High Dimensional Statistical Inference and Random Matrices

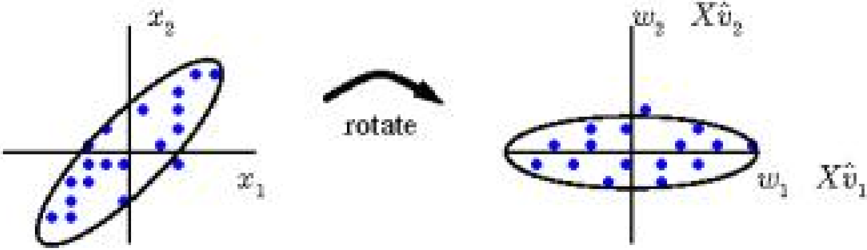
Iain Johnstone, Statistics, Stanford

imj@stanford.edu

PCA on Observed Data - 2

$$S\hat{\mathbf{v}}_j = \hat{\ell}_j \hat{\mathbf{v}}_j; \qquad \hat{w}_j = X\hat{\mathbf{v}}_j$$

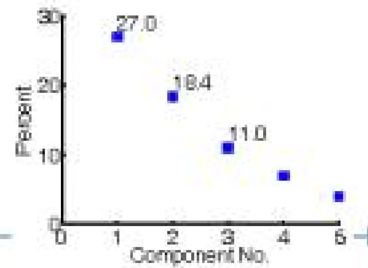
Sample PC eigenvalues $\hat{\ell}_j$; **Sample PC eigenvectors $\hat{\mathbf{v}}_j$:**



"% variance explained plot:" j **vs.** $\hat{\ell}_j / \sum \hat{\ell}_{j'}$

How many $\hat{\ell}_j$ are "significant"?

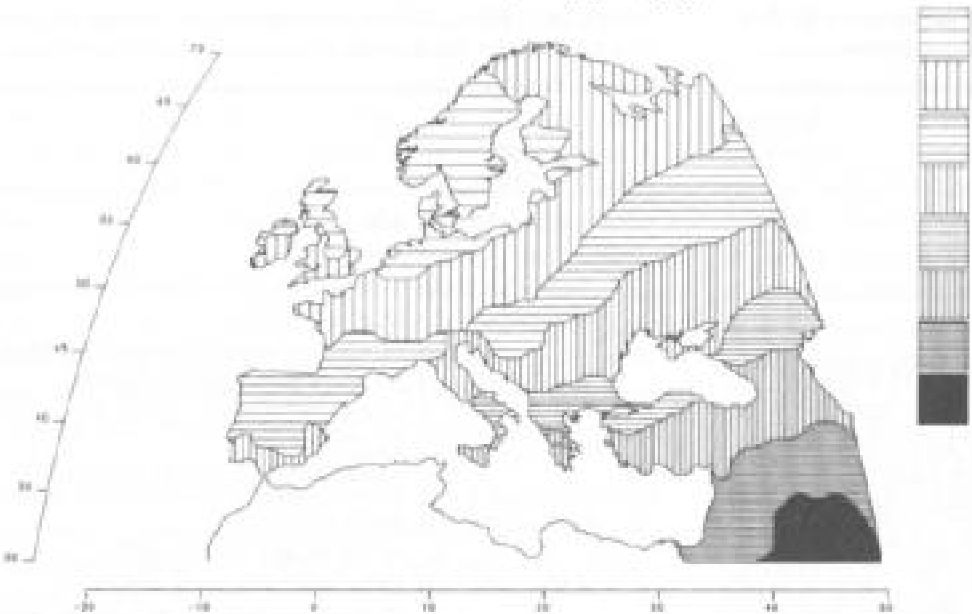
Discard $\hat{w}_j$, $j > j_0$.



CS-M-P: First Principal Component

Observations i have locations loc$[i]$, so for each PC $\hat{w}_j = X\hat{v}_j$,

contour plot of $(\text{loc}[i], \hat{w}_j[i])$

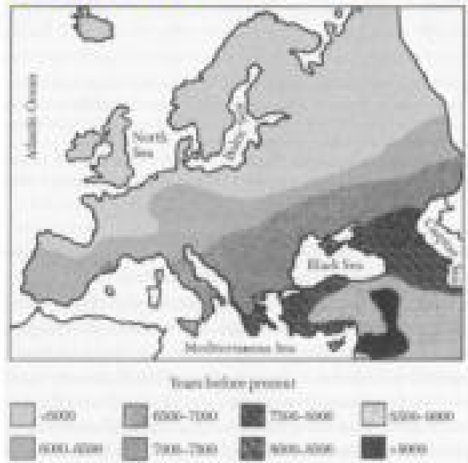


Scientific Significance

Did farmers or farming expand from Asia Minor into Europe?

Genetics: First PC

Archaeology of farming



Cavalli-Sforza, books 1994, 2000, esp. History & Geography of Human Genes.

Summary

	Population (unknown)	Data (observed)
Variables	p variables $X_1, \ldots, X_p$	$X = [\mathbf{x}_1 \cdots \mathbf{x}_p]$
Sample size		n
Covariance matrix	Σ	$S = n^{-1} X^T X$
P.C. eigenvalues	ℓ_j	$\hat{\ell}_j$
P.C. eigenvectors	$\mathbf{v}_j$	$\hat{\mathbf{v}}_j$

Data are noisy/variable/limited, so interest in estimation error

$$\hat{\ell}_j(X) - \ell_j, \qquad \hat{\mathbf{v}}_j(X) - \mathbf{v}_j$$

and how many components $\hat{\ell}_j$ are "significant"?

