Some aspects of large sample covariance matrices

Jianfeng Yao

Department of Statistics and Actuarial Sciences

The University of Hong Kong

Random Matrices and their Applications, Kyoto University, May 2018

Sample covariance matrix and problem of high-dimensionality

Random matrix theory for large sample covariance matrix

- Marčenko-Pastur distributions
- CLT's for linear spectral statistics

Problem 1: testing on high-dimensional covariance matrices

Problem 2: testing in high-dimensional regressions

An example where Marčenko-Pastur law does not hold

- High-dimensional theory fo eigenvalues of $\mathbf{S}_n$ from mixtures

Sample covariance matrix and problem of high-dimensionality

Sample variance/covariances from a multivariate population

- Let $\mathbf{x}, \ldots, \mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots$ an i.i.d. sequence of $\mathbb{R}^p$-valued random vectors with common distribution μ (population);
- Sample variance/covariance matrix: (assuming $\mathbb{E}(\mathbf{x}) = \mathbf{0}$)

$$\mathbf{S}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^T .$$

  That is, if we write $\mathbf{x}_k = (\xi_{1k}, \ldots, \xi_{pk})^T$,

$$\mathbf{S}_n(i,j) = \frac{1}{n}\sum_{k=1}^{n} \xi_{ik}\xi_{jk}, \quad 1 \le i,j \le p.$$

  [sample cross-moments between dimensions/variables i and j.]
- The population variance/covariance matrix is

$$\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^T], \quad (p \times p).$$

  Both $\mathbf{S}_n$ and $\mathbf{\Sigma}$ are nonnegative definite and trivially,

$$\mathbb{E}\mathbf{S}_n = \mathbf{\Sigma}.$$

Is $\mathbf{S}_n$ a "good enough" estimator of Σ ?

Large sample theory

Holding the dimension p while letting the sample size $n \to \infty$:

1. Law of large numbers: $\mathbf{S}_n \overset{\text{a.s.}}{\to} \mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^T]$, [once $\mathbb{E}[\|\mathbf{x}\|^2] < \infty$]
2. Central limit theorem: $\sqrt{n}\,[\mathbf{S}_n - \mathbf{\Sigma}] \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$,

   with some asymptotic variance/covariance matrix $\mathbf{\Lambda}$.

[once $\mathbb{E}[\|x\|^4] < \infty$]

- **A fundamental issue in statistics:**

  When analyzing a real "high-dimensional" data set with given (p, n)

  such that $p/n \gg 0$, *for example* $(p = 100, n = 500)$,

  approximation from this classical large sample theory becomes

  biased and inefficient!

(a). High-dimensional data is now common

- Many sources to high-dimensional data: electronic trading in finance; genomics;
- typical data dimensions and sample sizes:

	# variables p	sample size n	**ratio** p/n	**Small / Big**
portfolio	$\sim$ **100**	**500**	**0.2**	S
climate survey	320	600	0.21	S
speech analysis	$\sim 10^3$	$\sim 10^3$	~ 1	S
ORL face data base	1440	320	4.5	B
micro-arrays	10000	1000	10	B

(b). Example illustrating inefficiency of classical large sample limits

Consider

- a "white"/unit Gaussian populaiton $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_p)$, that is,

$$\mathbf{x} = (\xi_1, \ldots, \xi_p)^T, \quad \xi_\ell \text{ are i.i.d. } \mathcal{N}(0, 1).$$

- given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $\mathbf{x}$, the sample covariance matrix is,

$$\mathbf{S}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = \frac{1}{n} \mathbf{W}_n,$$

  Here

$$\mathbf{W}_n = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = (\mathbf{x}_1, \ldots, \mathbf{x}_n)(\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \quad \sim \quad \text{Wishart}(n, \mathbf{I}_p)$$

- let $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ be eigenvalues of $\mathbf{S}_n$.

Example (cont.)

Large sample limits: p **fixed while** $n \to \infty$

1. LLN: $\mathbf{S}_n \overset{a.s.}{\to} \mathbf{I}_p$; by continuity, $(\lambda_1, \ldots, \lambda_p) \overset{a.s.}{\to} 1$.
2. CLT:

$$\sqrt{n}\,(\mathbf{S}_n - \mathbf{I}_p) \;\Rightarrow\; \mathcal{N}(0, *),$$

   By delta method,

$$\sqrt{n}\,\left\{(\lambda_1^2 + \cdots + \lambda_p^2) - p\right\} \Rightarrow \mathcal{N}(0, *).$$

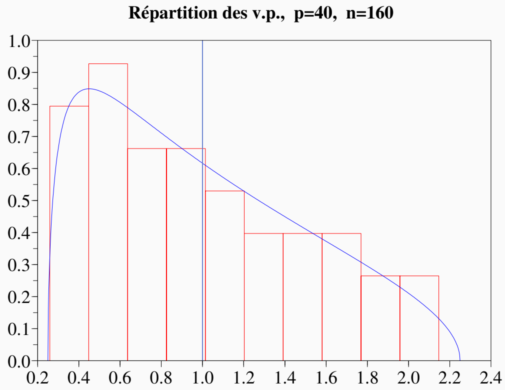
Random-matrix-theory (RMT) limits:

$n \to \infty$, $p = p_n \to \infty$ **such that** $p_n/n \to c > 0$

1. LLN: $\mathbf{S}_n \not\to \mathbf{I}_p$; $\frac{1}{p}\sum_{k=1}^{p} \delta_{\lambda_k} \;\Rightarrow$ Marčenko-Pastur law
2. CLT:

$$(\lambda_1^2 + \cdots, +\lambda_p^2) - p - p^2/n \Rightarrow \mathcal{N}(m, *).$$

Example (cont.)



