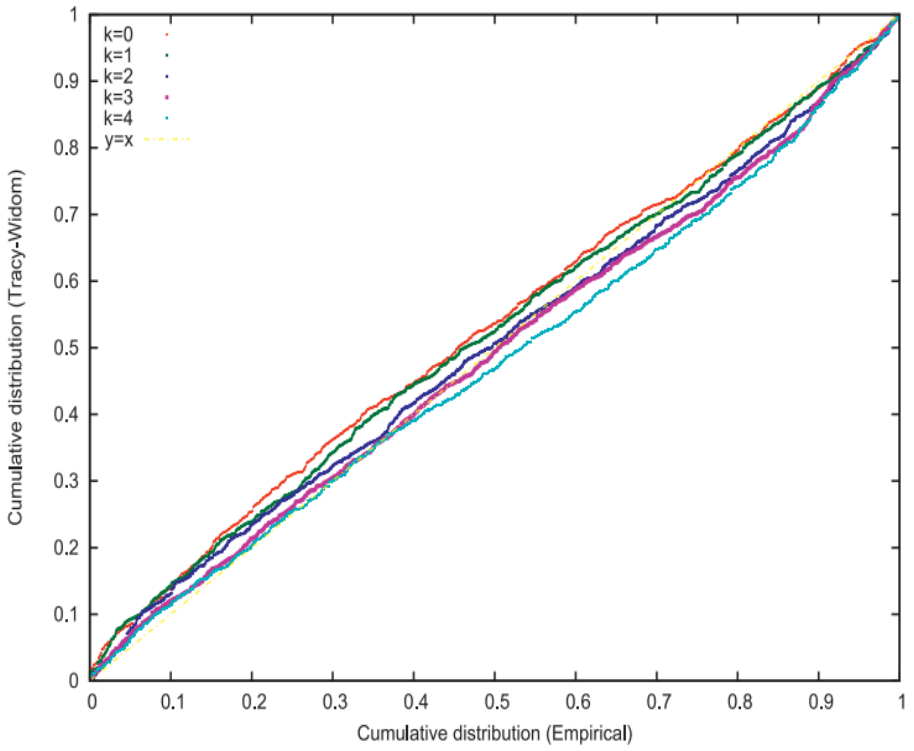


Nick Patterson et al. (2006). PLoS Genetics, Volume 2 , Issue 12, e190

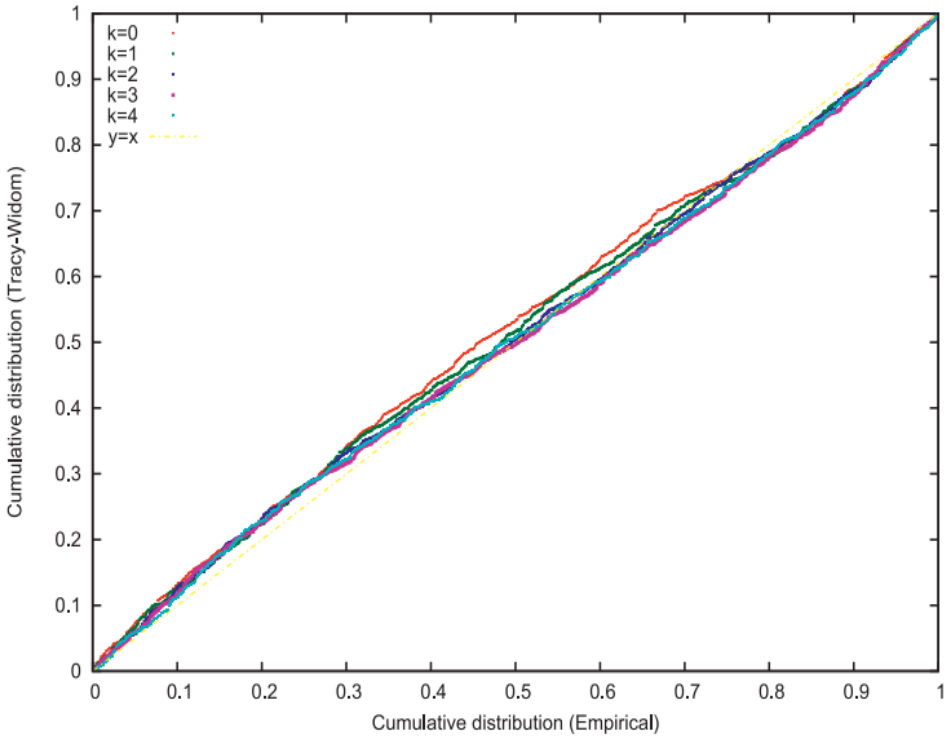
- There remain issues of SMARTPCA:
  1. Recent admixture generates large-scale LD which may cause difficulties in a dense dataset as the allele distributions are not independent. STRUCTURE 2.0 allows careful modeling.
  2. More ancient admixture, especially if the admixed population is genetically now homogeneous, may lead to a causal eigenvalue not very different from the values generated by the sampling noise.
  3. Methods require that divergence is small, and that allele frequencies are divergent primarily because of drift.
  4. If ‘‘admixture LD’’ is present, so that in admixed individuals long segments of the genome originate from one founder population, simple PCA methods will not be as powerful as programs such as STRUCTURE 2.0. LD will seriously distort the eigenvector/eigenvalue structure, making results difficult to interpret.
- Solution for “admixture LD” problem.
  1. Form matrix M $$M(i,j) = \frac{C(i,j) - \mu(j)}{\sqrt{p(j)(1-p(j))}}$$
  2. For each column j, set $$\mathbf{a} = a_s^{[j]}(1 \le s \le k)$$ $$R(i,j) = M(i,j) - \sum_{s=1}^{k} a_s^{[j]} M(i,j-s)(1 \le i \le m)$$
  3. Choose a to minimize $\sum_i R^2(i,j)$
  4. Calculate X = RR’ instead of MM’