Outline

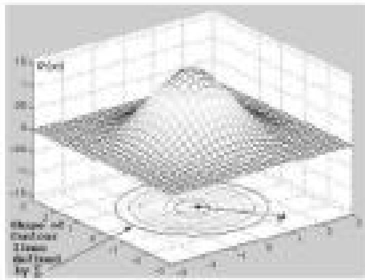
1. Principal Components Analysis
2. Gaussian & Wishart Distributions
3. Random matrices
4. Large p Asymptotics
5. Largest Eigenvalue Laws
6. Beyond the "Null Hypothesis"
7. Estimating Eigenvectors

Multivariate Gaussian distribution

$\vec{X} \sim N_p(\mu, \Sigma)$ **Mean** $\mu = \mathbb{E}X$

Covariance $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^T$

Density: $f(X) = |\sqrt{2\pi}\Sigma|^{-1/2} \exp\{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\}$



Standard model: n **independent draws**

$$\vec{X}_1, \ldots, \vec{X}_n \sim N_p(\mu, \Sigma).$$

Wishart Distribution $W_p(n, \Sigma)$

$$X = \begin{pmatrix} \vec{X}_1 \\ \vdots \\ \vec{X}_n \end{pmatrix}$$

n **rows, independent** $N_p(0, \Sigma)$

Definition: $A = X^T X \sim W_p(n, \Sigma)$

p **variate Wishart distribution.** n **"degrees of freedom"**

E.g. sample covariance matrix $A = nS$

Density function (Wishart, 1928):

$$f(A) = c_{np} |\Sigma|^{-n/2} |A|^{(n-p-1)/2} \text{etr}\{-\tfrac{1}{2}\Sigma^{-1} A\}$$

Wishart and PCA

With n independent Gaussian data, eigenstructure of Wishart distribution $\leftrightarrow$ PCA of X (Hotelling, 1933).

$$\underset{n\times p}{X} = \begin{pmatrix} \vec{X}_1 \\ \vdots \\ \vec{X}_n \end{pmatrix}$$

Thus, if $A = nS \sim W_p(n, \Sigma)$ and

$$Au_i = l_i u_i \qquad l_1 \geq \ldots \geq l_p \geq 0.$$

then $\begin{pmatrix} l_i = n\hat{\ell}_i \\ v_i = \hat{\mathbf{v}}_i \end{pmatrix}$ are PC $\begin{pmatrix} \text{eigenvalues} \\ \text{eigenvectors} \end{pmatrix}$ of X

Canonical Correlations

$[\vec{X} \mid \vec{Y}] = [X_1 \; \ldots \; X_p \mid Y_1 \; \ldots \; Y_q]$ **jointly** $p+q$**–var. Gaussian**

"Most predictable criterion": (Hotelling, 1935, 1936).

$$\max_{u_i, v_i} \textbf{Corr}\left(u_i^T \vec{X}, v_i^T \vec{Y}\right)$$

With n **samples** $(\vec{X}_i, \vec{Y}_i)$, $i = 1, \ldots, n$**,**

$$\Rightarrow \quad A v_j = r_j^2 (A + B) v_j, \qquad r_1^2 \geq \ldots \geq r_p^2.$$

Two *independent* **Wishart distributions:**

$$A \sim W_p(q, \Phi; \Omega), \qquad B \sim W_p(n - q, \Phi).$$

Double Wishart Setting

$$A \sim W_p(n_1, I)$$ **2 independent Wisharts,** $p \le n_1, n_2$

$$B \sim W_p(n_2, I)$$ **"null hypothesis" setting**

Common feature: roots $:= (x_i)_{i=1}^p$ **of generalized eigenproblem:**

$$\det[x(A+B) - A] = 0$$

Single Wishart

- Principal Component analysis
- Factor analysis
- Multidimensional scaling

Double Wishart

- Canonical correlation analysis
- Multivariate Analysis of Variance (MANOVA)
- Multivariate regression analysis
- Discriminant analysis
- Tests of equality of covariance matrices

Orientation

- How does mathematics influence statistics?

Joint density of eigenvalues, 1939

Fisher
(Cambridge)

Girshick
(Columbia)

Hsu
(London)

Mood
(Princeton)

Roy
(Calcutta)

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

Single Wishart: $w(x) = x^{n-p} e^{-x},$ **(Laguerre)**

Double Wishart: $w(x) = x^{p-q-1}(1-x)^{n-p-q-1}.$ **(Jacobi)**

Outline

1. Principal Components Analysis
2. Gaussian & Wishart Distributions
3. Random matrices
4. Large p Asymptotics
5. Largest Eigenvalue Laws
6. Beyond the "Null Hypothesis"
7. Estimating Eigenvectors

Random Matrices in Physics

Energy levels of nuclei $\leftrightarrow$ eigenvalues of Hamiltonian

$$H\psi_i = E_i\psi_i, \qquad E_0 \le E_1 \le \dots$$

Wigner: (1950s) ***statistical*** **description of higher energy levels**

Model $\{E_i\}$, i large by eigenvalues of

$$H_N = (H_{ij})_{N\times N}, \qquad \text{“random”, symmetric}$$

Semicircle Law: H_{ij} **i.i.d., mean 0, var σ^2, $i \le j$**

$F_N =$ **empirical d.f. of eigenvalues $x_1, \dots, x_N$**

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

E.P. Wigner

Ensembles and Orthogonal Polynomials

Joint density of eigenvalues $x_1, \ldots, x_N$**:**

$$c \prod_1^N w(x_i)^{\beta/2} \prod_{i<j} |x_i - x_j|^\beta.$$

Classical orthogonal polynomials (Fox-Kahn, 1964)

$w(x) = e^{-x^2/2}$	**Hermite**
$e^{-x} x^a$	**Laguerre**
$(1-x)^a (1+x)^b$	**Jacobi**