1. Histogram of 40 eigenvalues of S_n simulated with $p = 40$ and $n = 160$
2. blue curve = RMT limit: Marčenko-Pastur law with index $\frac{p}{n} = \frac{1}{4}$

$$f(x) = \frac{1}{2\pi c x}\sqrt{(b-x)(x-a)}, \quad x \in [a,b] = [0.25, 2.25]$$

3. large sample limit: sample eigenvalues $\simeq 1$

(c) Marčenko-Pastur paradigm for high-dimensional statistics

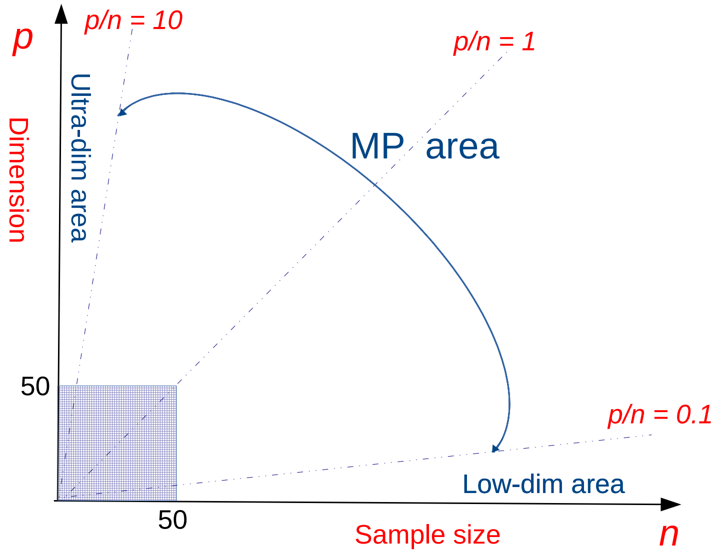
- Both the large sample limits and random matrix theory limits are mathematical theorems, are thus theoretically correct;
- But the question from a responsible statistician (now "data scientist") would be:

  Which theory to follow if data table has $(p, n) = (40, 160)$?

- Previous simulation shows clearly that

  RMT Marčenko-Pastur limit $\gg$ classical large sample limit !

Empirical performance of the Marčenko-Pastur limiting scheme



Random matrix theory for large sample covariance matrix

The Marčenko-Pastur distribution

Theorem. Assume : Marčenko & Pastur, 1967

- Population $\mathbf{x} = (\xi_1, \ldots, \xi_p)^T$ has i.i.d. components with mean 0 and variance 1; (so $\mathbf{\Sigma} = \mathbf{I}_p$);
- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is an i.i.d. sample of $\mathbf{x}$;
- $n \to \infty$, $p = p(n) \to \infty$ and $p/n \to y \in (0, 1]$;

Then, the eigenvalue distribution of

$$S_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^T = \frac{1}{n} \mathbf{X}\mathbf{X}^T = \frac{1}{n} (\mathbf{x}_1, \ldots, \mathbf{x}_n)(\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$$

converges to the distribution with density function

$$f(x) = \frac{1}{2\pi y x} \sqrt{(x-a)(b-x)}, \quad a \le x \le b,$$

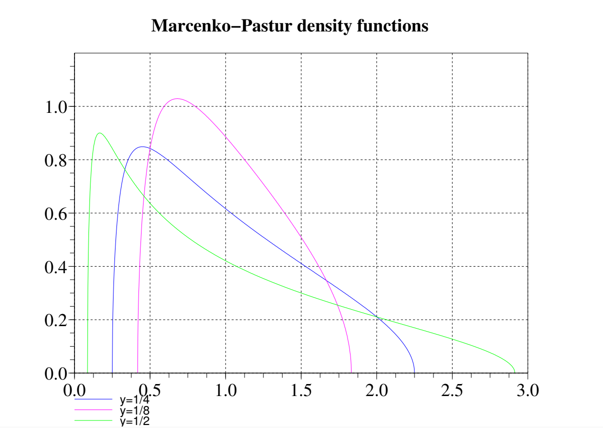
where

$$a = (1 - \sqrt{y})^2, \quad b = (1 + \sqrt{y})^2 .$$

The Marčenko-Pastur distribution

$$f(x) = \frac{1}{2\pi y x}\sqrt{(x-a)(b-x)}, \qquad (1-\sqrt{y})^2 = a \le x \le b = (1+\sqrt{y})^2.$$

$y \sim p/n$	a	b
1/8	0.42	1.83
1/4	0.25	2.25
1/2	0.09	2.91



The generalized Marčenko-Pastur distribution

Theorem. Assume : Marčenko & Pastur, (1967); Silverstein (1995)

- $\mathbf{X} = p \times n$ i.i.d. variables $(0, 1)$;
- $n \to \infty$, $p = p(n) \to \infty$ and $p/n \to y \in (0, 1]$;
- $(T_p)_{p \geq 1}$ is a sequence of non-negative Hermitian matrices whose eigenvalue distributions $(H_p)_p$ tend to a deterministic probability distribution H;

Then, the eigenvalue distribution of $S_n = \frac{1}{n} T_p^{1/2} \mathbf{X}\mathbf{X}^T T_p^{1/2}$ converges to a deterministic distribution $F_{y,H}$ characterized by its Stieltjes transform m which solves the following Marčenko-Pastur equation

$$m = \int \frac{1}{t(1 - y - yzm) - z} dH(t).$$

This solution is unique in the set $\{m \in \mathbb{C}^+ : -(1 - y)/z + ym \in \mathbb{C}^+\}$.