Population Structure and Eigenanalysis

Figure 9. LD Correction with no LD Present



(A) for $m = 500$, $n = 5{,}000$,



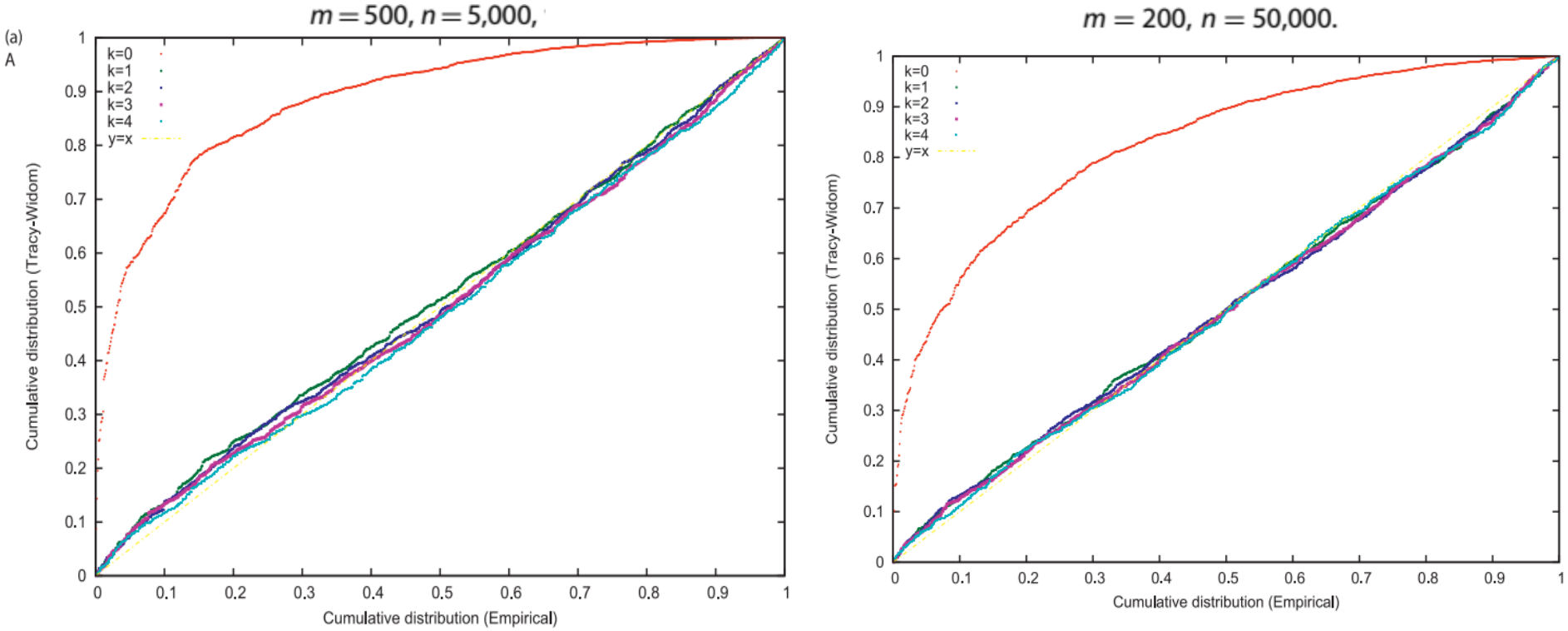
(B) for $m = 200$, $n = 50{,}000$.