Ensembles and Orthogonal Polynomials

Joint density of eigenvalues $x_1, \ldots, x_N$:

$$c \prod_1^N w(x_i)^{\beta/2} \prod_{i<j} |x_i - x_j|^\beta.$$

Classical orthogonal polynomials (Fox-Kahn, 1964)

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

Dyson's "threefold way"

	Symmetry Type	*Matrix entries*	
$\beta = 1$	**orthogonal**	**real**	
$\beta = 2$	**unitary**	**complex**	**[easiest]**
$\beta = 4$	**symplectic**	**quaternion**	

So, classical multivariate H_0 **distributions** $\Leftrightarrow$

$$\left\{ \begin{array}{c} \textbf{Gaussian} \\ \textbf{Laguerre} \\ \textbf{Jacobi} \end{array} \right\} \left\{ \begin{array}{c} \textbf{Orthogonal} \\ \textbf{Unitary} \\ \textbf{Symplectic} \end{array} \right\} \textbf{Ensemble : \quad LOE etc.}$$

Some uses of RMT in Statistics

Eigenvalues:

Bulk	Graphical methods [finance, communications]
Linear Statistics	Hypothesis tests, distribution theory
Extremes	Hypothesis tests, distribution theory, role in proofs
Spacings	[Few so far]
General	Computational tools, role in proofs

Eigenvectors:

Transforms	subspace estimation

Outline

1. Principal Components Analysis
2. Gaussian & Wishart Distributions
3. Random matrices

4. Large p Asymptotics
5. Largest Eigenvalue Laws
6. Beyond the "Null Hypothesis"
7. Estimating Eigenvectors

Asymptotic Regimes

Exact distributions complicated $\rightarrow$ asymptotic approximations

Traditionally: **Statistics:** $n \rightarrow \infty$, p **fixed** Anderson, Muirhead

RMT: $N \rightarrow \infty$ Mehta

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

Modern asymptotics:

- n, p **both large** (e.g. $p = cn$) $\rightarrow$
- **Plancherel-Rotach type** $(\alpha = (1-c)N)$.

Spread of Sample Eigenvalues

$$nS \sim W_p(n, \Sigma)$$

Phenomenon: sample eigenvalues are (much) more spread out than those of population.

Population: $\ell_j = \ell_j(\Sigma);$ for $\Sigma = I, \;\ell_j \equiv 1$

Sample: $\hat{\ell}_j = \hat{\ell}_j(S);$

Typical sample for $n = p = 10 \rightarrow$ sample eigenvalues of S

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

$\Rightarrow$ condition number $(\hat{\ell}_{max}/\hat{\ell}_{min}) \approx 1000!$

Orientation

- How does mathematics influence statistics?
- Challenges of high data volume, many dimensions

The Quarter Circle Law

Description of spreading phenomenon:

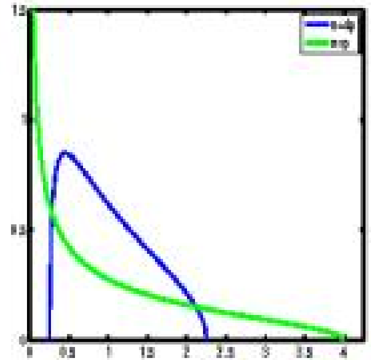
Empirical distribution function: for eigenvalues $\{\hat{\ell}_i\}_{j=1}^{p}$

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

Marčenko-Pastur, (67) For $A \sim W_p(n, I) \qquad p/n \to \gamma$

For $\Sigma = I$,

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



Outline

1. Principal Components Analysis
2. Gaussian & Wishart Distributions
3. Random matrices
4. Large p Asymptotics
5. Largest Eigenvalue Laws
6. Beyond the "Null Hypothesis"
7. Estimating Eigenvectors

Hypothesis Test for Largest Eigenvalue

Observed eigenvalue data for $n = p = 10$:

$$(\hat{\ell}_i) = (0.002, 0.06, \ldots 1.98, 2.52, 4.25)$$

Is observed largest value 4.25 consistent with $nS \sim W_p(n, I)$?

Terminology: **Null hypothesis:** $nS \sim W_p(n, \mathbf{I})$.

Alternative hypothesis: $nS \sim W_p(n, \mathbf{\Sigma}), \qquad \Sigma \neq I.$

Compare sample-to-sample variation: 3 samples from $W_p(n, I)$

0.03	0.12	…	1.38	2.11	2.91
0.0003	0.007	…	1.72	2.52	3.40
0.0000	0.12	…	1.90	2.11	3.50

$\Longrightarrow$ need **Null hypothesis distribution:**

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

Tracy Widom Limits

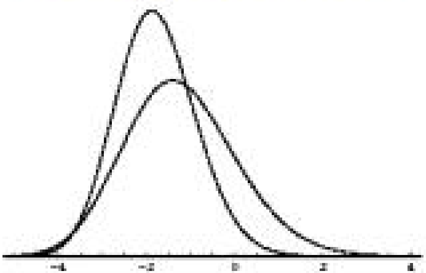
For {real, complex}, {single, double} Wishart matrices, if

$n/p \to \gamma,$ **[or** $(n_1/p, n_2/p) \to (\gamma_1, \gamma_2)$**,]** **then**

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

Johansson, IMJ

Tracy-Widom distributions (1994,1996):



$$F_2(s) = e^{-\int_s^\infty (x-s)^2 q(x) dx}$$

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

$q'' = sq + 2q^3$ **(Painlevé II)**

$q(s) \sim$ **Ai**(s) **as** $s \to \infty$

Second order accuracy

For {real, complex}, {single, double } Wishart matrices, if

$n/p \to \gamma$, [or $(n_1/p, n_2/p) \to (\gamma_1, \gamma_2)$,] then

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

Johansson, IMJ, El Karoui

E.g. Real, Single Wishart:

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

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

Approximation vs. Tables for $p = 5$

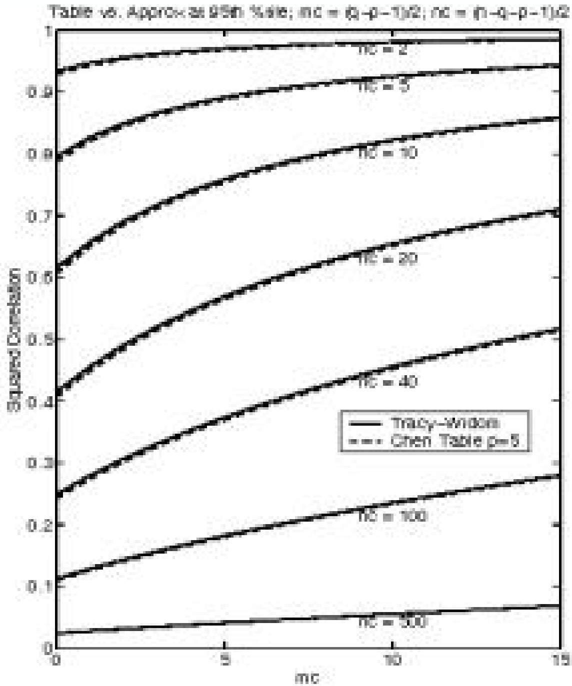
Double Wishart case:

Tables of 95th percentile:

[William Chen, IRS, 2002]

$$m_c = \frac{n_1 - p - 1}{2} \in [0, 15],$$

$$n_c = \frac{n_2 - p - 1}{2} \in [1, 1000]$$



A different domain of attraction

"Extreme value" theory approach to maxima is (classically) infeasible:

$$\{\max_{1\le i\le p} l_i \le t\} = \prod_{i=1}^{p} I\{l_i \le t\}$$

Key role: *determinants*, not independence:

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

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

$\cdots \Rightarrow$ **determinantal formula for** $P\{\max_{1\le i\le p} l_i \le t\}$

Correlation kernels & Statistics payoff

Complex data

$$\begin{cases} P\{\max l_i \le t\} & = \det(I - K_p \chi_t) \\ K_p(x,y) & = \sum_{k=1}^{p} \phi_k(x)\phi_k(y) \end{cases}$$

Distribution	Ensemble	(Weighted) Polynomials
Gaussian	Hermite	$\phi_k \propto \sqrt{w} H_k$
Single Wishart	Laguerre	$\phi_k \propto \sqrt{w} L_k^{\alpha}$
Double Wishart	Jacobi	$\phi_k \propto \sqrt{w} P_k^{\alpha,\beta}$

Real data

$$\begin{cases} P\{\max l_i \le t\} & = \sqrt{\det(I - K_p \chi_t)} \\ K_p(x,y) & = \begin{pmatrix} \tilde{K}_p & -D_2 \tilde{K}_p \\ \epsilon_1 \tilde{K}_p & \tilde{K}_p^T \end{pmatrix}, \\ & \tilde{K}_p = K_p + r_1 \end{cases}$$

Back to Example

Observed eigenvalue data for $n = p = 10$:

$$(\hat{\ell}_i/n) = (0.002, 0.06, \ldots, 1.98, 2.52, \mathbf{4.25})$$

Is observed largest value 4.25 consistent with $nS \sim W_p(n, I)$?

Yes! From Tracy-Widom approximation, $\mu_{np} = 3.8, \sigma_{np} = 0.53$,

$$P\{\hat{\ell}_1 > 4.25 \,|\, W_p(n, I)\} \approx 0.061$$

Next question: power of test: could a difference be detected?

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

e.g. if $\Sigma = (1, \ldots, 1, 3)$.

Outline

1. Principal Components Analysis
2. Gaussian & Wishart Distributions
3. Random matrices
4. Large p Asymptotics
5. Largest Eigenvalue Laws
6. Beyond the "Null Hypothesis"
7. Estimating Eigenvectors

Orientation

- How does mathematics influence statistics?
- Challenges of high data volume, many dimensions
- today – *one* example: multivariate statistics & RMT

Beyond the "Null Hypothesis"

Classical RMT ensembles (e.g. $W_p(n, I)$)

$\leftrightarrow$ "null hypothesis", symmetry, no structure.

Great interest for asymmetric situations, e.g. $W_p(n, \Sigma)$:

$$\frac{f_\Sigma(l_1, \ldots, l_p)}{f_I(l_1, \ldots, l_p)} = c|\Sigma|^{-n/2} {}_0F_0(-\tfrac{1}{2}\Sigma^{-1}, L)$$ James, 60, 64

- power of test: $P\{l_1 > t|\Sigma\}$ for $\Sigma \neq I$
- confidence interval for, e.g. $\lambda_1(\Sigma)$
- specific applications
  - block diagonal Σ (genes; stocks)
  - Toeplitz Σ (stationary processes)

Persistence of Tracy-Widom Limit

Back to PCA: For what conditions on Σ does

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

Some **answers:**

- **sufficiently many $\ell_k(\Sigma)$ accumulate near $\ell_1(\Sigma)$** El Karoui
- **small number of (not too!) isolated $\ell_i(\Sigma)$** [next]

