Eigenvalues of large sample covariance matrices; Lecture 1

Jinho Baik
University of Michigan, Ann Arbor

February 2006

Goal: Find the limiting distribution of the largest eigenvalue of sample covariance matrix for so-called spiked population model as a way to illustrate a method ('Fredholm determinant method'?) from random matrix theory

- Introduction: sample covariance matrix, spiked population model.

- Algebraic part: eigenvalue density function, Fredholm determinant formula

- Asymptotic analysis: steepest-decent method, limiting distributions

- Differential equations for limiting distributions

- Other related models: traffic model, queues in tandem, last passage percolation

Population covariance and sample covariance

(complex) sample vector $\vec{x}$ of dimension p

normalized: mean 0, variance 1

population covariance matrix $T_p = \mathbb{E}(\vec{x}\vec{x}^*)$: $p \times p$ positive Hermitian

n samples, sample matrix $X = [\vec{x}_1, \cdots, \vec{x}_n]$

n= sample size
p= population size (dimension of vectors)

Sample covariance matrix

$$B_p := \frac{1}{n}[\vec{x}_1, \cdots, \vec{x}_n] \begin{bmatrix} \vec{x}_1^* \\ \vdots \\ \vec{x}_n^* \end{bmatrix} = \frac{1}{n} X X^*$$

$B_p(a,b) = \frac{1}{n}\sum_{j=1}^{n} \vec{x}_j(a)\vec{x}_j(b)$

Population and sample eigenvalues

population eigenvalues (true eigenvalues) $t_1^{(p)}, \ldots, t_p^{(p)}$

sample eigenvalues $s_1^{(p)} \geq \cdots \geq s_p^{(p)} > 0$

Is B_p a good approximate of T_p?

Are s_j's good approximate of t_j's?

$p << n$: yes

$p \sim n$: not so

[Marchenko+Pastur 1967] e.g. $T_p = I$
$n = n(p) \to \infty$, $\frac{p}{n} \to c$

$$\frac{1}{p}\#\{s_j^{(p)} : s_j^{(p)} < x\} \to F(x),$$

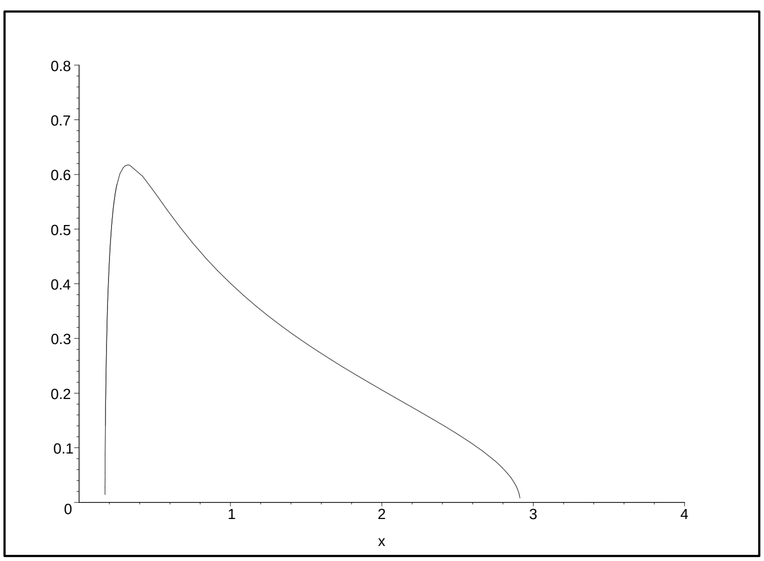
where

$$F'(x) = \frac{1}{2\pi x c}\sqrt{(b-x)(x-a)}, \quad a < x < b,$$

almost surely when $0 < c \le 1$. (mean 1, standard dev.$= \sqrt{1+c}$)

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

$c > 1$: Dirac measure at $x = 0$ of mass $1 - \frac{1}{c}$.



Marchenko-Pastur interval

$$I_{MP} := \left[(1-\sqrt{c})^2, (1+\sqrt{c})^2\right].$$

No stray sample eigenvalues!

$s_1^{(p)} \to (1+\sqrt{c})^2$ (Geman 1980)

$s_{\min\{p,n\}}^{(p)} \to (1-\sqrt{c})^2$ (Silverstein 1985)

($s_{n+1}^{(p)} = \cdots = s_p^{(p)} = 0$ when $n < p$)

Spiked population model (Johnstone)

T_p=finite-rank perturbation of I.

