Random matrices in statistics: testing in spiked models

Iain Johnstone

Statistics, Stanford

AustMS, Sydney, 1st October 2013
International Year of Statistics - 2013

www.statistics2013.org

- to promote public awareness of Statistics
- steered by professional societies [ASA, RSS, ISI, IBS, ...]
- ≥ 2000 participating organizations
  - 25 Australian sponsors at
    statistics2013.org/iyos/participants_countrysearch.cfm
- Videos, newsletters, website, events, conferences,
Outline

- Random matrices and covariances
  - Principal Components Analysis, examples

- Spiked covariance model
  - Phase transition
  - Weak signals
  - Strong signals
Random Matrix Theory (RMT)

Eigenvalues and vectors of large square random matrices:

\[ \mathbf{A} \mathbf{v}_j = \mu_j \mathbf{v}_j \quad \mathbf{A} = (A_{ij}) \quad n \times n. \]

Structured randomness:

- \( A_{ij} \) i.i.d. Hermitian, or [Wigner matrix]
- \( \mathbf{A} \) invariant for \( O(n), U(n) \) [GOE, GUE]

Interest in properties of eigenvalues:

- empirical distribution: \( F_n(x) = n^{-1} \# \{ i : \lambda_i \leq x \} \)
- extremes: \( \lambda_{(1)} = \max \lambda_j \)
- spacings ...
RMT: ‘Wishart’ case

Consider \( X = (X_{ij}) \quad n \times p \quad \text{rectangular, i.i.d entries} \)

Study eigenvalues \( \lambda_j \) of \( X^T X \)

Accessible because \( \lambda_j = \mu_j^2 \), with \( \{\mu_j\} \) eigenvalues of

\[
A = \begin{pmatrix}
0 & X \\
X^T & 0
\end{pmatrix}
\]

Link to statistics:
- \( n^{-1}X^T X \) is (simple form) of covariance matrix.
- goal: helpful approximations based on \( p/n \to \gamma > 0 \)
Covariance Matrices - Population

\[ X^T = (X, j) \in \mathbb{R}^p, \]  a random (row) vector distributed as \( P \).

Population mean:
\[ \mu = \mathbb{E}_P X \]

Pop. covariance matrix:
\[ \Sigma = \mathbb{E}(X - \mu)(X - \mu)^T \]
\[ = \mathbb{E}XX^T \quad \text{if } \mu = 0. \]
\[ \Sigma_{jj'} = \text{Cov}(X, j, X, j') \]

In general \( \Sigma (p \times p) \) has \( O(p^2/2) \) parameters. Too many!

Simpler models:
\[ \Sigma = \sigma^2 I_p \quad \text{‘white’} \]
\[ \Sigma = \sigma^2(I_p + \sum_{1}^{M} h_\nu v_\nu v_\nu^T) \quad \text{low rank (‘spiked’)} \]
Covariance Matrices - Sample

Data: \( X_1^T, \ldots, X_n^T \in \mathbb{R}^p \) (or \( \mathbb{C}^p \))

assumed to be independent draws from \( X^T \sim \mathcal{P}_\Sigma \)

Sample covariance matrix:

\[
S = n^{-1} \sum_{i=1}^{n} X_iX_i^T
\]

Use observed \( S \) to estimate or test unknown \( \Sigma \).

E.g. \( H_0: \Sigma = I \) “null” hypothesis

\( H_A: \Sigma = I + hvv^T \) “alternative” hypothesis

Link to RMT: \( nS = X^TX \) using the \( n \times p \) data matrix

\[
X = (X_{i,j}) = \begin{bmatrix}
X_1^T \\
\vdots \\
X_n^T
\end{bmatrix}
\]
Principal Components Analysis

Statistical interpretation of eigenstructure: \( Sv_j = \lambda_j v_j \).

Goal: reduce dimensionality of data from \( p \) (large) to \( k \) (small):

\[
2Xv = 2; Z_2 = Xv_2
\]
\[
1Xv = 1; Z_1 = Xv_1
\]

Interpret as directions \( v_j \) of maximum variance, with variances

\[
\lambda_j = \max\{v^T S v : v^T v_j', \|v\| = 1\}
\]

= ”principal component variances”
Outline

- Random matrices and covariances
  - Principal Components Analysis, two examples: genetics, finance

- Spiked covariance model
  - Phase transition
  - Weak signals
  - Strong signals
Example 1: PCA & population structure from genetic data

Gene ($Y$) vs. Phenotype ($X$) shows apparent correlation, but ...
Example 1: PCA & population structure from genetic data

Gene ($Y$) vs. Phenotype ($X$) shows apparent correlation, but ... 3 subpopulations — **Within** each population, no correlation exists!
Example 1: PCA & population structure from genetic data

Patterson et. al. (2006), Price et. al. (2006)

\[ n = \#\text{individuals}, \quad p = \#\text{markers (e.g. SNPs)} \]

\[ X_{ij} = \text{(normalized) allele count}, \quad \text{case } i = 1, \ldots, n, \quad \text{marker } j = 1, \ldots, p. \]

\[ H = n \times \text{sample covariance matrix of } X_{ij} \]