- With five levels (k = 1. . .5) of correction, absence of LD the suggested correction does not seriously distort the Tracy–Widom statistic.
- In both cases the LD correction makes little difference to the fit.

Nick Patterson et al. (2006). PLoS Genetics, Volume 2 , Issue 12, e190

Population Structure and Eigenanalysis

Figure 10. LD Correction with Strong LD



- Uncorrected (k=0), the TW statistic is hopelessly poor
- After (k=1,.., 4) correction, the fit TW is again good
- Recommend that before analyzing a very large dataset with dense genotyping, one should filter the data by removing a marker from every pair of markers that are in tight LD.

Nick Patterson et al. (2006). PLoS Genetics, Volume 2 , Issue 12, e190

Eigenvalue Significance Testing for Genetic Association

- **Notation and Definitions**
- X to denote the $p \times n$ genotype matrix after appropriate normalization/scaling, where each row $i = 1, \ldots, p$ represents an ancestry-informative genetic marker, column $j = 1, \ldots, n$ represents an individual.
- $\overline{X}$ be the $p \times n$ matrix where the entries in each column are the column means of X and $\widetilde{X} = X - \overline{X}$
- The $n \times n$ sample covariance matrix of columns of X

$$S = \frac{1}{p-1} \tilde{X}^T \tilde{X}$$

- The eigen decomposition

$$S = \hat{Q} \hat{\Lambda} \hat{Q}^{-1}$$

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **Existing Methods**
- **The "naive" approach**
  - With N, p go to infinity, the empirical distribution function of sample eigenvalues approaches the standard Marcenko–Pastur (MP) distribution
  - The re-centered and scaled largest sample eigenvalue of *S*

$$T = (\hat{\lambda}_1 - \mu)/\sigma$$

μ is approximately the 0.832 quantile of $\hat{\lambda}_1$

$$\sigma = \frac{1}{p}(\sqrt{n} + \sqrt{p-1})(1/\sqrt{n} + 1/\sqrt{p-1})^{1/3}$$

↓

Tracy–Widom law of order 1

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **Existing Methods**
- **The robust approach**
  - Start with a genotype matrix *G*, which is centered and scaled for each marker *i* using $\bar{g}_j = \sum_i g_{ij}/2n$
  - A scaled matrix *M* with elements $m_{ij} = (g_{ij} - 2\bar{g}_j)/\sqrt{\bar{g}_j(1-\bar{g}_j)}$.
  - Eigenvalues are computed on the covariance matrix M'M/n.
  - The “effective" number of markers (Patterson et al. (2006)) is computed as

$$p' = \frac{(n+1)(\sum_j \hat{\lambda}_j)^2}{((n+1)\sum_j \hat{\lambda}_j^2) - (\sum_j \hat{\lambda}_j)^2}$$

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **Block Permutation and Residualization**
- **The first step**

1. For each marker i, the algorithm looks ahead L markers, and calculates $p_{i,i+l}$, the p-value for correlation between the genotypes at marker i versus each of markers i + l, l = 1, . . . , L.
2. If all of $p_{i,i+l} > p_{thresh}$ for l = 1, . . . , L, then the algorithm considers i to be the beginning of a new block of markers. Otherwise the algorithm extends the current block, increases i by 1, and repeats the comparison to the next L markers.
3. Once the blocks have been identified, block permutation is performed for the entire dataset at least 100 times (and preferably 1000) to determine the appropriate significance threshold.