Remark: More complete results for complex data due to

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

Harish-Chandra
Itzykson-Zuber

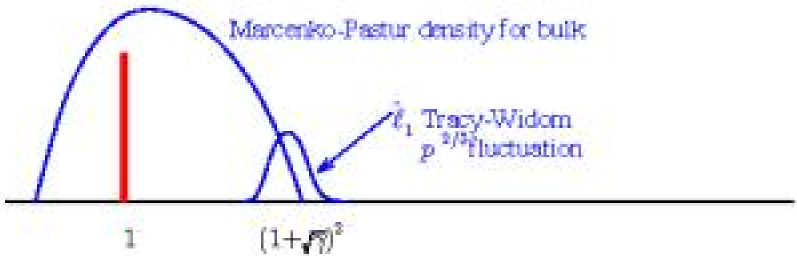
with $\sigma(\Sigma^{-1}) = \mathbf{diag}(\pi_j); \quad L = \mathbf{diag}(l_k); \quad V(l) = \prod_{j<k}(l_j - l_k).$

Finite rank perturbations: heuristics

"Spiked" model: $\Sigma = \mathbf{diag}(\ell_1, \ldots, \ell_M, 1, \ldots, 1)$

$\ell_1 \geq \ell_2 \geq \cdots \geq \ell_M \geq 1,$ M **fixed as** $p \nearrow \infty,\ p/n \to \gamma.$

$\Sigma = I$

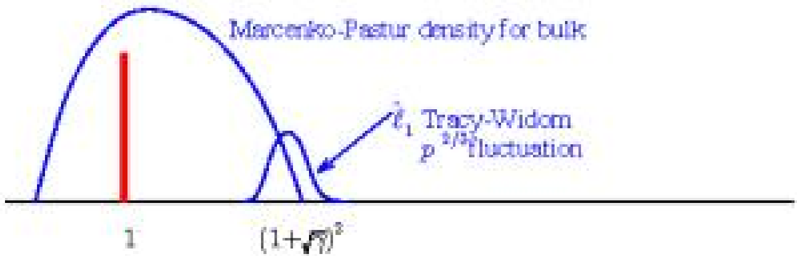


Finite rank perturbations: heuristics

"Spiked" model: $\Sigma = \textbf{diag}(\ell_1, \ldots, \ell_M, 1, \ldots, 1)$

$\ell_1 \geq \ell_2 \geq \cdots \geq \ell_M \geq 1,$ $\quad M$ **fixed as** $p \nearrow \infty, \; p/n \to \gamma.$

$\Sigma = I$



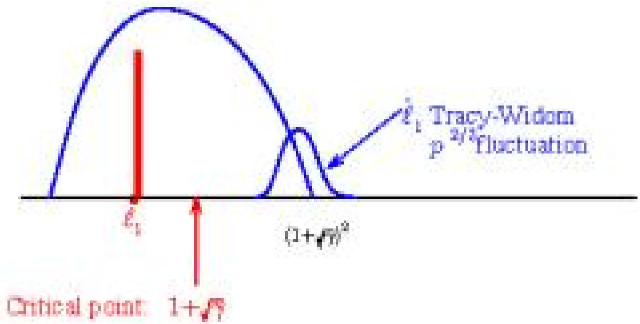
$\ell_1 \gg 1$

$M = 1$

Finite rank model: phase transition

$\Sigma =$ **diag**$(\ell_1, \ldots, \ell_M, 1, \ldots, 1) \qquad p/n \to \gamma.$ Baik-Ben Arous-Peche, Paul,

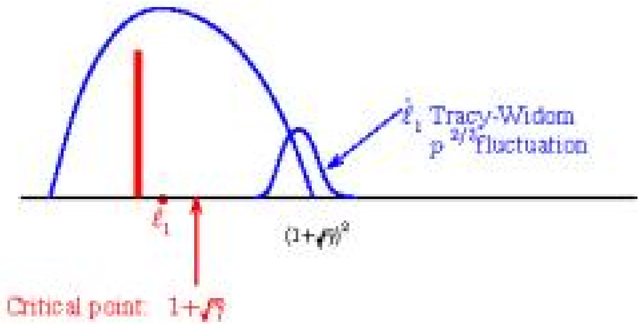
Interior point transition at $\ell_1 = 1 + \sqrt{\gamma}$: Baik-Silverstein



Finite rank model: phase transition

$\Sigma =$ **diag**$(\ell_1, \ldots, \ell_M, 1, \ldots, 1) \qquad p/n \to \gamma$. Baik-Ben Arous-Peche, Paul,

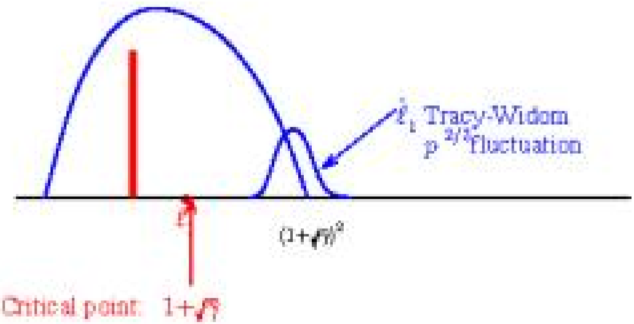
Interior point transition at $\ell_1 = 1 + \sqrt{\gamma}$: Baik-Silverstein



Finite rank model: phase transition

$\Sigma =$ **diag**$(\ell_1, \ldots, \ell_M, 1, \ldots, 1) \qquad p/n \to \gamma$. Baik-Ben Arous-Peche, Paul,

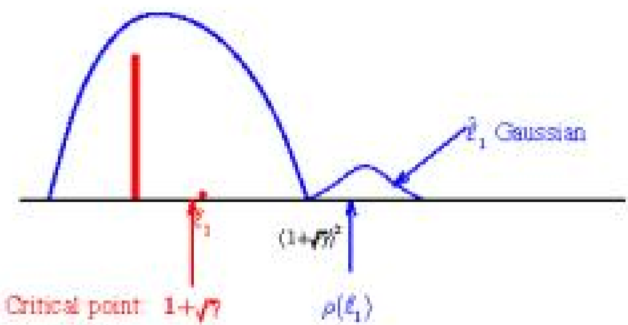
Interior point transition at $\ell_1 = 1 + \sqrt{\gamma}$: Baik-Silverstein