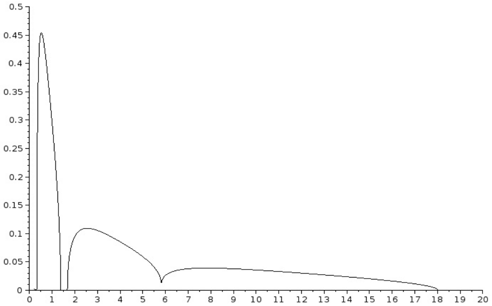
An example of generalized Marčenko-Pastur distribution

Assuming that $T_p = \mathrm{diag}\{\underbrace{1,\ldots,1}_{1/3},\underbrace{4,\ldots,4}_{1/3},\underbrace{10,\ldots,10}_{1/3}\}.$

Then the limiting Stieltjes transform m solves:

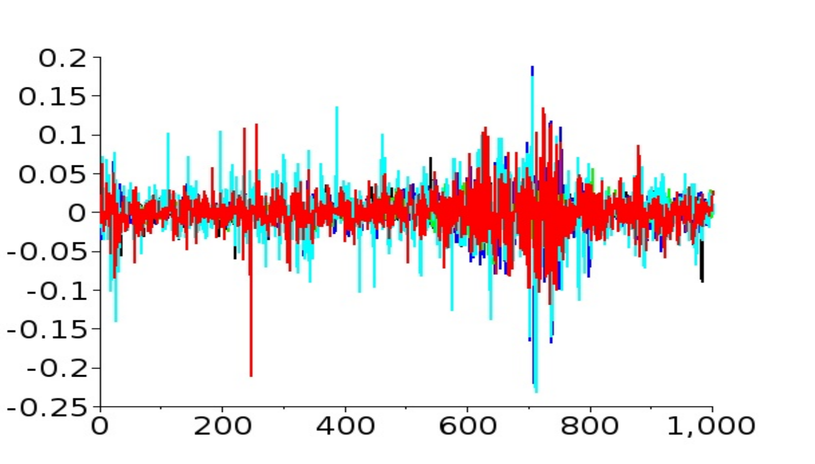
$$m = \frac{1/3}{1-y-yzm-z} + \frac{1/3}{4(1-y-yzm)-z} + \frac{1/3}{10(1-y-yzm)-z} \; .$$

By inversion of Stieltjes transform, density function is:



Example of stock data

- SP 500 daily stock prices ; $p = 488$ stocks;
- $n = 1000$ daily returns $\mathbf{r}_t(i) = \log p_t(i)/p_{t-1}(i)$ from 2007-09-24 to 2011-09-12;



- Let the SCM (488× 488)

$$\mathbf{S}_n = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{r}_t - \bar{\mathbf{r}})(\mathbf{r}_t - \bar{\mathbf{r}})^T \ .$$

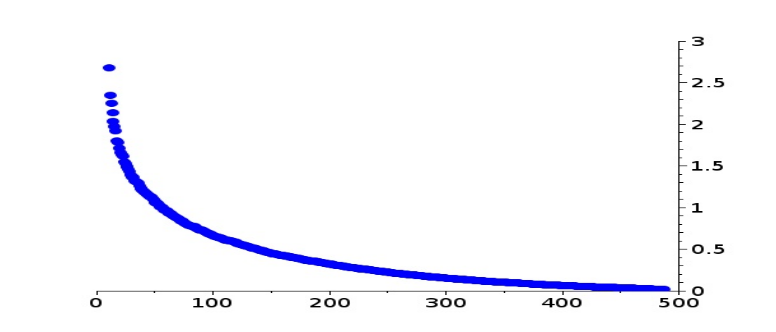
- We consider the sample correlation matrix $\mathbf{R}_n$ with

$$\mathbf{R}_n(i,j) = \frac{S_n(i,j)}{[S_n(i,i)S_n(j,j)]^{1/2}}.$$

- The 10 largest and 10 smallest eigenvalues of $\mathbf{R}_n$ are:

237.95801	4.8568703	...	0.0212137	0.0178129
17.762811	4.394394	...	0.0205001	0.0173591
14.002838	3.4999069	...	0.0198287	0.0164425
8.7633113	3.0880089	...	0.0194216	0.0154849
5.2995321	2.7146658	...	0.0190959	0.0147696

Sample eigenvalues of stock returns



[excluding the 10 largest: $\lambda_{11}, \ldots, \lambda_{488}$]

- Two important questions:
  - Explanation the largest sample eigenvalues (spikes, perturbation);
  - Provide a model for bulk correlation structure between the 488 returns.
- Both successfully analysed using

  Genelized Marčenko-Pastur distribution + spiked outliers

General issue:

- Assume that for a sequence of E.S.D F_n $\quad F_n = \frac{1}{p}\sum_{j=1}^{p} \delta_{\lambda_j}$, we have proved the existence of a limiting distribution F;
- Given a "smooth" function g, e.g. $g(x) = x - 1 - \log x$, consider the linear spectral statistic (LSS):

$$F_n(g) = \frac{1}{p}\sum_{j=1}^{p} g(\lambda_j)$$

- Problem: find a_n, b_n s.t.

$$a_n\left[F_n(g) - b_n\right] \Longrightarrow \mathcal{N}(m, V)$$

for some asymptotic mean m and variance V.

CLT for LSS of sample covariance matrices

- Consider a sequence of sample covariance matrices S_n s.t $F^{S_n} \Longrightarrow F_y$, the Marcčenko-Pastur distribution of index y;
- CLT's for regular functions g have a long history

  Arharov (1971); Jonsson (1982) ; Johnsson (1998); Sinai & Soshnikov (1998);
  Bai & Silverstein (2004); Bai and Y. (2005); Lytova & Pastur (2009)

Following Bai & Silverstein '04, let

- an open set $\mathcal{U}$ of $\mathbb{C}$ including the support $[a, b] = [(1-\sqrt{y})^2, (1+\sqrt{y})^2]$ of the LSD
- for any g analytic on $\mathcal{U}$: $G_n(g) = p\,[F_n(g) - \mu^{y_n}(g)]$
  where μ^{y_n} is the MP distribution of index $y_n \in (0, 1)$.

Bai and Silverstein (2004)

Theorem

Assume that

- $g_1, \cdots, g_k$ *are* k *analytic functions on* $\mathcal{U}$;
- *the matrix entries* x_{ij} *are i.i.d. real-valued random variables such that* $Ex_{ij} = 0,\ Ex_{ij}^2 = 1,\ E{x_{ij}}^4 = 3.$
- *as* $n, p \to \infty$, $y_n = \frac{p}{n} \to y \in (0, 1)$;

Then,

$$(G_n(g_1), \cdots, G_n(g_k)) \Rightarrow \mathcal{N}_k(m, V),$$

with a given mean vector $m = m(g_1, \ldots, g_k)$ *and asymptotic covariance matrix* $V = V(g_1, \ldots, g_k)$.