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

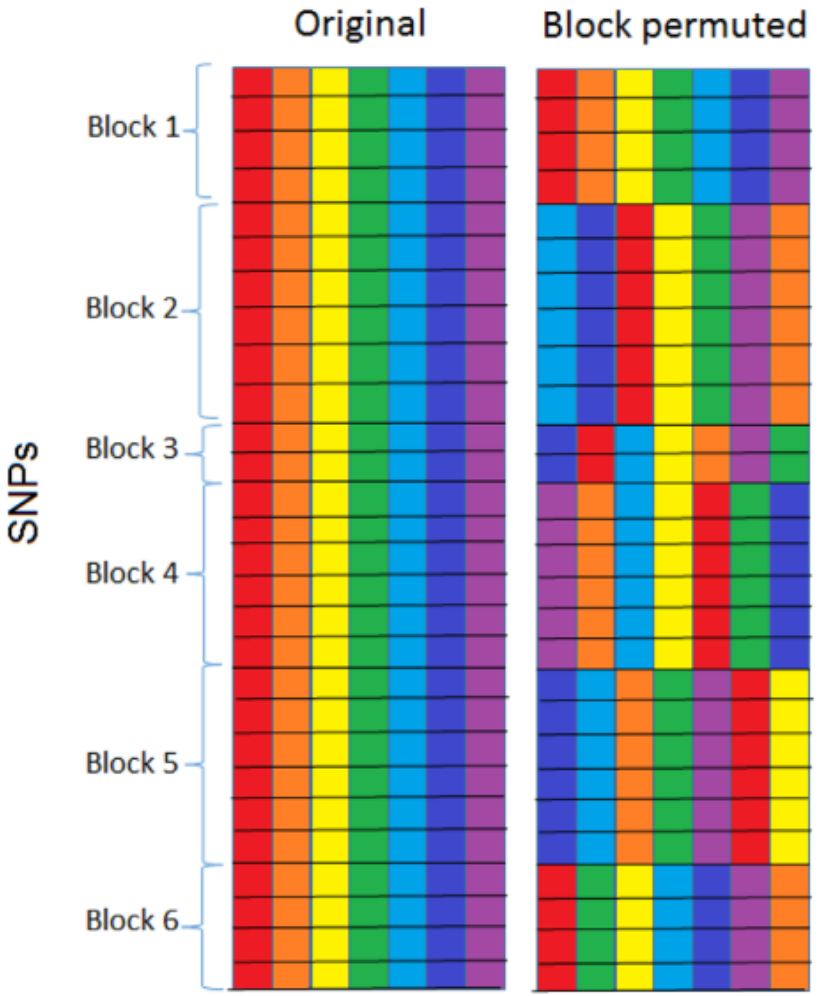


Figure 1. Illustration of block permutation, showing original genotype data on the left and a single permuted data set on the right. Colors represent the original individual source of the genotypes. After block boundaries are identified, within each block (submatrix) the order of columns is permuted.

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

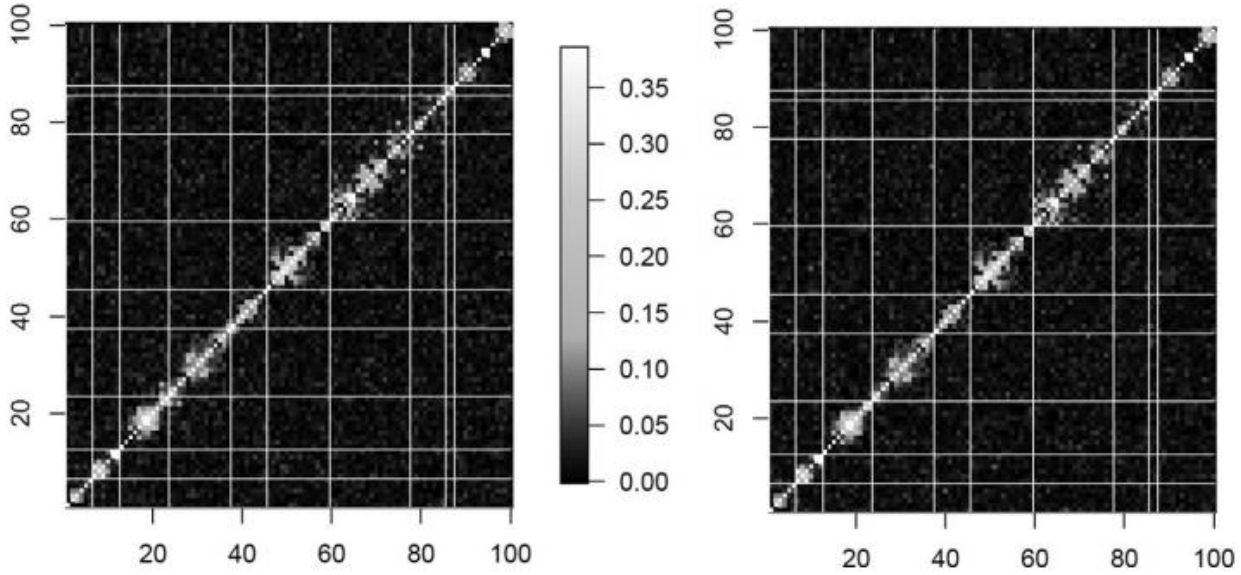


Figure 2. Illustration that block permutation preserves local correlation. For the first 100 markers in the CF data set, the absolute correlation $|r|$ is shown in the left panel. The white lines shows the boundaries from the block-finding method. The right panel shows the result after a single block permutation. The local correlation structure is identical before and after permutation (within blocks along the diagonal).