- Eigenvalues of \( H \): \( \lambda_1 > \lambda_2 > \ldots > \lambda_{\min(n,p)} \)
- How many \( \lambda_i \) are significant?
- Under \( H_0 \), distribution of \( \lambda_1 \) if \( H \sim W_p(n, I) \)?
Example 1: PCA & population structure from genetic data

- PPR (2006) example: 3 African populations, $n = 67, p = 993$
- Tracy-Widom theory $\implies$ 2 “significant” eigenvectors, separates populations
Example 2: finance

Arbitrage Pricing Theory → a few factors “explain” returns

\[ R = \sum_{\nu=1}^{M} b_{\nu} f_{\nu} + e \]
Example 2: finance

Arbitrage Pricing Theory → a few factors “explain” returns
Given $j = 1, \ldots, p$ securities, $t = 1, \ldots, T$ observation times,
(and $M = 1$),

$$R_{jt} = b_{j1} f_{1t} + e_{jt}.$$
Example 2: finance

Arbitrage Pricing Theory → a few factors “explain” returns.
Given $j = 1, \ldots, p$ securities, $t = 1, \ldots, T$ observation times,

$$R_{jt} = \sum_{\nu=1}^{M} b_{j\nu} f_{\nu t} + e_{jt}.$$
Example 2: finance

Arbitrage Pricing Theory → a few factors “explain” returns.
Given $j = 1, \ldots, p$ securities, $t = 1, \ldots, T$ observation times,

$$R_{jt} = \sum_{\nu=1}^{M} b_{j\nu} f_{\nu t} + e_{jt}.$$ 

Under Gaussian assumptions*, $\Sigma = \text{Cov} (R)$ has eigenvalues

$$(\ell_1 > \ell_2 = \cdots = \ell_M > \sigma_e^2, \ldots, \sigma_e^2).$$

$$(*) \quad b_{j\nu} \sim N(\beta, \sigma_b^2); \quad f_{\nu t} \sim N(0, \sigma_f^2); \quad e_{jt} \sim N(0, \sigma_e^2) \quad \text{all independent}$$
Example 2: finance

Arbitrage Pricing Theory → a few factors “explain” returns.

Given \( j = 1, \ldots, p \) securities, \( t = 1, \ldots, T \) observation times,

\[
R_{jt} = \sum_{\nu=1}^{M} b_{j\nu} f_{\nu t} + e_{jt}.
\]

Under Gaussian assumptions*, \( \Sigma = \text{Cov}(R) \) has eigenvalues

\[
(\ell_1 > \ell_2 = \cdots = \ell_M > \sigma_e^2, \ldots, \sigma_e^2).
\]

Form *sample* covariance matrix \( S \)

\[
S_{jk} = T^{-1} \sum_t (R_{jt} - \bar{R}_j)(R_{kt} - \bar{R}_k).
\]

Use (largest) sample eigenvalues \( \hat{\ell}_i(S) \) to estimate \( \ell_i \).

\[(*) \quad b_{j\nu} \sim N(\beta, \sigma_b^2); \quad f_{\nu t} \sim N(0, \sigma_f^2); \quad e_{jt} \sim N(0, \sigma_e^2) \quad \text{all independent}\]
Example 2: finance

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

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

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

Use $\hat{\ell}_i(S)$ to estimate $\ell_1, \ldots, \ell_4$. 

Empirical puzzle (Brown, 1989): many sample eigenvalues swamp $\ell_2, \ell_3, \ell_4$. 

Illustration: vary $p = 50(1)200$ ($T = 80$) 

Plot theoretical $\ell_i(p)$ and simulated $\hat{\ell}_i(p)$ versus $p$. 

Example 2: finance

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

4 factor model → \( \Sigma = \text{diag}(\ell_1, \ldots, \ell_4, \sigma_e^2, \ldots, \sigma_e^2) \)

\[ \ell_1 > \ell_2 = \ell_3 = \ell_4 > \sigma_e^2 \]

Use \( \hat{\ell}_i(S) \) to estimate \( \ell_1, \ldots, \ell_4 \).

Empirical puzzle (Brown, 1989):

*many sample* eigenvalues swamp \( \ell_2, \ell_3, \ell_4 \).

- Illustration: vary \( p = 50(1)200 \quad (T = 80) \)
- Plot theoretical \( \ell_i(p) \) and simulated \( \hat{\ell}_i(p) \) versus \( p \).
Example 2: theoretical $\ell_i(p)$ values
Example 2: Brown (1989) plot
Example 2: finance

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

4 factor model* → \( \Sigma = \text{diag}(\ell_1, \ldots, \ell_4, \sigma_e^2, \ldots, \sigma_e^2) \)

\( \ell_1 > \ell_2 = \ell_3 = \ell_4 > \sigma_e^2 \)

Use \( \hat{\ell}_i(S) \) to estimate \( \ell_1, \ldots, \ell_4 \).

Empirical puzzle (Brown, 1989): 

- many sample eigenvalues swamp \( \ell_2, \ell_3, \ell_4 \).

  ▶ Illustration: vary \( p = 50(1)200 \)

  ▶ Plot theoretical \( \ell_i(p) \) and simulated \( \hat{\ell}_i(p) \) versus \( p \).