CLT for LSS of random Fisher matrices

- two independent samples:

$$\mathbf{x}_1, \ldots, \mathbf{x}_{n_1} \sim (0, I_p), \qquad \mathbf{y}_1, \ldots, \mathbf{y}_{n_2} \sim (0, I_p)$$

  with i.i.d coordinates of mean 0 and variance 1

- Associated sample covariance matrices:

$$S_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_i \mathbf{x}_i^T, \qquad S_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{y}_j \mathbf{y}_j^T.$$

- Fisher matrix: $V_n = S_1 S_2^{-1}$ where $n_2 > p$.

- Assume

$$y_{n_1} = \frac{p}{n_1} \to y_1 \in (0,1), \qquad y_{n_2} = \frac{p}{n_2} \to y_2 \in (0,1) \ .$$

- Under mild moment conditions, the ESD $F_n^{V_n}$ of V_n has a LSD F_{y_1,y_2} with density (Wachter distribution):

$$\ell(x) = \begin{cases} \dfrac{(1-y_2)\sqrt{(b-x)(x-a)}}{2\pi x(y_1+y_2x)}, & a \le x \le b, \\ 0, & \text{otherwise} \end{cases}$$

where

$$a = (1-y_2)^{-2}\left(1-\sqrt{y_1+y_2-y_1y_2}\right)^2, \quad b = (1-y_2)^{-2}\left(1+\sqrt{y_1+y_2-y_1y_2}\right)^2.$$

CLT for LSS of random Fisher matrices

- let $\widetilde{\mathcal{U}} \subset \mathbb{C}$ be an open set including the interval

$$\left[I_{(0,1)}(y_1)\frac{(1-\sqrt{y_1})^2}{(1+\sqrt{y_2})^2}, \quad \frac{(1+\sqrt{y_1})^2}{(1-\sqrt{y_2})^2} \right],$$

- for an analytic function f on $\widetilde{\mathcal{U}}$, define

$$\widetilde{G_n}(f) = p\left[F_n^{V_n}(g) - F_{y_{n_1}, y_{n_2}}(g) \right] ,$$

where $F_{y_{n_1}, y_{n_2}}$ is the LSD with indexes y_{n_k}, $k = 1, 2$.

Zheng (2008)

Theorem

Assume $E\mathbf{x}_{11}^4 = E\mathbf{y}_{11}^4 < \infty$ *and let* $\beta = E|\mathbf{x}_{11}|^4 - 3$. *Then for any analytic functions* $f_1, \cdots, f_k$ *defined on* $\widetilde{\mathcal{U}}$,

$$\left[\widetilde{G}_n(f_1), \cdots, \widetilde{G}_n(f_k)\right] \Longrightarrow \mathcal{N}_k(m, \upsilon) \ .$$

Zheng (2008)

Limiting mean function m

$$m(f_j) = \lim_{r \to 1_+} [(2.1) + (2.2) + (2.3)]$$

$$\frac{1}{4\pi i} \oint_{|\zeta|=1} f_j(z(\zeta)) \left[\frac{1}{\zeta - \frac{1}{r}} + \frac{1}{\zeta + \frac{1}{r}} - \frac{2}{\zeta + \frac{y_2}{hr}} \right] d\zeta \quad (2.1)$$

$$+ \frac{\beta \cdot y_1 (1 - y_2)^2}{2\pi i \cdot h^2} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{1}{(\zeta + \frac{y_2}{hr})^3} d\zeta \quad (2.2)$$

$$+ \frac{\beta \cdot y_2 (1 - y_2)}{2\pi i \cdot h} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{\zeta + \frac{1}{hr}}{(\zeta + \frac{y_2}{hr})^3} d\zeta, \quad (2.3)$$

where

$$z(\zeta) = (1 - y_2)^{-2} \left[1 + h^2 + 2h\mathcal{R}(\zeta) \right], \qquad h = \sqrt{y_1 + y_2 - y_1 y_2}. \quad (2.4)$$

Zheng (2008)

Limiting covariance function υ

$$\upsilon(f_j, f_\ell) = \lim_{1<r_1<r_2\to 1_+} [(2.5) + (2.6)]$$

$$-\frac{1}{2\pi^2}\oint_{|\zeta_2|=1}\oint_{|\zeta_1|=1}\frac{f_j(z(r_1\zeta_1))f_\ell(z(r_2\zeta_2))r_1r_2}{(r_2\zeta_2 - r_1\zeta_1)^2}d\zeta_1 d\zeta_2, \tag{2.5}$$

$$-\frac{\beta\cdot(y_1+y_2)(1-y_2)^2}{4\pi^2h^2}\oint_{|\zeta_1|=1}\frac{f_j\left(z(\zeta_1)\right)}{(\zeta_1+\frac{y_2}{hr_1})^2}d\zeta_1\oint_{|\zeta_2|=1}\frac{f_\ell\left(z(\zeta_2)\right)}{(\zeta_2+\frac{y_2}{hr_2})^2}d\zeta_2 \tag{2.6}$$

$j, \ell \in \{1, \cdots, k\}.$

Problem 1: testing on high-dimensional covariance matrices

Testing structure of a large covariance matrix

- a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- want to test hypothesis about structure of $\boldsymbol{\Sigma}$:
  - $\boldsymbol{\Sigma} = \mathbf{I}_p$ (identity test)
  - $\boldsymbol{\Sigma} = c \times \mathbf{I}_p$, c unknown (sphericity test)
  - $\boldsymbol{\Sigma}$ is diagonal, block diagonal, Toeplitz, band, etc.
- in high-dimensional case, several previous work exist:
  Ledoit & Wolf '02; Schott '07; Srivastava '05 ...
- we focus on the simplest case of identity test $H_0 : \quad \boldsymbol{\Sigma} = \mathbf{I}_p$
- LR statistic:

$$T_n = n\left[trS_n - \log|S_n| - p\right], \quad S_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})',$$

Classical LRT -large sample limit:

- when $n \to \infty$, $T_n \Longrightarrow \chi^2_{p(p+1)/2}$ (data dimension p is fixed)
- Procedure based on this limit is rapidly deficient when p is not "small".