For some fixed r,

$$U_T T_p U_T^{-1} = diag(t_1, t_2, \ldots, t_r, 1, 1, 1, \ldots 1)$$

$n = n(p) \to \infty$, $\frac{p}{n} \to c$, r fixed

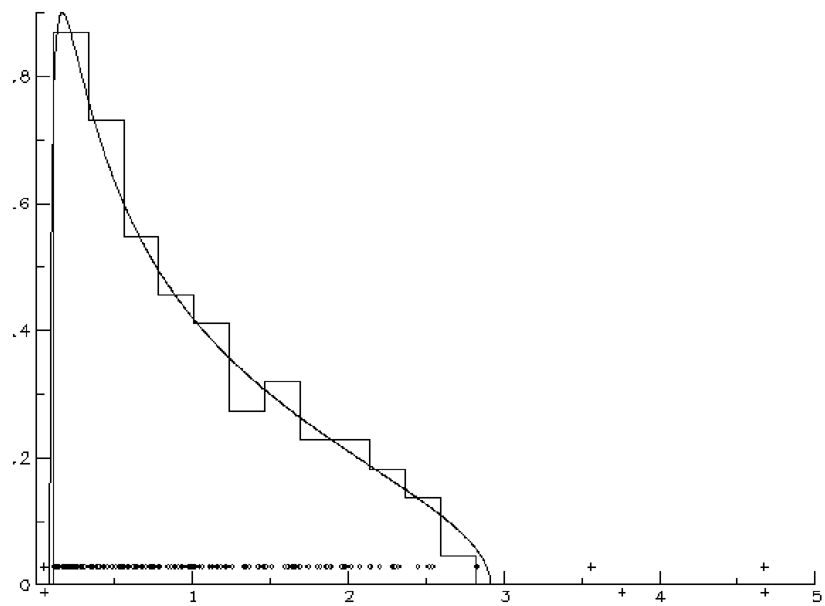
Limiting empirical distribution is same as before (Marchenko-Pastur)

But there may be some sample eigenvalues outside the Marchenko-Pastur interval.

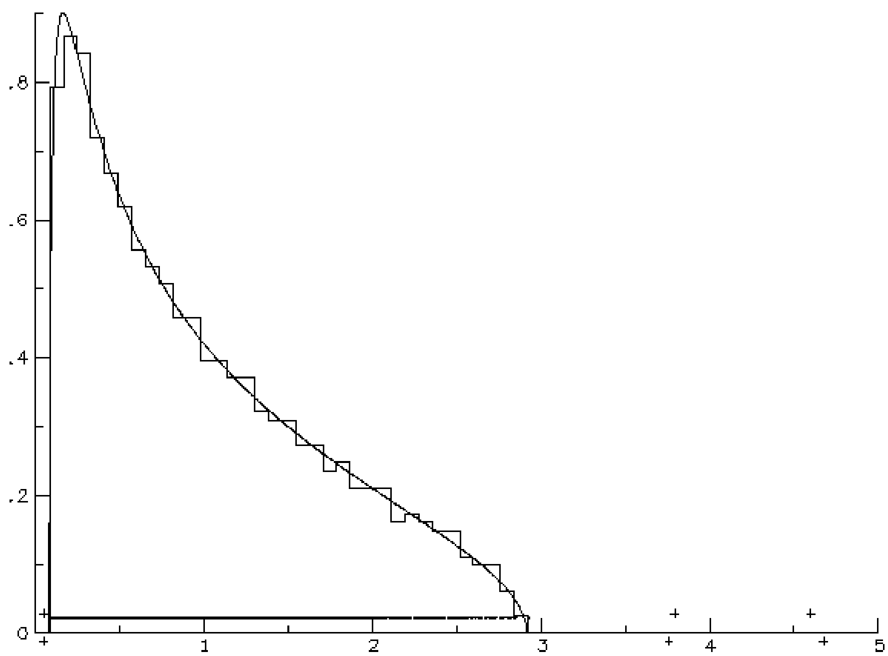
[Avellaneda + Park] Dynamic Risk Factor model for the dynamics of the cross correlation of a large financial system