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **Block Permutation and Residualization**
- **The second step**
  - D_e denote the diagonal matrix of $\{d_1, \ldots, d_{n_e}, 0, \ldots, 0\}$ that is, only the extreme eigenvalues (subset largest eigenvalues) appear.
  - $\hat{X} = UD_eV^T$ is the standard rank-n_e matrix approximation to X
  - $X_{\text{resid}} = X - \hat{X}$ is a residual matrix after removing the "effects" of the extreme eigenvalues/eigenvectors.
  - For each block permutation indexed by π, $X_{\text{resid},\pi}$ is the permuted data, so $X_\pi = \hat{X} + X_{\text{resid},\pi}$ with eigenvalues $\hat{\lambda}_\pi = \{\hat{\lambda}_{1,\pi}, \ldots, \hat{\lambda}_{n_e,\pi}, \hat{\lambda}_{n_e+1,\pi}, \ldots, \hat{\lambda}_{n,\pi}\}$
  - Correct the bias by computing $\tilde{\lambda}_{j,\pi} = c\hat{\lambda}_{j,\pi}$, where $c = (\sum_{j=n_e+1}^{n} \hat{\lambda}_j)/(\sum_{j=n_e+1}^{n} \hat{\lambda}_{j,\pi})$
  - $\tilde{\bar{\lambda}}_{\pi,n_e+1} = \eta(\tilde{\lambda}_{\pi,n_e+1} - \bar{\tilde{\lambda}}_{n_e+1}) + \bar{\tilde{\lambda}}_{n_e+1}$ with $\eta = 1/\sqrt{1-\rho^2}$
  - For *K* random permutations, we compute the *p*-value

$$p_{perm} = \frac{1}{K}\sum_{\pi=1}^{K} I(\tilde{\bar{\lambda}}_{n_e+1,\pi} \geqslant \hat{\lambda}_{n_e+1}).$$

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

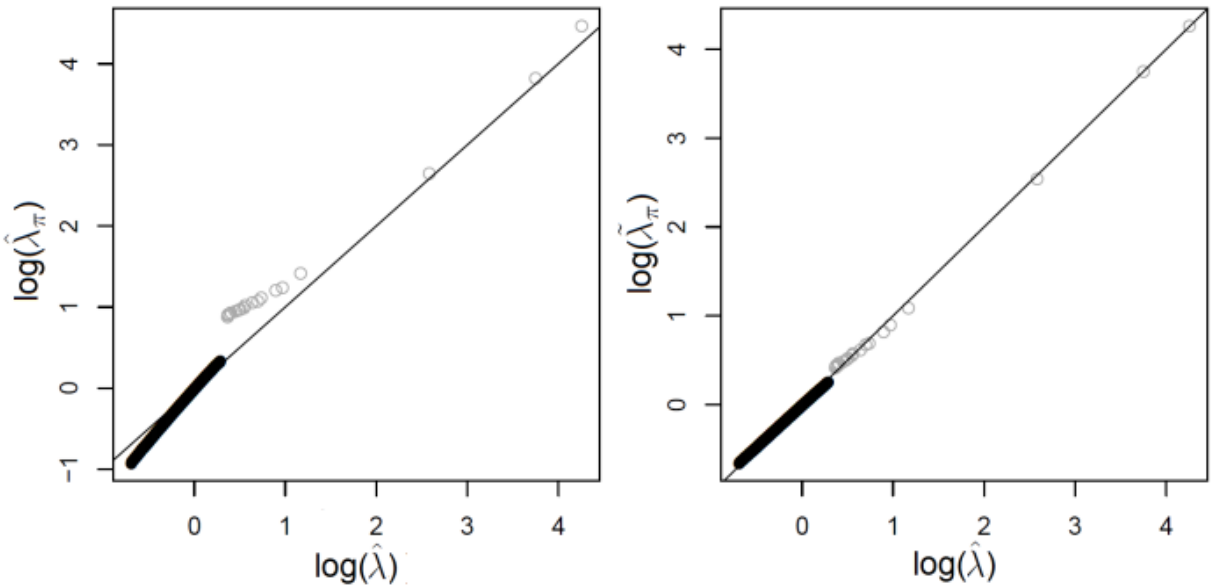


Figure S2: Illustration of bias of extreme eigenvalues under block permutation, for a simulation in Simulation Set 2 ($n = 2000, p = 19,681, n_e = 19$). Left panel: on the log-log scale, the extreme eigenvalues (grey circles) become even more extreme under the permutation scheme (y-axis) than the original sample eigenvalues (x-axis), due to the theoretical difference between true and sample eigenvalues. The phenomenon affects the non-extreme eigenvalues (black), for which inference about $\hat{\lambda}_{n_e+1}$ can be adversely affected. Right panel: Following a correction scheme in which bias is first assessed, corrected, and block permutation run again, the bias nearly disappears.