Explanation (Harding, 2008):

- \( \ell_2, \ell_3, \ell_4 \) are below a phase transition predicted by RMT.
Outline

- Random matrices and covariances
  - Principal Components Analysis, two examples
- Spiked covariance model
  - Phase transition
  - Weak signals
  - Strong signals
Spiked Covariance Model

- $n$ (independent) observations on $p$-vectors: $X_i$
- correlation structure is “white + low rank”:

$$\Sigma = \text{Cov}(X_i) = \sigma^2 I + \sum_{\nu=1}^{M} h_\nu v_\nu v_\nu^T$$

Interest in

- testing/estimating $h_\nu$ [today]
- determining $M$
- estimating $v_\nu$
Some motivating models

1. Economics: \( X_i \) = vector of stocks (indices) at time \( i \)
   \( \mathbf{v}_\nu \) = factor loadings, \( f_{\nu i} \) factors, \( Z_i \) idiosyncratic terms.

2. ECG: \( X_i \) = \( i \)th heartbeat (\( p \) samples per cycle)
   \( \mathbf{v}_\nu \) = may be sparse in wavelet basis.

3. Microarrays: \( X_i \) = expression of \( p \) genes in \( i \)th patient.
   \( \mathbf{v}_\nu \) = may be sparse few genes involved in each factor.

4. Genetics: \( X_i \) = allele count at \( p \) SNPs in \( i \)th individual.

5. Sensors: \( X_i \) = observations at sensors
   \( \mathbf{v}_\nu \) = cols. of steering matrix, \( f_{\nu i} \) signals

6. Climate: \( X_i \) = measurements from global network at time \( i \)
   \( \mathbf{v}_\nu \) = (empirical) orthogonal functions (EOF)
Outline

- Spiked Covariance model
  - Examples
    \[ \Sigma = I + hvv^T \]

- Wishart eigenvalues and Phase Transition
  \[ p/n \to \gamma \]
  \[ 0 \quad \sqrt{\gamma} \quad h \]

- Weak Signals
  - Contiguity

- Strong Signals
  - Approximations to Power
$p$ and $n$ and all that

$p = \# \text{ variables/parameters}$

$n = \# \text{ of (independent?) observations}$

- $p = o(n)$ classical statistics
- $n = o(p)$ (nominally) high-dimensional data, sparsity

This talk:
$p$ and $n$ and all that

$p = \# \text{ variables/parameters}$
$n = \# \text{ of (independent?) observations}$

- $p = o(n)$ classical statistics
- $n = o(p)$ (nominally) high-dimensional data, sparsity

This talk:
- $p/n \to \gamma > 0$ less ambitious; important phenomena appear
  - for e.g. $p = 5$, $n = 20$, this limit may yield better approximation than $p$ fixed, $n$ large.
\( p \) and \( n \) and all that

\[
p = \# \text{ variables/parameters} \\
n = \# \text{ of (independent?) observations}
\]

- \( p = o(n) \) classical statistics
- \( n = o(p) \) (nominally) high-dimensional data, sparsity

This talk:
- \( p/n \to \gamma > 0 \) less ambitious; important phenomena appear
  - for e.g. \( p = 5, n = 20 \), this limit may yield better approximation than \( p \) fixed, \( n \) large.
- \( p, n \) fixed (strong signal asymptotics).
Wishart Distribution

\[ \mathbf{X} = (X_{i,j}) \quad n \times p \]

Rows \( \mathbf{X}_i^T = (X_{i,j}) \) \( \text{indep} \sim \mathcal{N}_p(\mu, \Sigma) \)

**Definition:** sample covariance, unnormalized:

\[ H = \mathbf{X}^T \mathbf{X} \sim \mathcal{W}_p(n, \Sigma) \quad \text{if } \mu = 0 \]

\( p \) variables, \( n \) degrees of freedom. \( \text{“Null case:” } \Sigma = I \)

Eigenvalues of \( H \): \( \lambda_1 > \lambda_2 > \cdots > \lambda_{n \wedge p} \geq 0 \)

J. Wishart 1898-1956
Wishart Eigenvalues, Null case

2 draws of eigenvalues from $W_{15}(60, I)$

- Spreading of sample eigenvalues from 4 to [1,9].

2 draws of 15 independent $U(1, 9)$ variates – very different!
The Quarter Circle Law

Description of spreading phenomenon in null case:

Marčenko-Pastur, (67) For $H \sim W_p(n, I)$ \quad $p/n \to \gamma \leq 1$

Empirical distribution function: for eigenvalues $\{n\lambda_j\}_{j=1}^p$ of $H$,

$$F_p(x) = p^{-1} \#\{\lambda_j \leq x\} \to F(x) = f(x)dx.$$ 

For $\Sigma = I$,

$$f^{MP}(x) = \frac{1}{2\pi\gamma x} \sqrt{(b_+ - x)(x - b_-)},$$

$$b_{\pm} = (1 \pm \sqrt{\gamma})^2.$$
Largest eigenvalue: Null case

Square root singularity:

\[ f^{MP}(x) \sim c \sqrt{b_+ - x}, \quad x \rightarrow b_+ \]