RMT Corrected LRT:

Bai, Jiang, Y. and Zheng (2009)

Theorem

Assume $p/n \to y \in (0,1)$ *and let* $g(x) = x - \log x - 1$. *Then, under* H_0 *and when* $n \to \infty$

$$\left[\frac{T_n}{n} - p \cdot F^{y_n}(g)\right] \Rightarrow \mathcal{N}(m(g), \upsilon(g)),$$

where F^{y_n} *is the Marčenko-Pastur law of index* y_n *and*

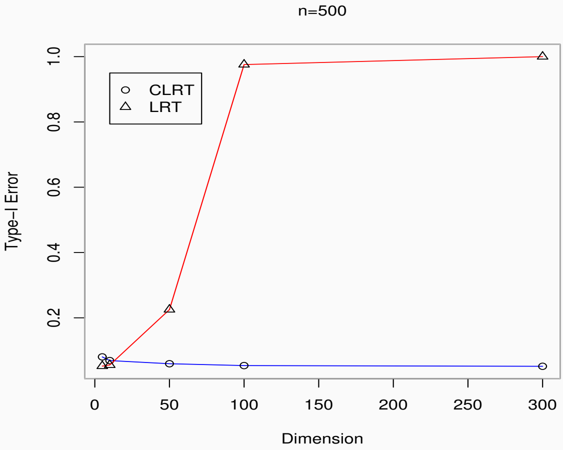
$$\begin{aligned} m(g) &= -\frac{\log(1-y)}{2}, \\ \upsilon(g) &= -2\log(1-y) - 2y. \end{aligned}$$

Comparison of LRT and Corrected LRT by simulation

- nominal test level $\alpha = 0.05$;
- for each (p, n), 10,000 independent replications with real Gaussian variables.
- Powers are estimated under the alternative H_1:
  $\Sigma = \mathrm{diag}(1, 0.05, 0.05, 0.05, \ldots, 0.05)$.

	CLRT			LRT	
(p, n)	Size	Difference with 5%	Power	Size	Power
(5, 500)	0.0803	0.0303	0.6013	0.0521	0.5233
(10, 500)	0.0690	0.0190	0.9517	0.0555	0.9417
(50, 500)	0.0594	0.0094	1	0.2252	1
(100, 500)	0.0537	0.0037	1	0.9757	1
(300, 500)	0.0515	0.0015	1	1	1

On a plot


n=500
CLRT
LRT
Type–I Error
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
Dimension

Problem 2: testing in high-dimensional regressions

A general linear hypothesis in a multivariate regression

A p-th dimensional regression model:

$$\mathbf{x}_i = \mathbf{B}\mathbf{z}_i + \varepsilon_i, \quad i = 1, \ldots, n$$

where

$$\varepsilon_i \sim \mathcal{N}_p(0, \boldsymbol{\Sigma}), \quad \mathbf{x}_i \in \mathbb{R}^p, \quad \mathbf{z}_i \in \mathbb{R}^q, \quad n \geq p + q.$$

A general linear hypothesis:

- Write a bloc decomposition $\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2)$ with q_1 and q_2 columns ($q = q_1 + q_2$)
- To test

$$H_0 : \ \mathbf{B}_1 = \mathbf{M} \ ,$$

  with a given $\mathbf{M}$.

Wilk's $\boldsymbol{\Lambda}$

- Let $\widehat{\boldsymbol{\Sigma}}_0$ and $\widehat{\boldsymbol{\Sigma}}_1$ be the likelihood "estimator" of $\boldsymbol{\Sigma}$ under H_0 and the alternative, respectively
- LRT statistic equals

$$\mathscr{L}_0/\mathscr{L}_1 = (\boldsymbol{\Lambda}_n)^{n/2}, \quad \boldsymbol{\Lambda}_n = \frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_0|} ,$$

where $\boldsymbol{\Lambda}_n$ is the celebrated Wilk's $\boldsymbol{\Lambda}$: Wilks '32, '34 ; Bartlett '34.

- Classic (low dimensional) approximation of LRT: for fixed p and q, $n \to \infty$ and under H_0:

$$U_n = -n \log \boldsymbol{\Lambda}_n \Rightarrow \chi^2_{pq_1} .$$

- Less biased Bartlett's correction:

$$\tilde{U}_n = -k \log \boldsymbol{\Lambda}_n, \quad k = n - q - \frac{1}{2}(p - q_1 + 1) .$$

Bai, Jiang, Y. and Zheng (2010)

Theorem

Let $p \to \infty,\ q_1 \to \infty,\ n-q \to \infty$ and

$$y_{n_1} = \frac{p}{q_1} \to y_1 \in (0,1), \quad y_{n_2} = \frac{p}{n-q} \to y_2 \in (0,1).$$

Then, under H_0,

$$T_n = \upsilon(f)^{-\frac{1}{2}} \left[-\log \boldsymbol{\Lambda}_n - p \cdot F_{y_{n_1}, y_{n_2}}(f) - m(f)\right] \Rightarrow \mathscr{N}(0,1),$$

where $m(f), \upsilon(f)$ and $F_{y_{n_1}, y_{n_2}}(f)$ are suitable constants computed from

$$f(x) = \log(1 + \frac{y_{n_2}}{y_{n_1}} x) \ .$$

The centering term:

$$F_{y_{n_1},y_{n_2}}(f) = \frac{y_{n_2}-1}{y_{n_2}}\log c_n + \frac{y_{n_1}-1}{y_{n_1}}\log(c_n - d_n h_n)$$
$$= +\frac{y_{n_1}+y_{n_2}}{y_{n_1}y_{n_2}}\log\left(\frac{c_n h_n - d_n y_{n_2}}{h_n}\right),$$

where

$$h_n = \sqrt{y_{n_1}+y_{n_2}-y_{n_1}y_{n_2}}$$
$$a_n, b_n = \frac{(1\mp h_n)^2}{(1-y_{n_2})^2}$$
$$c_n, d_n = \frac{1}{2}\left[\sqrt{1+\frac{y_{n_2}}{y_{n_1}}b_n} \pm \sqrt{1+\frac{y_{n_2}}{y_{n_1}}a_n}\right], c_n > d_n,$$

