Random matrix theory and contributions of Prof. P. L. Hsu

Zhidong Bai

An International Conference, *in memory to PL Hau's 100th birthday*,
Peking University, July 5–7, 2010

Academic Family Tree

Jerzy Neyman $\overset{(1938)}{\Longrightarrow}$ PL Hsu $\overset{(1958)}{\Longrightarrow}$ YQ Yin $\overset{(1982)}{\Longrightarrow}$ ZD Bai

WQ Liang $\overset{(1984)}{\Longrightarrow}$ KL Chung $\Longrightarrow$ PL Hsu

YQ Yin $\overset{(1982)}{\Longrightarrow}$ ZD Bai $\Longrightarrow$ RMT

Overview

Hsu's Publications

Publications results for "Items authored by Hsu, P. L."

MR0060453 (15,671g) Hsu, P. L. On symmetric, orthogonal, and skew-symmetric matrices. *Proc. Edinburgh Math. Soc. (2)* **10,** (1953). 37--44. (Reviewer: I. Reiner) 09.0X

MR0072373 (17,274c) Hsu, P. L. Absolute moments and characteristic functions. *J. Chinese Math. Soc. (N.S.)* **1** (1951), 257--280. (Reviewer: K. L. Chung) 60.0X

MR0062979 (16,53a) Hsu, P. L. A lemma on the coefficient of reduction of a sum of independent variates. *Acad. Sinica Science Record* **4,** (1951). 197--200. (Reviewer: K. L. Chung) 60.0X

MR0027998 (10,387f) Hsu, P. L. The limiting distribution of functions of sample means and application to testing hypotheses. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, 1945, 1946,* pp. 359--402. *University of California Press, Berkeley and Los Angeles,* 1949. (Reviewer: J. L. Doob) 62.0X

Hsu's Publications

MR0019852 (8,470e) Hsu, P. L.; Robbins, Herbert Complete convergence and the law of large numbers. *Proc. Nat. Acad. Sci. U. S. A.* **33,** (1947). 25--31. (Reviewer: R. Fortet) 60.0X

MR0017497 (8,161m) Hsu, P. L. On the asymptotic distributions of certain statistics used in testing the independence between successive observations from a normal population. *Ann. Math. Statistics* **17,** (1946). 350--354. (Reviewer: H. Cramér) 62.0X

MR0017266 (8,129c) Hsu, P. L. On a factorization of pseudo-orthogonal matrices. *Quart. J. Math., Oxford Ser.* **17,** (1946). 162--165. (Reviewer: J. Williamson) 09.0X

MR0016563 (8,36c) Hsu, P. L.; Chung, K. L. Sur un théorème de probabilités dénombrables. (French) *C. R. Acad. Sci. Paris* **223,** (1946). 467--469. (Reviewer: J. Wolfowitz) 60.0X

MR0013883 (7,212m) Hsu, P. L. On the power functions of the E^2-test and the T^2-test. *Ann. Math. Statistics* **16,** (1945). 278--286. (Reviewer: C. C. Craig) 62.0X

Hsu's Publications

MR0012384 (7,18j) Hsu, P. L. On the approximate distribution of ratios. *Ann. Math. Statistics* **16,** (1945). 204--210. (Reviewer: W. Feller) 60.0X

MR0011911 (6,233b) Hsu, P. L. The approximate distributions of the mean and variance of a sample of independent variables. *Ann. Math. Statistics* **16,** (1945). 1--29. (Reviewer: W. Feller) 60.0X

MR0008667 (5,43a) Hsu, P. L. Some simple facts about the separation of degrees of freedom in factorial experiments. *Sankhyā* **6,** (1943). 253--254. (Reviewer: H. B. Mann) 62.0X

MR0008666 (5,42i) Hsu, P. L. The limiting distribution of a general class of statistics. *Acad. Sinica Science Record* **1,** (1942). 37--41. (Reviewer: C. C. Craig) 62.0X

MR0005576 (3,174b) Hsu, P. L. On the limiting distribution of roots of a determinantal equation. *J. London Math. Soc.* **16,** (1941). 183--194. (Reviewer: C. C. Craig) 62.0X

MR0004444 (3,8c) Hsu, P. L. On the problem of rank and the limiting distribution of Fisher's test function. *Ann. Eugenics* **11,** (1941). 39--41. (Reviewer: S. S. Wilks) 62.0X

Hsu's Publications

MR0004443 (3,8b) Hsu, P. L. Canonical reduction of the general regression problem. *Ann. Eugenics* **11,** (1941). 42--46. (Reviewer: S. S. Wilks) 62.0X

MR0003548 (2,236c) Hsu, P. L. Analysis of variance from the power function standpoint. *Biometrika* **32,** (1941). 62--69. (Reviewer: J. Neyman) 62.0X

MR0003539 (2,234g) Hsu, P. L. On the limiting distribution of the canonical correlations. *Biometrika* **32,** (1941). 38--45. (Reviewer: S. S. Wilks) 62.0X