Real Gaussian, 3 non-unit eigenvalues $\frac{1}{10}, 3, 4$

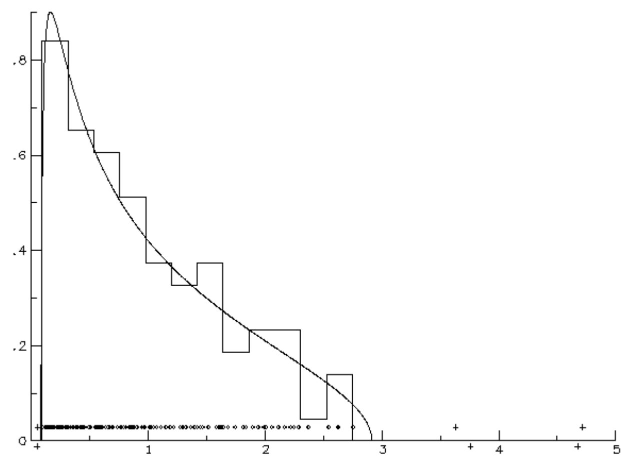
$p = 100$, $n = 200$



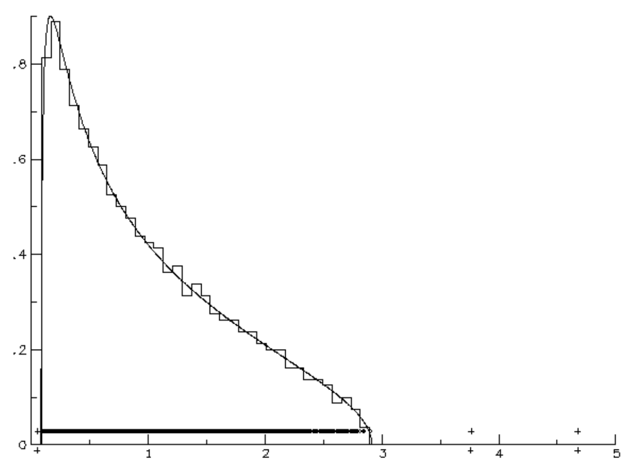
$p = 1000,\ n = 2000$



Bernoulli samples (values ± 1), $p = 100, n = 200$



Bernoulli $p = 1000, n = 2000$



Almost sure limits ($0 < c = \frac{p}{n} \leq 1$) [B.+Silverstein] (from [Bai+Silverstein])

Samples of form $\vec{x} = T_p^{1/2}\vec{z}$, entries of $\vec{z}$ are independent

Critical value of population eigenvalue $= 1 \pm \sqrt{c}$: population eigenvalues in $[1-\sqrt{c}, 1+\sqrt{c}]$ have no effect on sample eigenvalues.

To each population eigenvalue outside $[1-\sqrt{c}, 1+\sqrt{c}]$, there is one corresponding sample eigenvalue outside $I_{MP} = [(1-\sqrt{c})^2, (1+\sqrt{c})^2]$.

Examples ($c = \frac{1}{2}$): $1+\sqrt{c} \simeq 1.70711$, $1-\sqrt{c} \simeq 0.29289$

	$s_p^{(p)}$	$s_p^{(2)}$	$s_p^{(1)}$
theoretical	0.044	3.750	4.667
Gaussian $p = 1000$	0.044	3.784	4.591
Gaussian $p = 100$	0.040	3.554	4.662
Bernoulli $p = 1000$	0.046	3.757	4.666
Bernoulli $p = 100$	0.050	3.623	4.708

Limiting distributions: null case $T_p = I$

Complex Gaussian [Forrester 1993, Johansson 2000]

$$\lim_{n\to\infty} \mathbb{P}\left(\left(s_{\max} - (1+\sqrt{c})^2\right) \cdot \frac{c^{1/6} n^{2/3}}{(1+\sqrt{c})^{4/3}} \le x\right) = F_0(x)$$

for an explicit distribution function $F_0(x)$

(Note: F_0 is usually denoted by F_2 or F_{GUE}. Here we reserve F_2 for something else.)

Non-Gaussian rv's [Soshnikov 2002] ($c = 1$)

Real Gaussian [Johnstone 2001, Tracy+Widom 1996]: different limiting distribution

Goal: Spiked model with complex Gaussian samples. Determine the critical value of population eigenvalue. Find the limiting mean and limiting distribution function. What is the proper scaling ($n^{2/3}$ vs $n^{1/2}$)?

References

[Johnstone (2001) Ann. Stat.] Spiked models

[Péché (2003) Thesis] Complex Gaussian, $s_{\max}$, lower bound of critical value + limiting distributions

[Baik+Ben Arous+Péché (2004) Ann. Prob. 33] Complex Gaussian, $s_{\max}$, full phase transition, limiting distributions.

[Baik (2005) DMJ] Differential equations for limiting distributions.

[Baik+Silverstein (2004) JMVA] Real & complex, general rv, almost sure limits. [Z. Bai + Silverstein 1998, 1999]

[Paul 2004] Real Gaussian, above critical value, normal distribution for multiplicity 1 [Maida+Péché]

[Baik+Silverstein 2005] Real & complex, general rv, above critical value, limiting distribution for higher multiplicity

Things need to be done:

• limiting distribution of other rows for (sub-)critical case

• limiting distribution for other than complex Gaussian (e.g. real Gaussian) for (sub-)critical case

• other choices of T_p, such as $T_p = \begin{pmatrix} aI_{p/2} & 0 \\ 0 & bI_{p/2} \end{pmatrix}$ [Ben Arous+Péché] or 'random' T_p