Finite rank model: phase transition

$\Sigma =$ **diag**$(\ell_1, \ldots, \ell_M, 1, \ldots, 1)$ $\quad p/n \to \gamma$. Baik-Ben Arous-Peche, Paul,

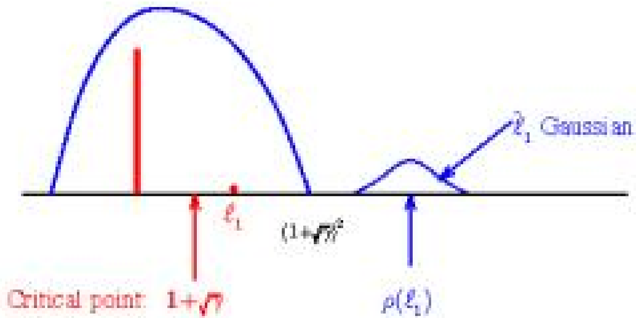
Interior point transition at $\ell_1 = 1 + \sqrt{\gamma}$: Baik-Silverstein



Finite rank model: phase transition

$\Sigma =$ **diag**$(\ell_1, \ldots, \ell_M, 1, \ldots, 1)$ $\quad p/n \to \gamma$. Baik-Ben Arous-Peche, Paul,

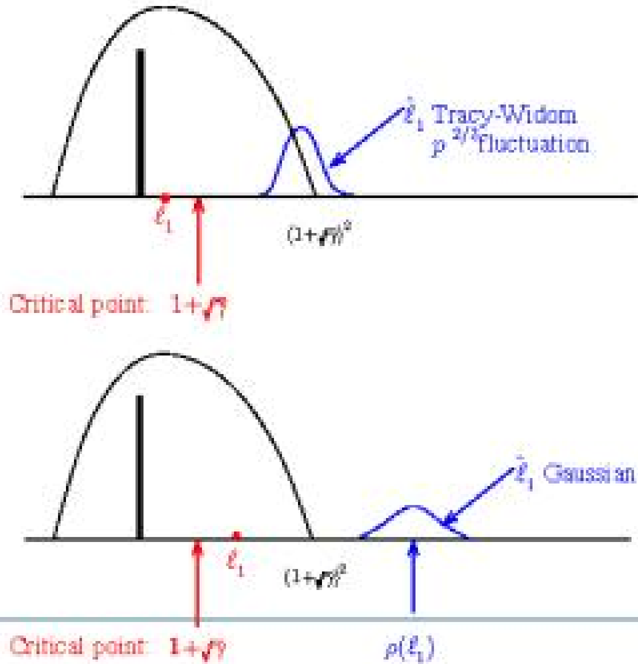
Interior point transition at $\ell_1 = 1 + \sqrt{\gamma}$: Baik-Silverstein



Finite rank model: phase transition

$\Sigma = \textbf{diag}(\ell_1, \ldots, \ell_M, 1, \ldots, 1) \qquad p/n \to \gamma.$ Baik-Ben Arous-Peche, Paul,

Interior point transition at $\ell_1 = 1 + \sqrt{\gamma}$: Baik-Silverstein



Outline

1. Principal Components Analysis (PCA)
2. Gaussian & Wishart Distributions
3. Random matrices
4. Large p Asymptotics
5. Largest Eigenvalue Laws
6. Beyond the "Null Hypothesis"
7. Estimating Eigenvectors

Thanks: Persi Diaconis, Peter Forrester, Matthew Harding, Plamen Koev, Arno Kuijlaars, Craig Tracy, Maarten Vanlessen,
Support: NSF, NIH

Recent example: economics

How many factors are present in security returns? Use PCA??

S.J. Brown (1989) simulations, calibrated to NYSE data

4 factor model* $\rightarrow \Sigma =$ diag$(\ell_1, \ldots, \ell_4, \sigma_e^2, \ldots, \sigma_e^2)$

$$\ell_1 > \ell_2 = \ell_3 = \ell_4 > \sigma_e^2$$

Goal: Use PCA to estimate $\ell_1, \ldots, \ell_4$.

Empirical puzzle (Brown, 1989):

many sample eigenvalues swamp ℓ_2, ℓ_3, ℓ_4.

(*) $R_{it} = \Sigma_{k=1}^4 b_{ik} f_{kt} + e_{it}; \quad i = 1, \ldots, p$ securities; $\quad t = 1, \ldots, T$ times.

$b_{ik} \sim N(\beta, \sigma_b^2); \quad f_{kt} \sim N(0, \sigma_f^2); \quad e_{it} \sim N(0, \sigma_e^2)$ all independent

Recent example: economics

How many factors are present in security returns? Use PCA??

S.J. Brown (1989) simulations, calibrated to NYSE data

4 factor model* $\rightarrow \Sigma = \mathbf{diag}(\ell_1, \ldots, \ell_4, \sigma_e^2, \ldots, \sigma_e^2)$

$$\ell_1 > \ell_2 = \ell_3 = \ell_4 > \sigma_e^2$$

Goal: Use PCA to estimate $\ell_1, \ldots, \ell_4$.

Empirical puzzle (Brown, 1989):

many sample eigenvalues swamp ℓ_2, ℓ_3, ℓ_4.