MR0002758 (2,111c) Hsu, P. L. On generalized analysis of variance. I. *Biometrika* **31,** (1940). 221--237. (Reviewer: S. S. Wilks) 62.0X

MR0002746 (2,109g) Hsu, P. L. An algebraic derivation of the distribution of rectangular coordinates. *Proc. Edinburgh Math. Soc. (2)* **6,** (1940). 185--189. (Reviewer: C. C. Craig) 62.0X

MR0001500 (1,248e) Hsu, P. L. On the distribution of roots of certain determinantal equations. *Ann. Eugenics* **9,** (1939). 250--258. (Reviewer: C. C. Craig) 62.0X

Hsu's first work

- Doogle.scholar: Citation=145 times,
- Web of Science citation=98 times,
- Almost every textbook on multivariate analysis cite this work, named Fisher-Hsu-Roy distribution
- Four people: RA Fisher, MA Girshick, PL Hsu and SN Roy, independently found it in 1939. (See "Random Matrix Theory and Wireless Communications" by A. Tulino and S. Verdu (2004)).
- R. A. Fisher, The sampling distribution of some statistics obtained from non-linear equations, *Annals of Eugenics,* **vol. 9**, pp. 238–249, 1939.
- M. A. Girshick, On the sampling theory of roots of determinantal equations, *Annals of Math. Statistics*, **vol. 10**, pp. 203–204, 1939.
- S. N. Roy, p-statistics or some generalizations in the analysis of variance appropriate to multivariate problems, *Sankhya: Indian J. of Statistics*, **vol. 4**, pp. 381–396, 1939.

Hsu's first work

The Setup

- Model: Eigenvalues θ defined by $|A - \theta(A+B)| = 0$, where
- $0 \leq \theta \leq 1$,
- $A = (a_{ij})$, $a_{ij} = \sum_{k=1}^{n_1} y_{ik} y_{jk}$
- $B = (b_{ij})$, $b_{ij} = \sum_{k=1}^{n_2} z_{ik} z_{jk}$
- Distribution $\mathrm{constant} \cdot \exp\left[-\frac{1}{2} \sum_{i,j=1}^{p} \alpha_{ij}(a_{ij} + b_{ij}) \right]$, where $(\alpha_{ij})^{-1}$ is the common nonsingular covariance matrix of both ys and zs.
- Assume $n_2 \geq p$, so that B is positive definite.
- Theorem Hsu 1. *The distribution of θs (in descending order) is*

$$C \prod_{i<j\leq p} (\theta_i - \theta_j) \left(\prod_{i=1}^{p} \theta_i \right)^{\frac{1}{2}(n_1 - p - 1)} \left(\prod_{i=1}^{p} (1 - \theta_i) \right)^{\frac{1}{2}(n_2 - p - 1)} \qquad \textit{if } p \leq n_1,$$

$$C_1 \prod_{i<j\leq n_1} (\theta_i - \theta_j) \left(\prod_{i=1}^{n_1} \theta_i \right)^{\frac{1}{2}(p - n_1 - 1)} \left(\prod_{i=1}^{n_1} (1 - \theta_i) \right)^{\frac{1}{2}(n_2 - p - 1)} \qquad \textit{if } p \geq n_1,$$

Comments on the work

- The distribution of θs consists of two part, one is the pdf of $Q = A(A+B)^{-1}$, $C|Q|^{\frac{1}{2}(n_1-p-1)}|I-Q|^{\frac{1}{2}(n_2-p-1)}$ and the other is P_1 where $P_\beta = \prod_{i<j\le\min(p,n_1)} |\theta_i - \theta_j|^\beta$, $\beta = 1$ (real), 2 (complex) and 4 (quarternion).
- This results was extended to general cases: **Proposition Hsu 1** *If the pdf of the matrix A is a function of its eigenvalues, i.e., $f(A) = g(\theta)$, then the pdf of eigenvalues of A is given by $P_1(\theta)g(\theta)$ up to a normalizing constant.*
- Further extension: **Proposition Hsu 2** *If the pdf of the matrix A is a function of its eigenvalues, i.e., $f(A) = g(\theta)$, then the pdf of eigenvalues of A is given by $P_\beta(\theta)g(\theta)$ up to a normalizing constant, for complex or quarternion cases.*
- This is even true for non-symmetric Gaussian matrix. (Ginibre (1965) For $n \times n$ matrix with iid complex Gaussian entries, the pdf of eigenvalues is

$$C \prod_{i<j\le n} |\lambda_i - \lambda_j|^2 \exp\left(-\sum_{i=1}^{n} |\lambda_i|^2\right).$$

(Circular Law (Mehta (1967)).)