Heuristically,

\[ \lambda_1 - b_+ = O_p(n^{-2/3}), \]

and

\[
\frac{n^{2/3} \gamma^{1/6}}{(1 + \sqrt{\gamma})^{4/3}} (\lambda_1 - b_+) \overset{D}{\Rightarrow} TW_\beta,
\]

the \emph{Tracy-Widom} distributions [\( \beta = 1 \) for \( \mathbb{R} \), \( \beta = 2 \) for \( \mathbb{C} \).]
Largest eigenvalue: Non-null cases

Rank 1 for simplicity: $\Sigma = I + hvv^T$

For $0 \leq h < \sqrt{\gamma}$,

$$\frac{n^{2/3}\gamma^{1/6}}{(1 + \sqrt{\gamma})^{4/3}} (\lambda_1 - (1 + \sqrt{\gamma})^2) \xrightarrow{\mathcal{D}} TW_\beta,$$

Limit does not depend on $h$.

“Fundamental asymptotic limit of sample eigenvalue based detection” (?)

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\mathbb{R}$</th>
<th>$\mathbb{C}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 0$</td>
<td>$J$ (01)</td>
<td>Johannson (00)</td>
</tr>
<tr>
<td>$h \in (0, \sqrt{\gamma})$</td>
<td>Féral-Péché (09)</td>
<td>Baik-Ben Arous-Péché (05)</td>
</tr>
</tbody>
</table>
Largest eigenvalue: Phase transition

Different rates, limit distributions:

For $h < \sqrt{\gamma}$: $n^{2/3} \left[ \frac{\lambda_1 - \mu(\gamma)}{\sigma(\gamma)} \right] \overset{D}{\Rightarrow} TW_\beta,$

For $h > \sqrt{\gamma}$: $n^{1/2} \left[ \frac{\lambda_1 - \rho(h, \gamma)}{\tau(h, \gamma)} \right] \overset{D}{\Rightarrow} N(0, 1)$
Largest eigenvalue: Phase transition

Different rates, limit distributions:

For $h < \sqrt{\gamma}$:

$$n^{2/3} \left[ \frac{\lambda_1 - \mu(\gamma)}{\sigma(\gamma)} \right] \xrightarrow{D} TW_\beta,$$

For $h > \sqrt{\gamma}$:

$$n^{1/2} \left[ \frac{\lambda_1 - \rho(h, \gamma)}{\tau(h, \gamma)} \right] \xrightarrow{D} N(0, 1)$$

with

$$\rho(h, \gamma) = (1 + h) \left( 1 + \frac{\gamma}{h} \right)$$
$$\tau^2(h, \gamma) = 2(1 + h)^2 \left( 1 - \frac{\gamma}{h^2} \right)$$

Statistical physics lit, 94-Baik-Ben Arous-Peche(05)
Example: finance

How many factors are present in security returns? Use PCA??
S.J. Brown (1989) simulations, calibrated to NYSE data

4 factor model* → \( \Sigma = \text{diag}(\ell_1, \ldots, \ell_4, \sigma_e^2, \ldots, \sigma_e^2) \)
\[
\ell_1 > \ell_2 = \ell_3 = \ell_4 > \sigma_e^2
\]

Goal: Use \( \hat{\ell}_i(S) \) to estimate \( \ell_1, \ldots, \ell_4 \).

Empirical puzzle (Brown, 1989):
many sample eigenvalues swamp \( \ell_2, \ell_3, \ell_4 \).
Example: finance

How many factors are present in security returns? Use PCA??
S.J. Brown (1989) simulations, calibrated to NYSE data

4 factor model* → \( \Sigma = \text{diag}(\ell_1, \ldots, \ell_4, \sigma_e^2, \ldots, \sigma_e^2) \)
\[
\ell_1 > \ell_2 = \ell_3 = \ell_4 > \sigma_e^2
\]

Goal: Use \( \hat{\ell}_i(S) \) to estimate \( \ell_1, \ldots, \ell_4 \).

Empirical puzzle (Brown, 1989):
many sample eigenvalues swamp \( \ell_2, \ell_3, \ell_4 \).

Explanation (Harding, 2008):
\( \ell_2, \ell_3, \ell_4 \) are below the \( 1 + \sqrt{\gamma} \) phase transition.
Brown(1989) plot

Marcenko-Pastur & phase transition

Outline

- Spiked Covariance model
  - Examples
    \[ \Sigma = I + h \mathbf{v}\mathbf{v}^T \]

- Phase Transition

- Weak Signals
  - Contiguity
    \[ \frac{p}{n} \to \gamma \]

- Strong Signals
  - Approximations to Power
Detecting weak signals?

For \( h < \sqrt{\gamma} \), distribution of largest eigenvalue

\[
\lambda_1 \approx \mu(\gamma) + n^{-2/3} \sigma(\gamma) TW_1
\]

does not depend on \( h \).