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **The General Marcenko–Pastur Equation**
- For the stochastic behavior of the largest eigenvalue, for $c \in [0, 1/\lambda_1)$ satisfying the relation

$$\int \left(\frac{\lambda c}{1-\lambda c} \right)^2 dX = n/p$$

$$\mu = \frac{1}{c}\left(1 + \frac{p}{n} \int \frac{\lambda c}{1-\lambda c} dX \right)$$

$$\sigma^3 = \frac{1}{c^3}\left(1 + \frac{p}{n} \int \left(\frac{\lambda c}{1-\lambda c} \right)^3 dX \right)$$

The statistic $T = (\hat{\lambda}_1 - \mu)/\sigma$ tends to the TW distribution

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **The General Marcenko–Pastur Equation**
- Use the efficient SPECTRODE algorithm for the density computation and modify the R code at https://github.com/dobriban/Spectrode-R to incorporate and fit eigenvalue model

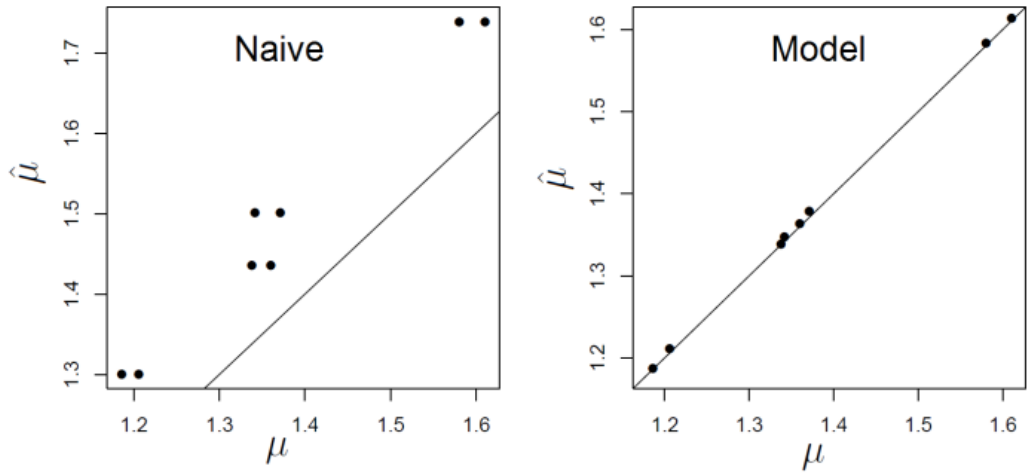


Figure S3: For the eight settings of Simulation Set 2, estimated μ values (averaged over simulations) vs. true μ values, computed by simulation as the 0.82 quantile of $\hat{\lambda}_{n_e+1}$. Left panel: results for the naive MP approach. Right panel: results from the SPECTRODE algorithm and using $\hat{\delta}$ for the discrete true eigenvalue model, fit separately for each simulation. The results illustrate the dramatic improvement of the discrete population eigenvalue model compared to the naive model.

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **Simulation Set 1**
  i. Sample sizes n = {1000,2000,4000}
  ii. Block sizes {10, 50}
  iii. ρ = {0.1,0.2}
  iv. FST = {0,0.01}
  v. Number of population strata ne + 1 = {2,6,11} (when FST > 0), keeping in mind that only ne eigenvectors are necessary to discriminate ne + 1 strata
  vi. The number of markers was p = 20,000 throughout.
  vii. For each of the 48 settings, we performed 1000 simulations.