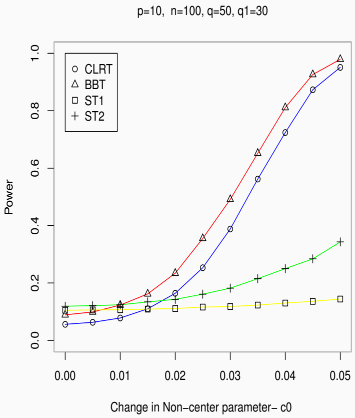
The limiting parameters:

$$
\begin{aligned}
m(f) &= \frac{1}{2}\log\frac{(c^2-d^2)h^2}{(ch-y_2d)^2}\,,\\
\upsilon(f) &= 2\log\left(\frac{c^2}{c^2-d^2}\right)\,,
\end{aligned}
$$

where

$$
\begin{aligned}
h &= \sqrt{y_1+y_2-y_1y_2}\\
a_0, b_0 &= \frac{(1\mp h)^2}{(1-y_2)^2}\\
c, d &= \frac{1}{2}\left[\sqrt{1+\frac{y_2}{y_1}b_0} \pm \sqrt{1+\frac{y_2}{y_1}a_0}\right], c > d.
\end{aligned}
$$

A simulation experiment



- Gaussian entries,
- non central parameter $c_0 \sim d(H, H_0)$.

An example where Marčenko-Pastur law does not hold

Multivariate normal mixture

- p-dimensional *multivariate normal mixture* (MNM):

$$f(\mathbf{x}) = \sum_{j=1}^{K} \alpha_j \phi(\mathbf{x};\, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{5.1}$$

  where

  - (α_j): K mixing weights
  - $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$: parameters of the jth Gaussian component (ϕ is the multivariate Gaussian density function)

- high-dimensional situations: p is large compared to the sample size n.

Statistical testing problem

- Test for the covariance matrix in the MNM model

$$f(\mathbf{x}) = \sum_{j=1}^{K} \alpha_j \phi(\mathbf{x};\, \boldsymbol{\mu}_j, \sigma_j^2 \mathbf{T}_p^2) \quad \text{with} \quad \boldsymbol{\mu}_j = 0 \tag{5.2}$$

in high-dimensional situations.

- This model is a special case of a p-dimensional *scale mixture*,

$$\mathbf{x} = w \mathbf{T}_p \mathbf{z}, \tag{5.3}$$

where

  - $\mathbf{z} = (z_1, \ldots, z_p)'$ are i.i.d. $E(z_i) = 0$, $E(z_i^2) = 1$;
  - $w > 0$ is a random scale, independent of $\mathbf{z}$;
  - $\mathbf{T}_p \in \mathbb{R}^{p \times p}$, $\mathbf{T}_p > 0$, $\mathrm{tr}(\mathbf{T}_p^2)/p = 1$;

Indeed: (5.3) $\Longrightarrow$ (5.2) if $\mathbf{z} \sim N(0, \mathbf{I}_p)$, $\mathbf{T}_p = \mathbf{I}_p$ and $P(w^2 = \sigma_j^2) = \alpha_j$.

- Terminology: distribution of w^2, denoted G, referred as *Population Mixing Distribution* (PMD).

- Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sample from the mixture $\mathbf{x}$, with population covariance matrix $\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^T]$
- Sample covariance matrix: $\mathbf{S}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'.$
- Random matrix theory: for p, n large,

$$\text{eigenvalues of } \mathbf{\Sigma} \quad \rightsquigarrow \quad \text{eigenvalues of } \mathbf{S}_n$$

Terminology. *Empirical spectral distribution* (ESD) of a $p \times p$ symmetric matrix $\mathbf{A}$:

$$\mu_{\mathbf{A}} = \frac{1}{p} \sum_{j=1}^{p} \delta_{\lambda_j},$$

where $(\lambda_j)_{1 \le j \le p}$ are the eigenvalues of $\mathbf{A}$, (δ_b: the Dirac mass at b).

Existing random matrix theory

population eigenvalues sample eigenvalues

$\mu_{\boldsymbol{\Sigma}}$: ESD of $\boldsymbol{\Sigma}$ $\rightsquigarrow$ $\mu_{\mathbf{S}_n}$: ESD $\mathbf{S}_n$

Findings

Mixtures are not a usual high-dimensional population:

normal population with $\boldsymbol{\Sigma} = \mathbf{I}_p$: $\mu_{\mathbf{S}_n} \sim$ Marčenko-Pastur law

mixture of normals with $\boldsymbol{\Sigma} = \mathbf{I}_p$: $\mu_{\mathbf{S}_n} \neq$ Marčenko-Pastur law

(Both populations have uncorrelated components!)

Case of uncorrelated population I

- Consider the simplest case of $\mathbf{x} = \mathbf{z}$: $E(\mathbf{x}) = 0$, $\text{cov}(\mathbf{x}) = I_p$.
- Assume the Marčenko-Pastur regime:

$$p = p_n, \quad \text{and} \quad p_n/n \to c > 0 \quad \text{as} \quad n \to \infty.$$

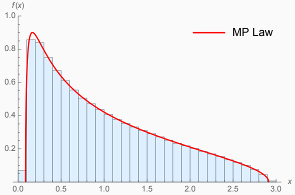
- We have that

$$\mu_{\mathbf{S}_n} \xrightarrow[w]{a.s.} \nu \text{ (MP law).}$$

- $\nu(dx) = f(x)dx + (1-1/c)\delta_0(dx)1_{\{c>1\}}$ where

$$f(x) = \frac{\sqrt{(b-x)(x-a)}}{2\pi cx} 1_{[a,b]}(x),$$

where $a = (1-\sqrt{c})^2$ and $b = (1+\sqrt{c})^2$.