Onatski-Moreira-Hallin, AOS (2013):

- can detect \( h < \sqrt{\gamma} \), with error
- use all eigenvalues
- contiguity ideas yield limit distributions for \( h \in (0, \sqrt{\gamma}) \).
Likelihood Ratio Test

\[ X_i \sim N_p(0, I + h\mathbf{v}\mathbf{v}^T), \quad H_0 : h = 0 \text{ vs. } H_1 : h > 0, \mathbf{v} \text{ unspecified.} \]

Invariant under rotations, so consider

\[ p(\lambda; h) = \text{joint density of sample eigenvalues } \lambda = (\lambda_1, \ldots, \lambda_n). \]

Likelihood ratio test against fixed \( h > 0 \):

\[ L(\lambda; h) = \frac{p(\lambda; h)}{p(\lambda; 0)} \]
Likelihood Ratio Test

\[ X_i \sim N_p(0, I + h\mathbf{v}\mathbf{v}^T), \quad H_0 : h = 0 \text{ vs. } H_1 : h > 0, \, \mathbf{v} \text{ unspecified.} \]

Invariant under rotations, so consider

\[ p(\lambda; h) = \text{joint density of sample eigenvalues } \lambda = (\lambda_1, \ldots, \lambda_n). \]

\[
\frac{\gamma(n, p, \lambda)}{(1 + h)^{n/2}} \int_{S(p)} e^{\frac{n}{2(1+h)}x_p'\Lambda x_p} \,(dx_p)
\]

with \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p) \).

Likelihood ratio test against fixed \( h > 0 \):

\[
L(\lambda; h) = \frac{p(\lambda; h)}{p(\lambda; 0)} = \frac{1}{(1 + h)^{n/2}} \int_{S(p)} e^{\frac{n}{2(1+h)}x_p'\Lambda x_p} \,(dx_p)
\]
Asymptotic normality of likelihood ratio

Under $H_0$ ($h = 0$), for $0 \leq h \leq \bar{h} < \sqrt{\gamma}$, and $p/n \to \gamma$

$$\log L(h; \lambda) \Rightarrow \mathcal{L}(h; \lambda), \quad (O-M-H, 2013)$$

a Gaussian process [by Bai-Silverstein CLT in RMT], with
Asymptotic normality of likelihood ratio

Under $H_0$ ($h = 0$), for $0 \leq h \leq \bar{h} < \sqrt{\gamma}$, and $p/n \to \gamma$

\[
\log L(h; \lambda) \Rightarrow \mathcal{L}(h; \lambda), \quad (O-M-H, 2013)
\]
a Gaussian process [by Bai-Silverstein CLT in RMT], with

\[
E\mathcal{L}(h; \lambda) = \frac{1}{4} \log \left(1 - \frac{h^2}{\gamma}\right)
\]

\[
\text{Cov}\{\mathcal{L}(h_1; \lambda), \mathcal{L}(h_2; \lambda)\} = -\frac{1}{2} \log \left(1 - \frac{h_1 h_2}{\gamma}\right)
\]

[ if $h \geq H > \sqrt{\gamma}$, $L(h; \lambda) = O_p(e^{-n\delta})$. ]
Detecting weak signals

Reparametrize: \[ \theta = \sqrt{-\log(1 - h^2/\gamma)} \quad \text{for} \ h < \sqrt{\gamma}. \]

Seek optimal test of \( H_0 : \theta = 0 \) vs. \( H_A : \theta = \theta_1 > 0 \)

Recap: Likelihood ratio: \[ L_n(\lambda; \theta_1) = \frac{p(\lambda; \theta_1)}{p(\lambda; 0)} \] satisfies

\[
\begin{align*}
\log L_n & \xrightarrow{P_{n,0}} N(-\theta_1^2/4, \theta_1^2/2) \\
\xrightarrow{P_{n,\theta_1}} & N(+\theta_1^2/4, \theta_1^2/2) \quad \text{(Contiguity!)}
\end{align*}
\]

So for asymptotically optimal test,

Reject \[ \iff \log L_n > C_{n,\alpha} = \frac{\theta_1 z_\alpha}{\sqrt{2} - \theta_1^2/4} \]
Asymptotic Power

Compute ‘Power’ \( \beta(\theta_1) = P_{\theta=\theta_1}(\text{Reject}) = \lim P_{n,\theta_1}(\log L_n > C_{n,\alpha}) \)

if \( P_{\theta=0}(\text{Reject}) = \alpha \)
Asymptotic Power

Fig 5. Asymptotic powers \( \beta_{J}(\theta_1), \beta_{LW}(\theta_1), \beta_{CLR}(\theta_1) \) of the tests described in Examples 1 (John), 2 (Ledoit and Wolf), and 3 (Baie et al.).

The asymptotic power functions of the tests from Examples 1, 2, and 3 are non-trivial. Figure 5 compares these power functions to the corresponding power envelopes. Since John's test is invariant with respect to orthogonal transformations and scalings, \( \beta_{J}(\theta_1) \) is compared to the power envelope \( \beta(\theta_1; \mu) \). The asymptotic power functions \( \beta_{LW}(\theta_1) \) and \( \beta_{CLR}(\theta_1) \) are compared to the power envelope \( \beta(\theta_1; \lambda) \) because the Ledoit-Wolf test of \( \Sigma = I \) and the "corrected" likelihood ratio test are invariant only with respect to orthogonal transformations.