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **Simulation Set 1**

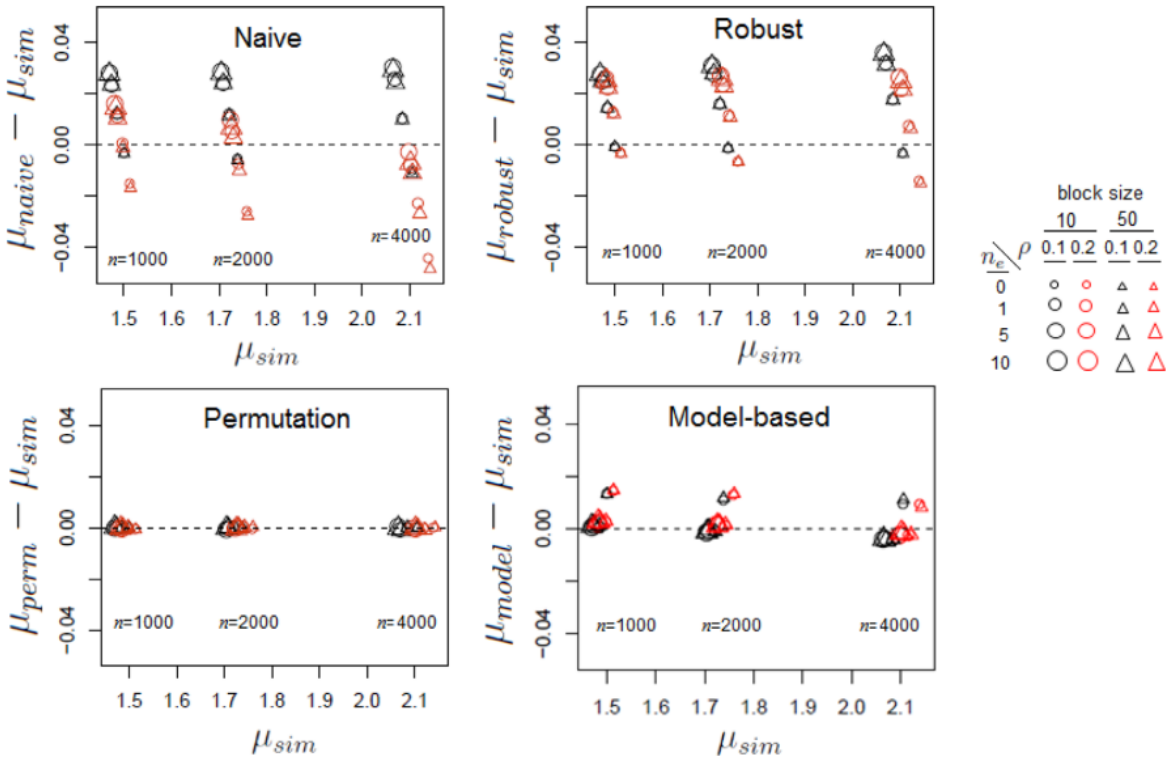


Figure 3. For Simulation Set 1 (48 simulation settings), block permutation and the model-based method are superior to existing methods. The figures show average estimates of μ versus the true values obtained by simulation, expressed as differences between estimates and true values. Upper left: Naive Marčenko–Pastur estimation of μ. Upper right: The results from the "robust" method, following the reasoning and sequential estimation from Patterson et al. (2006). Lower left: results from the proposed block permutation approach. Lower right: results from the proposed model-based approach.

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **Simulation Set 2**
- Implemented the method of Zhou et al. (2017), in which 4000 phased haploid genomes from 1000 Genomes Project Phase 3 v5 (1000 Genomes Consortium, 2015) were re-sampled, using an artificial recombination process to represent meiosis

Table 1

Results for Simulation Set 2, using resampling from ne populations/subpopulations using 1000 Genomes haploid genomes

n	p	n_e	$\bar{\hat{n}}_e$	$\hat{\alpha}_{\text{naive}}$	$\bar{R}_{\text{naive}}$	$\hat{\alpha}_{\text{robust}}$	$\bar{R}_{\text{robust}}$	$\hat{\alpha}_{\text{model}}$	$\bar{R}_{\text{model}}$	$\hat{\alpha}_{\text{perm}}$	$\bar{R}_{\text{perm}}$
1000	19,681	3	3.0	0	0	0	0	0.005	0.0	0.015	0.01
		19	18.6	0	0	0	0	0.003	0.0	0.007	0.01
	50,827	3	3.0	0	0	0	0	0.005	0.01	0.016	0.02
		19	19.0	0	0	0	0	0.048	0.03	0.051	0.04
2000	19,681	3	3.0	0	0	0	0	0.003	0.0	0.025	0.01
		19	19.0	0	0	0	0	0.020	0.0	0.010	0.02
	50,827	3	3.0	0	0	0	0	0.048	0.03	0.010	0.00
		19	19.0	0	0	0	0	0.042	0.04	0.049	0.03