The Marčenko-Pastur law (red line). The dimensions are $(p, n, c) = (500, 1000, 0.5)$ and $rep = 100$.

Case of uncorrelated population II

- Consider a simple mixture $\mathbf{x} = w\mathbf{z}$ where $E(w^2) = 1$;

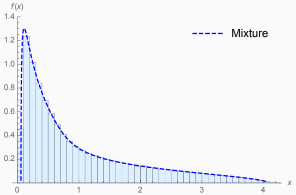
  we have $E(\mathbf{x}) = 0$, $\text{cov}(\mathbf{x}) = \mathbf{I}_p$.

- Assume again the Marčenko-Pastur regime: $p_n/n \to c > 0$.

- We have that

$$\mu_{\mathbf{S}_n} \xrightarrow[w]{a.s.} F^{c,G} \neq \text{MP law}.$$

- Example: The MNM is $f(\mathbf{x}) = 0.25\phi(\mathbf{x};\, 0, 2.5I_p) + 0.75\phi(\mathbf{x};\, 0, 0.5I_p)$ with $c = 1/2$.



The LSD (blue line) from the MNM. The dimensions are $(p, n, c) = (500, 1000, 0.5)$ and $rep = 100$. The support interval is $[0.0576, 4.0674]$.

Comparison between the two cases

Uncorrelated populaitons: $\mathbf{x} = \mathbf{z}$ versus $\mathbf{x} = w\ \mathbf{z}$

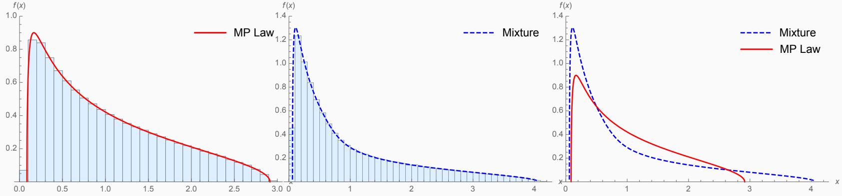


Figure 1: The Marčenko-Pastur law (red line) v.s. the LSD (blue line) from an MNM with identity covariance. The dimensions are $(p, n, c) = (500, 1000, 0.5)$ and $rep = 100$. The support intervals are $[0.0858, 2.9142]$ and $[0.0576, 4.0674]$, respectively.

High-dimensional mixtures needs new random matrix theory

Why mixtures are different?

- Main reason: coordinates of $\mathbf{x}$ could be uncorrelated but strongly dependent in the sense that:

$$\mathrm{var}(\|\mathbf{x}\|^2) \propto p^2, \quad p \to \infty.$$

- Consequence: much we have done so far for high-dimensional covariance matrices do not apply to high-dimensional mixtures.

- **Remark**
  - It is known that if for any bounded sequence (in spectral norm) $(\mathbf{A}_p)$, we have

$$\mathrm{var}\mathbf{x}^T\mathbf{A}_p\mathbf{x} = o(p^2),$$

then the corresponding sample covariance $\mathbf{S}_n$ satisfies the Marčenko-Pastur law. Bai and Zhou (2008)

Also called "good vector" by Pastur and Pajor (2009)

Setting of a general scale mixture

Assumption (a). The sample and population sizes n, p both tend to infinity with their ratio $c_n = p/n \to c \in (0, \infty)$.

Assumption (b). There are two independent arrays of i.i.d. random variables $(z_{ij})_{i,j\geq 1}$ and $(w_i)_{i\geq 1}$, satisfying

$$\mathbb{E}(z_{11}) = 0, \quad \mathbb{E}(z_{11}^2) = 1, \quad \mathbb{E}(z_{11}^4) < \infty, \tag{5.4}$$

such that for each p and n the observation vectors can be represented as $\mathbf{x}_i = w_i \mathbf{T}_p \mathbf{z}_i$ with $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})'$, $i = 1, \ldots, n$.

Assumption (c). The spectral distribution H_p of the matrix $\mathbf{T}_p^2$ weakly converges to a probability distribution H, as $p \to \infty$, referred as *Population Spectral Distribution* (PSD).

Assumption (d). The support set S_G of the MD G is bounded above and from below, that is $S_G \subset [a, b]$ for some $0 < a < b < \infty$.

Theorem

Suppose that Assumptions (a)-(c) hold. Then, almost surely, the empirical spectral distribution $\mu_n := \mu_{\mathbf{S}_n}$ *converges in distribution to a probability distribution* $F^{c,G,H}$ *whose Stieltjes transform* $m = m_{F^{c,G,H}}(z)$ *is a solution to the following system of equations, defined on the upper complex plane* $\mathbb{C}^+$*,*

$$\begin{cases} zm(z) = -1 + \int \frac{p(z)t}{1+cp(z)t} dG(t), \\ zm(z) = -\int \frac{1}{1+q(z)t} dH(t), \\ zm(z) = -1 - zp(z)q(z), \end{cases} \tag{5.5}$$

where $p(z)$ *and* $q(z)$ *are two auxiliary analytic functions. The solution is also unique in the set*

$$\{m(z) : -(1-c)/z + cm(z) \in \mathbb{C}^+,\ zp(z) \in \mathbb{C}^+,\ q(z) \in \mathbb{C}^+,\ z \in \mathbb{C}^+\}.$$

Li and Y. (2017)

Some special cases of limiting spectral distributions

- When the distributions H and/or G degenerate to some Dirac mass, the system (5.5) simplifies to a single equation leading to several well-known LSDs.
  - Case 1. If $H = G = \delta_1$, then the equations become
    $$z = -\frac{1}{m} + \frac{1}{1 + cm},$$
    which defines the standard MP law (Marčenko-Pastur, 1969).
  - Case 2. If $G = \delta_1$, then the equations turn into
    $$m = \int \frac{1}{t(1 - c - cmz) - z} dH(t) \ ,$$
    which defines the generalized MP law (Silverstein 1995).
  - Case 3. If $H = \delta_1$, then the equations reduce to
    $$z = -\frac{1}{m} + \int \frac{t}{1 + ctm} dG(t) \ , \tag{5.6}$$
    which defines an LSD corresponding to a scale-mixture population with spherical covariance matrix.