Interestingly, whereas \( \beta_{J}(\theta_1) \) and \( \beta_{LW}(\theta_1) \) depend only on \( \alpha \) and \( \theta_1 \), \( \beta_{CLR}(\theta_1) \) depends also on \( c \). As \( c \) converges to one, \( \beta_{CLR}(\theta_1) \) converges to \( \alpha \), which corresponds to the case of trivial power. As \( c \) converges to zero, \( \beta_{CLR}(\theta_1) \) converges to \( \beta_{J}(\theta_1) \). In Figure 5, we provide the plot of \( \beta_{CLR}(\theta_1) \) that corresponds to \( c = 0 \).

The left panel of Figure 5 shows that the power function of John's test is very close to the power envelope \( \beta(\theta_1; \mu) \) in the vicinity of \( \theta_1 = 0 \). Such behavior is consistent with the fact that John's test is locally most powerful invariant. However, for large \( \theta_1 \), the asymptotic power functions of all the tests from Examples 1, 2, and 3 are lower than the corresponding asymptotic power envelopes. We should stress here that these tests have power against general alternatives as opposed to the "spiked" alternatives that maintain the assumption that the population covariance matrix of data has the form \( \sigma^2 (I_p + hvv_0) \).

For the "spiked" alternatives, the \( \lambda \)-and \( \mu \)-based LR tests may be more

\[
LW = p^{-1} \text{tr}[(\hat{\Sigma} - I)^2] - \gamma_n[p^{-1} \text{tr} \hat{\Sigma}]^2 + \gamma_n, \quad \gamma_n = p/n, \quad \hat{\Sigma} = H/n \quad [\text{Ledoit-Wolf}]
\]

\[
CLR = \text{tr} \hat{\Sigma} - \log \det \hat{\Sigma} - p(1 - (1 - \gamma_n^{-1}) \log(1 - \gamma_n)) \quad [\text{Bai et. al.}]
\]
Recall that $\theta = \sqrt{-\log(1 - h^2/\gamma)}$ ...
In original parameter $h$, power is good only very close to $\sqrt{\gamma}$. 

According to Le Cam’s third lemma, under a specific alternative $\theta = \theta_1 \leq M$, the asymptotic distribution of the LR statistic equals the distribution of $2 \sup_{\theta \in (0, M]} \tilde{X}_\theta$, where $\tilde{X}_\theta$ is a Gaussian process with the same covariance function as that of $X_\theta$, but with a different mean: $E(\tilde{X}_\theta) = E(X_\theta) + \text{Cov}(X_\theta, X_{\theta_1})$.

Therefore, to numerically evaluate the asymptotic power function of the $\lambda$-based LR test, we simulate 500,000 observations of $X_\theta$ on a grid of 1,000 equally spaced points in $\theta \in [0, M]$, where $M = 6$ is chosen as the upper limit of the grid because it is large enough for the power envelopes to reach the value of 99%. For each observation, we save its supremum on the grid, and use the empirical distribution of two times the suprema as the approximate asymptotic distribution of the likelihood ratio statistic under the null. We denote this distribution as $\hat{F}_0$. Its 95% quantile equals 4.3982. For each $\theta_1$ on the grid, we repeat the simulation for process $\tilde{X}_\theta$ to obtain the approximate asymptotic distribution of the likelihood ratio statistic under the alternative $\theta = \theta_1$, which we denote as $\hat{F}_1$. We evaluate $\hat{F}_1$ at the 95% quantile of $\hat{F}_0$ as a numerical approximation to the asymptotic power at $\theta_1$ of the $\lambda$-based LR test with asymptotic size 0.05.
A dose of reality

In original parameter $h$, power is good only very close to $\sqrt{\gamma}$.

Figure 3. The maximal asymptotic power of the $\lambda$-based tests (dashed lines) and $\mu$-based tests (solid lines) of $\theta = 0$ against $\theta = \theta_1$. Left panel: $\theta$-parametrization. Right panel: $h$-parametrization.

.... “It is not done well; but you are surprised to find it done at all.”
Outline

- Spiked Covariance model
  - Examples \( \Sigma = I + hvv^T \)

- Wishart eigenvalues and Phase Transition

- Weak Signals
  - Contiguity

- Strong Signals
  - Approximations to Power [with Boaz Nadler]
Strong signals

\[ H \sim W_p(n, \sigma^2 I + h\mathbf{v}\mathbf{v}^T) \]

So far:  
- \( h < \sqrt{\gamma} \):  
  \[ \lambda_1(H) \sim \mu_{TW} + \sigma_{TW} TW / n^{2/3} \]
- \( h > \sqrt{\gamma} \):  
  \[ \lambda_1(H) \sim N(\mu_{h,\gamma}, \sigma^2_{h,\gamma} / n) \]

In strong signal regime:  
- \( h \gg \sqrt{\gamma} \),

\[ \log L(h; \lambda) = n^2 \left[ \lambda_1 h + 1 - \log(1 + h) \right] + o(1) \]
Strong signals

\[ H \sim W_p(n, \sigma^2 I + h\mathbf{v}\mathbf{v}^T) \]

So far: \( h < \sqrt{\gamma} \) : \( \lambda_1(H) \sim \mu_{TW} + \sigma_{TW} TW / n^{2/3} \)