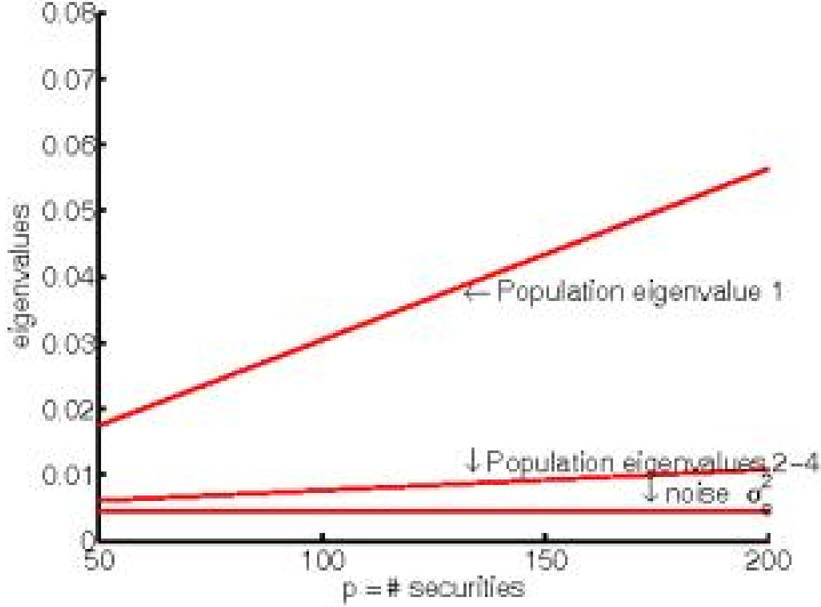
Explanation (Harding, 2006):

ℓ_2, ℓ_3, ℓ_4 are below the $1 + \sqrt{\gamma}$ phase transition.

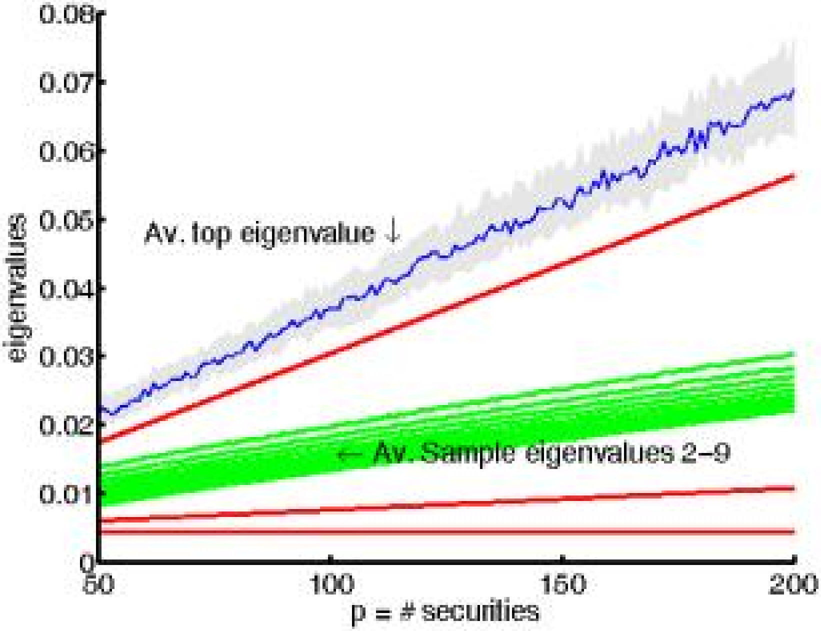
(*) $R_{it} = \Sigma_{k=1}^{4} b_{ik} f_{kt} + e_{it}; \quad i = 1, \ldots, p$ securities; $\quad t = 1, \ldots, T$ times.

$b_{ik} \sim N(\beta, \sigma_b^2); \quad f_{kt} \sim N(0, \sigma_f^2); \quad e_{it} \sim N(0, \sigma_e^2)$ all independent

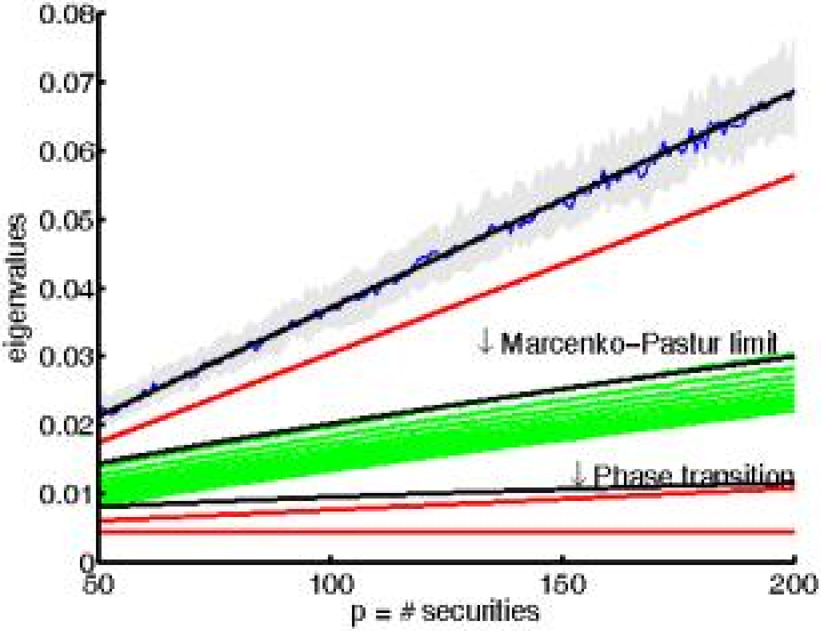
Population values



Brown(1989) plot



Marcenko-Pastur & phase transition



Source: Harding(2006).