Fluctuations of eigenvalue statistics

- We study the fluctuation of linear spectral statistics (LSS) of $\mathbf{S}_n$ under the simplest spherical mixture model:

$$\mathbf{x} = w\mathbf{z},$$

  that is $\mathbf{T}_p = \mathbf{I}_p$ and the PSD $H = \delta_1$.

- By the previous theorem, $\mu_n := \mu_{\mathbf{S}_n} \xrightarrow{\mathscr{D}} F^{c,G}$.

- Linear spectral statistics (LSS) are of the form

$$\frac{1}{p}\sum_{j=1}^{p} f(\lambda_j) = \int f(x) d\mu_{\mathbf{S}_n}(x) = \int f d\mu_{\mathbf{S}_n}$$

  where f is a function on $[0, \infty)$.

- In Bai and Silverstein (2004), the LSS under their settings are proved to be asymptotically normal distributions:

$$\sum_{j=1}^{p} f(\lambda_j) - p \cdot \kappa(n, p) \xrightarrow{D} N(a, s^2)$$

  However, we show that this CLT does not apply to the present model of scale mixtures.

Fluctuations of eigenvalue statistics

- Express the sample as $\mathbf{x}_j = w_j \mathbf{z}_j$, $j = 1, \ldots, n$, and let

$$G_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{w_j^2}, \quad \text{ESD } \mu_n \approx \begin{cases} F^{c,G} & c_n \to c,\, G_n \xrightarrow{w} G, \\ F^{c_n,G} & c \text{ is replaced with } c_n, \\ F^{c_n,G_n} & (c, G) \text{ is replaced with } (c_n, G_n) \end{cases}$$

- The aim here is to study the fluctuation of

$$\boxed{\frac{1}{p}\sum_{j=1}^{p} f(\lambda_j) - \int f(x) dF^{c_n,G}(x) = \int f \cdot d(\mu_n - F^{c_n,G})}$$

through the decomposition

$$\int f \cdot d(\mu_n - F^{c_n,G}) = \int f \cdot d(\mu_n - F^{c_n,G_n}) + \int f \cdot d(F^{c_n,G_n} - F^{c_n,G})$$

Write it as:

$$\int f \cdot d\mathcal{F}_n = \int f \cdot d\mathcal{F}_{n1} + \int f \cdot d\mathcal{F}_{n2}.$$

A central limit theorem

Theorem

Suppose that Assumptions (a)-(d) hold. Let $f_1, \dots, f_k$ *be functions on* $\mathbb{R}$ *analytic on an open interval containing* $\left[aI_{(0,1)}(1/c)(1-\sqrt{1/c})^2, b(1+\sqrt{1/c})^2\right]$. *Write* $\Delta = E(z_{11}^4) - 3$, *then the random vectors*

$$n\left(\int f_1 \cdot d\mathcal{F}_{n1}, \ \dots, \int f_k \cdot d\mathcal{F}_{n1}\right) \xrightarrow{D} N_k(\mu, \Gamma_1),$$

$$\sqrt{n}\left(\int f_1 \cdot d\mathcal{F}_{n2}, \ \dots, \int f_k \cdot d\mathcal{F}_{n2}\right) \xrightarrow{D} N_k(0, \Gamma_2).$$

Li and Y. (2017)

- Notice that

  $\mathcal{F}_{n1} = F_n - F^{c_n, G_n}$ is "asymptotically independent" of $\mathcal{F}_{n2} = F^{c_n, G_n} - F^{c_n, G}$,

  which leads to a finite-sample corrected CLT

$$\sqrt{n}\left(\int f_1 \cdot d\mathcal{F}_n, \ \dots, \int f_k \cdot d\mathcal{F}_n\right) \dot{\sim} N_k(\mu/\sqrt{n}, \Gamma_1/n + \Gamma_2). \qquad (5.7)$$

Applications to empirical moments

- Example: For $\widehat{\beta}_{n2} = \sum_{j=1}^{p} \lambda_j^2/p$,

$$\sqrt{n}\left(\widehat{\beta}_{n2} - \beta_2\right) \dot{\sim} N\left(v_2/\sqrt{n}, \psi_{122}/n + \psi_{222}\right) \tag{5.8}$$

where the parameters are respectively

$$\beta_2 = c_n\gamma_2 + \gamma_1^2, \quad v_2 = (1+\Delta)\gamma_2,$$
$$\psi_{122} = 4((2+\Delta)\gamma_1^2\gamma_2/c + 8(2+\Delta)\gamma_1\gamma_2 + 4(\gamma_2^2 + c(2+\Delta)\gamma_4)),$$
$$\psi_{222} = c^2(\gamma_4 - \gamma_2^2) + 4c\gamma_1\gamma_3 + 4(1-c)\gamma_1^2\gamma_2 - 4\gamma_1^4.$$

Here, $\gamma_j = \int t^j dG(t)$ are the moments of the limiting mixing distribution G (not observed in a mixture !)

- Numerical results: PMD $G = 0.4\delta_1 + 0.6\delta_3$, $z_{ij} \sim \sqrt{1/6} \cdot (\chi_3^2 - 3)$.

Statistic	(p, n)	limiting distribution	correction
$\sqrt{n}(\widehat{\beta}_{n2} - \tilde{\beta}_2)$	(200,400)	$N(0, 39.32)$	$N(3.48, 48.88)$