\( h > \sqrt{\gamma} \) : \( \lambda_1(H) \sim N(\mu_{h,\gamma}, \sigma_{h,\gamma}^2 / n) \)

In strong signal regime: \( h \gg \sqrt{\gamma} \), for rank one alternatives:

largest eigenvalue test is actually best test

e.g. \( p \) fixed, \( n \) large \[ \text{so } \gamma = p/n \sim 0 \]

\[
\log L(h; \lambda) = \frac{n}{2} \left[ \lambda_1 \frac{h}{h+1} - \log(1 + h) \right] (1 + o(1)).
\]
Change perspective

\[ H \sim W_p(n, \sigma^2 I + hh^T) \]

Consider \( n, p \) fixed

Strong signal: \( h \) large \( \Leftrightarrow \) \( \sigma^2 \) small

Goal: power approximation for \textquote{\text{``Roy's largest root test''}}:

\[
\text{find } P_h(\lambda_1 > \lambda^{(\alpha)}) \quad \text{where} \quad P_0(\lambda_1 > \lambda^{(\alpha)}) = \alpha.
\]
Change perspective

\[ H \sim W_p(n, \sigma^2 I + hh^T) \]

Consider \( n, p \) fixed

Strong signal: \( h \) large \( \Leftrightarrow \) \( \sigma^2 \) small

Goal: power approximation for “Roy’s largest root test”:

find \( P_h(\lambda_1 > \lambda^{(\alpha)}) \) where \( P_0(\lambda_1 > \lambda^{(\alpha)}) = \alpha. \)

An old open issue:

A.T. James (64): ‘For numerical evaluation ... power series expansions of hypergeometric functions are of very limited value.’

T.W. Anderson (84): ‘No straightforward method exists for computing powers for Roy’s statistic itself.’

O’Brien and Shieh (92): ‘To date, no acceptable method has been developed for transforming Roy’s largest root test statistic to an \( F \) or \( \chi^2 \) statistic.’

Dozens of textbooks; G*Power3 software (07): power for linear statistics, not \( \lambda_1 \).
Small $\sigma$ perturbation approach

Initial reductions: $\Sigma = I$, $\nu = e_1$

Suppose, at first deterministically

$$X_i = \begin{pmatrix} u_i \\ 0 \end{pmatrix} + \sigma \begin{pmatrix} 0 \\ \xi_i \end{pmatrix}$$

Then

$$H_\sigma = X^T X = \begin{bmatrix} z & 0^T \\ 0 & 0_{m-1} \end{bmatrix} + \sigma \sqrt{z} \begin{bmatrix} 0 & b^T \\ b & 0_{m-1} \end{bmatrix} + \sigma^2 \begin{bmatrix} 0 & 0^T \\ 0 & Z \end{bmatrix}$$

$$= A_0 + \sigma A_1 + \sigma^2 A_2$$
Small $\sigma$ perturbation approach

Initial reductions: $\Sigma = I, \quad \nu = e_1$

Suppose, at first deterministically

$$X_i = \begin{pmatrix} u_i \\ 0 \end{pmatrix} + \sigma \begin{pmatrix} 0 \\ \xi_i \end{pmatrix}$$

Then

$$H_\sigma = X^T X = \begin{bmatrix} z & 0^T \\ 0 & 0_{m-1} \end{bmatrix} + \sigma \sqrt{z} \begin{bmatrix} 0 & b^T \\ b & 0_{m-1} \end{bmatrix} + \sigma^2 \begin{bmatrix} 0 & 0^T \\ 0 & Z \end{bmatrix}$$

$$= A_0 + \sigma A_1 + \sigma^2 A_2$$

Stochastic assumptions: [to get $W_p(n, \sigma^2 I + h \nu \nu^T)$]

$$u_i \sim N(0, \sigma^2 + h) \quad \xi_i \sim N_{m-1}(0, I)$$
Proposition:

**SP:** Assume $H \sim W_p(n, \sigma^2 I + hvv^T)$. Then

$$\lambda_1(H_\sigma) \sim V_0 + \sigma^2 V_2 + \sigma^4 V_4 + o_p(\sigma^4),$$

with

$$V_0 = (h + \sigma^2)\chi^2_n, \quad V_2 = \chi^2_{p-1}, \quad V_4 = (V_2/V_0)\chi^2_{n-1}$$

and each $\chi^2$ is independent.
Example: Signal Detection, \( h = 10 \)

Density of \( h_1, SP, m = 5, n_H = 4 \)

\[
\frac{(\ell_1 - E[\ell_1])}{\sigma(\ell_1)}
\]

Sim. Gaussian Theory
[Classical] Multivariate Analysis

**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
[Classical] Multivariate Analysis

**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

A. T. James, 1924–2013
A. James (1964): Five-fold Way

Unified view of multivariate eigenvalue distributions:

single matrix,

<table>
<thead>
<tr>
<th>Multivariate</th>
<th>Univariate Analog</th>
<th>Wisharts</th>
<th>Typical Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0F_0$</td>
<td>$\chi^2$</td>
<td>$H \sim W_m(n, \Sigma + \Omega)$</td>
<td>Signal Detection \Sigma known</td>
</tr>
<tr>
<td>$0F_1$</td>
<td>non-central $\chi^2$</td>
<td>$H \sim W_m(n, \Sigma, \Omega)$</td>
<td>Equality of Means \Sigma known</td>
</tr>
</tbody>
</table>
A. James (1964): Five-fold Way