Zhou et al. (2018). Biometrics 74, 439–447

Eigenvalue Significance Testing for Genetic Association

- **Simulation Set 2**

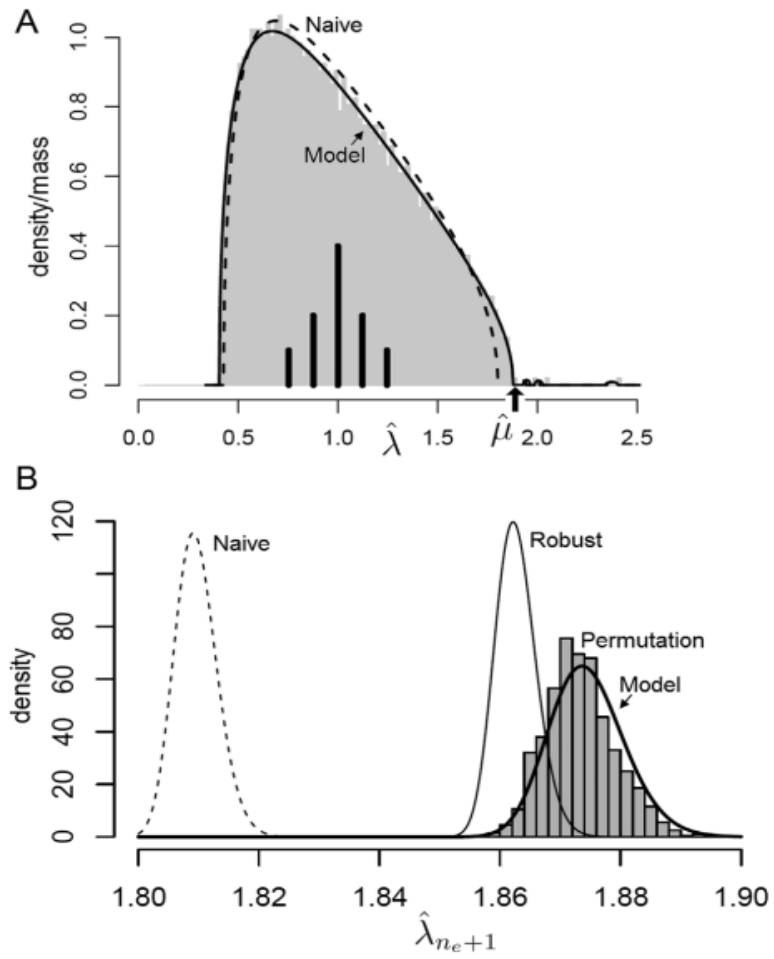


Figure 4. Marcenko–Pastur fitting and Tracy–Widom testing for the CF data and $\hat{\lambda}_{n_e+1}$, where $n_e = 6$. Upper panel: Histogram of all eigenvalues, overlaid with fits from standard (naive) WP, and the fit from the proposed model ($\hat{\delta} = 0.12$). $\hat{\mu}$ marks the value where the bulk portion of the fitted density $f_{\hat{\delta}}$ reaches zero, although the density is positive again for larger values driven by the extreme eigenvalues. The black bars under the histogram show $h_{\hat{\delta}}$. Lower panel: Densities for the naive, robust, and model-based approaches use a Tracy–Widom distributional approximation. The model-based density closely matches the permutation approach. The naive method declares 26 eigenvalues to be significant at intended $\alpha = 0.05$, including the 6 extreme eigenvalues, followed by robust (8 declared significant), corrected block permutation (7), and the model-based approach (7).

Zhou et al. (2018). Biometrics 74, 439–447

Reference

1. Johnstone, Iain M. "On the distribution of the largest eigenvalue in principal components analysis." The Annals of statistics 29.2 (2001): 295-327.
2. Soshnikov, Alexander. "A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices." Journal of Statistical Physics 108.5-6 (2002): 1033-1056.
3. Baik, Jinho, Gérard Ben Arous, and Sandrine Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices." The Annals of Probability 33.5 (2005): 1643-1697.
4. Kritchman, Shira, and Boaz Nadler. "Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory." IEEE Transactions on Signal Processing 57.10 (2009): 3930-3941.
5. Bianchi, Pascal, et al. "Performance of statistical tests for single-source detection using random matrix theory." IEEE Transactions on Information theory 57.4 (2011): 2400-2419.
6. Patterson, Nick, Alkes L. Price, and David Reich. "Population structure and eigenanalysis." PLoS genetics 2.12 (2006): e190.
7. Zhou, Yi-Hui, J. S. Marron, and Fred A. Wright. "Eigenvalue significance testing for genetic association." Biometrics 74.2 (2018): 439-447