Outline

1. Principal Components Analysis
2. Gaussian & Wishart Distributions
3. Random matrices

4. Large p Asymptotics
5. Largest Eigenvalue Laws
6. Beyond the "Null Hypothesis"
7. Estimating Eigenvectors

Estimation of Eigenvectors

$S \sim W_p(n, \Sigma), \qquad \Sigma = \sigma^2 I + \sum_{\nu=1}^{M} \lambda_\nu \theta_\nu \theta_\nu^T$

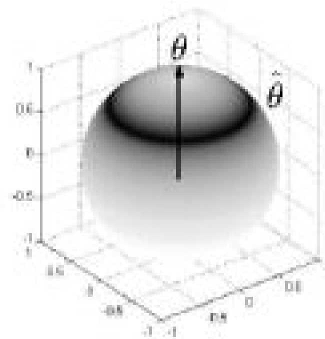
Estimation and inference for θ_ν??

Classical: p **fixed,** n **large:** $\sqrt{n}(\hat{\theta}_\nu - \theta_\nu) \to N_p(0, \Gamma_\nu)$

BUT: inconsistency when $p/n \to \gamma > 0$:

Reimann, v.d.Broeck, Bcx, Hoyle, Rattray, Paul, Baik, Silverstein

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



[Signal processing literature]

Eigenvectors: Elements of an Estimation Theory

- **Assume** $\exists$ a basis with **sparse** representation:

$$\theta \in \Theta_q(C): \quad \text{e.g.} \quad |\theta_{\nu,(\mu)}| \le C|\mu|^{-1/q} \qquad q < 2$$

- **Approximate** by "signal in Gaussian noise" model

LEMMA ($M = 1$) Let $\hat{C} = \langle \hat{\theta}, \theta \rangle$ and $\hat{\theta}^{\perp} = \hat{\theta} - \hat{C}\theta$. Then

$$\hat{\theta} = \hat{C}\theta + \hat{S}U \qquad \text{(Paul)}$$

  - $U = \hat{\theta}^{\perp}/\|\hat{\theta}^{\perp}\|$ is **uniform** on "S^{p-2}" $\Rightarrow$ **nearly Gaussian**.
  - move from **eigenvectors** to **sparse mean** estimation.

$\Longrightarrow$ near sharp upper & lower bounds for **minimax risk**:

$$\inf_{\hat{\theta}} \sup_{\theta_\nu \in \Theta_q(C)} E\|\hat{\theta}_\nu - \theta_\nu\|^2.$$

Finale

- Role of eigenstructure in high-dimensions will grow in statistics
  - Many topics not covered
- Benefit from many areas of math & physics
  - Role for both rigorous and non-rigorous results

THANK YOU

INTERNATIONAL CONGRESS OF MATHEMATICIANS MADRID 2006

Abstract PCA

Variables $X_1, \ldots, X_p$; Reduce dimension using

$$\Sigma = (\sigma_{kk'}) = \text{Cov}(X_k, X_{k'}) \quad = E(X_k - \mu_k)(X_{k'} - \mu_{k'})$$

Derived variable: $W = \sum_k v_k X_k$ has

$$\text{Var}(W) = \sum_{k,k'} v_k \sigma_{k,k'} v_{k'} = \mathbf{v}^T \Sigma \mathbf{v}$$

Successive maximization of variance

$$\ell_j = \max\{\mathbf{v}^T \Sigma \mathbf{v} \ : \ \mathbf{v}^T \mathbf{v}_{j'} = 0; j' < j, \ |\mathbf{v}| = 1\}$$

$\rightarrow$ **principal component eigenvalues** ℓ_j and **p.c. eigenvectors** $\mathbf{v}_j$:

Hope: small j capture most variance ($\ell_1 \geq \ell_2 \geq \ldots$).

Observed Data and Estimation

- Σ and hence $(\ell_j, \mathbf{v}_j)$ are **unknown**.
- **observe data** $\mathbf{x}_k \in \mathbb{R}^n$ on each variable X_k:
- $\rightarrow$ data matrix $\underset{n\times p}{X} = (x_{ik}) = [\mathbf{x}_1 \ldots \mathbf{x}_p]$.
- n **observations** on each of p **variables**.

$$\mathbf{x}_k = \begin{pmatrix} x_{1k} \\ \vdots \\ x_{nk} \end{pmatrix}$$

Observed Data and Estimation

- Σ and hence $(\ell_j, \mathbf{v}_j)$ are **unknown.**
- **observe data** $\mathbf{x}_k \in \mathbb{R}^n$ on each variable X_k:
- $\rightarrow$ data matrix $\underset{n\times p}{X} = (x_{ik}) = [\mathbf{x}_1 \ldots \mathbf{x}_p]$.
- n **observations** on each of p **variables.**

$$\mathbf{x}_k = \begin{pmatrix} x_{1k} \\ \vdots \\ x_{nk} \end{pmatrix}$$

Example: Allele frequencies (after Cavalli-Sforza, Menozzi, Piazza, 1978 ff)

$p = 38$ genes

		A	*Rh*	*HLA*	$\cdots$
	Bilbao	.4	.7	.8	
$n = 400$ locations	*Helsinki*	.2	.5	.3	
	Budapest..	.3	.3	.5	

$X \ (n \times p)$

Centering: Subtract means: $x_{ik} \leftarrow x_{ik} - \bar{x}_k, \qquad \bar{x}_k = n^{-1} \sum_i x_{ik}$

PCA on Observed Data - 1

	A	Rh	HLA	...
Bilbao	.1	.2	.3	
Helsinki	-.1	0	-.2	
Budapest ..	0	-.2	-.05	

Assume mean centered $X = (x_{ik})$

Sample Covariance Matrix $S = (s_{kk'}) = (n^{-1} \sum_i x_{ik} x_{ik'})$

$$S = n^{-1} X^T X$$

Derived variable: $\mathbf{w} = X\mathbf{v} = \sum_k v_k \mathbf{x}_k$

$$\widehat{\text{Var}}(\mathbf{w}) = \mathbf{v}^T S \mathbf{v}.$$

Directions of maximum variance:

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