Unified view of multivariate eigenvalue distributions:

two matrix,

<table>
<thead>
<tr>
<th>Multivariate</th>
<th>Univariate Analog</th>
<th>Wisharts</th>
<th>Typical Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1F_0$</td>
<td>$F$</td>
<td>$H \sim W_m(n, \Sigma + \Omega)$</td>
<td>Signal Detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E \sim W_m(n', \Sigma)$</td>
<td>$\Sigma$ unknown</td>
</tr>
<tr>
<td>$1F_1$</td>
<td>non-central $F$</td>
<td>$H \sim W_m(n, \Sigma, \Omega)$</td>
<td>Equality of Means</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E \sim W_m(n', \Sigma)$</td>
<td>$\Sigma$ unknown</td>
</tr>
</tbody>
</table>
A. James (1964): Five-fold Way

Unified view of multivariate eigenvalue distributions:

\[ \text{canonical correlations} \]

<table>
<thead>
<tr>
<th>Multivariate</th>
<th>Univariate Analog</th>
<th>Wisharts</th>
<th>Typical Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>$r^2/(1 - r^2)$</td>
<td>$H \sim W_m(q, \Sigma, \Omega)$</td>
<td>Canonical Correlations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E \sim W_m(n - q, \Sigma)$</td>
<td></td>
</tr>
</tbody>
</table>
A. James (1964): Five-fold Way

Unified view of multivariate eigenvalue distributions:

- single matrix, two matrix, canonical correlations

<table>
<thead>
<tr>
<th>Multivariate</th>
<th>Univariate Analog</th>
<th>Wisharts</th>
<th>Typical Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0F_0$</td>
<td>$\chi^2$</td>
<td>$H \sim W_m(n, \Sigma + \Omega)$</td>
<td>Signal Detection $\Sigma$ known</td>
</tr>
<tr>
<td>$0F_1$</td>
<td>non-central $\chi^2$</td>
<td>$H \sim W_m(n, \Sigma, \Omega)$</td>
<td>Equality of Means $\Sigma$ known</td>
</tr>
<tr>
<td>$1F_0$</td>
<td>$F$</td>
<td>$H \sim W_m(n, \Sigma + \Omega)$</td>
<td>Signal Detection $\Sigma$ unknown</td>
</tr>
<tr>
<td>$1F_1$</td>
<td>non-central $F$</td>
<td>$H \sim W_m(n, \Sigma, \Omega)$</td>
<td>Equality of Means $\Sigma$ unknown</td>
</tr>
<tr>
<td>$2F_1$</td>
<td>$r^2/(1 - r^2)$</td>
<td>$H \sim W_m(q, \Sigma, \Omega)$</td>
<td>Canonical Correlations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E \sim W_m(n - q, \Sigma)$</td>
<td></td>
</tr>
</tbody>
</table>
Power, MANOVA example

Small $\sigma$ perturbation approach extends to rank one alternatives in all James’ cases. One example: ($F_1$ case, MANOVA)
Power, MANOVA example

Small $\sigma$ perturbation approach extends to rank one alternatives in all James’ cases. One example: ($F_1$ case, MANOVA)

<table>
<thead>
<tr>
<th>dim</th>
<th>groups</th>
<th>samples</th>
<th>non-cent $\omega$</th>
<th>power simulated</th>
<th>power approx</th>
<th>relative error</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>0.483</td>
<td>0.477</td>
<td>-.012</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>10</td>
<td>20</td>
<td>0.847</td>
<td>0.852</td>
<td>.006</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>10</td>
<td>40</td>
<td>0.995</td>
<td>0.996</td>
<td>.001</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>0.320</td>
<td>0.304</td>
<td>-.050</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>10</td>
<td>20</td>
<td>0.671</td>
<td>0.668</td>
<td>-.004</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>10</td>
<td>40</td>
<td>0.964</td>
<td>0.967</td>
<td>.003</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>20</td>
<td>10</td>
<td>0.208</td>
<td>0.136</td>
<td>-.346</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>20</td>
<td>20</td>
<td>0.520</td>
<td>0.442</td>
<td>-.150</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>20</td>
<td>40</td>
<td>0.932</td>
<td>0.912</td>
<td>-.021</td>
</tr>
</tbody>
</table>

($\alpha = .05$) (SE $\leq .0016$).

Approximation better for larger $\omega$, smaller $m, p, n$ and plausible power
Conclusion- I

- Spiked Covariance model
  - Examples,

- Phase Transition

- Weak Signals
  - Contiguity

- Strong Signals
  - Power Approximations

- Many extensions possible in other multivariate settings

\[ \Sigma = I + hvv^T \]
McKinsey report 2011: projects excess demand for 140,000 - 190,000 “deep analytical positions”

Maths + Stats + Computing → good jobs
McKinsey report 2011: projects excess demand for 140,000 - 190,000 “deep analytical positions”

Maths + Stats + Computing → good jobs

THANK YOU!