Wigner 1955, 1958

Wigner Gaussian Matrix

- $\frac{1}{\sqrt{n}}X$, independent entries above and on the diagonal $x_{ii} \sim N(0,2)$ and $x_{ij} \sim N(0,1)$, $i<j$, symmetric or $x_{ii} \sim N(0,1)$ and $x_{ij} \sim CN(0,1)$ for Hermitian.
- pdf of eigenvalues, $\beta = 1$ or 2,

$$C_n \prod_{i<j\le n} (\lambda_i - \lambda_j)^\beta \exp\Big(-\frac{n\beta}{4}\sum_{i=1}^n \lambda_i^2\Big).$$

- $\prod_{i<j\le n}(\lambda_i - \lambda_j)$ is the Vandermonde determinant.
- If $\beta = 2$ (complex Wigner), no need to order the eigenvalues, then it is easy to derive the marginal pdf of λ_1 is given by $\left(\frac{1}{\sqrt{2\pi n}}\sum_{k=0}^{n-1} H_k^2(\sqrt{n}\lambda_1)\right)\exp(-n\lambda_1^2/2)$ (where H_k, the k-th normalized Hermit polynomial (orthonormal polynomials w.r.t. $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$), i.e. $H_k(x) = \frac{(-1)^k}{\sqrt{k!}}e^{x^2/2}\frac{d^k}{dx^k}e^{-x^2/2}$.)
- From here one can obtain the semicircular law whose pdf is $\dfrac{1}{2\pi}\sqrt{4-x^2}, \quad |x|<2.$

Marčenko and Pastur (1967)

Sample Covariance Matrix

- $\frac{1}{n}XX^*$, independent entries of X ($p \times n$) with $x_{ij} \sim N(0,1)$ or $CN(0,1)$.
- pdf of eigenvalues, $\beta = 1$ or 2,

$$C_n \prod_{i<j\leq r} (\lambda_i - \lambda_j)^\beta \prod_{i=1}^{r} \lambda_i^{(|n-p|-1)/2} \exp\left(-\frac{n\beta}{2} \sum_{i=1}^{r} \lambda_i \right).$$

  where $s = \max(n,p)$ and $r = \min(n,p)$.
- $\prod_{i<j\leq r}(\lambda_i - \lambda_j)$ is an $r \times r$ Vandermonde determinant.
- If $\beta = 2$ (complex Sample Covariance), no need to order the eigenvalues, then it is easy to derive the marginal pdf of λ_1 is given by $\left(\frac{n}{\sqrt{2\pi} r} \sum_{k=0}^{r-1} H_k^2(n\lambda_1)\right) \exp(-n\lambda_1/2)$ (where H_k, the k-th orthonormal polynomials w.r.t. $\varphi(x) = \frac{1}{2^{|n-p|/2}\Gamma(|n-p|/2)} x^{(|n-p|-1)/2} e^{-x/2}, \quad x > 0.$)
- By similar way, one can obtain the MP law as $p/n \to y > 0$. The pdf is given by $\frac{1}{2\pi xy}\sqrt{(b-x)(x-a)}, \quad a < x < b$, where $a, b = (1 \mp \sqrt{y})^2$. And a point mass $1 - 1/y$ at 0 when $y > 1$.

Tracy and Widom (1994, 1996)

Wigner Gaussian Matrix

- Bai and Yin (1988) showed that (under 4th moment)
- $\lambda_{\max}(n^{-1/2}W) \to 2, \lambda_{\min}(n^{-1/2}W) \to -2,\ a.s.$
- Tracy and Widom (1994,1996,2000)showed that (under normality)
- $n^{2/3}(\lambda_{\max}(n^{-1/2}W) - 2) \Rightarrow$ TW$_\beta$ Law whose distribution is given by

$$\begin{aligned} F_2(x) &= \exp\left(-\int_x^\infty (t-x)q^2(t)dt \right), \\ F_1(x) &= \exp\left(-\frac{1}{2}\int_x^\infty q(t)dt \right)[F_2(x)]^{1/2}, \\ F_4(2^{-1/2}x) &= \cosh\left(-\frac{1}{2}\int_x^\infty q(t)dt \right)[F_2(x)]^{1/2} \end{aligned}$$

  where $q(t)$ satisfies the ODE $q'' = tq + 2q^3$ with the marginal condition $q(t) \sim Ai(t), \quad$ as $t \to \infty$.
- The proof uses the exact pdf eigenvalues of Gaussian Wigner matrix.

TW law for Sample Covariance Matrix

Johnstone (2001) and Johansson (2000)

- Bai and Yin (1988) showed that (under 4th moment)
- $\lambda_{\max}(n^{-1/2}W) \to 2, \lambda_{\min}(n^{-1/2}W) \to -2,\ a.s.$
- Tracy and Widom (1994,1996,2000)showed that (under normality)
- $n^{2/3}(\lambda_{\max}(n^{-1/2}W) - 2) \Rightarrow$ TW$_\beta$ Law whose distribution is given by

$$\begin{aligned} F_2(x) &= \exp\left(-\int_x^\infty (t-x)q^2(t)dt\right), \\ F_1(x) &= \exp\left(-\frac{1}{2}\int_x^\infty q(t)dt\right)[F_2(x)]^{1/2}, \\ F_4(2^{-1/2}x) &= \cosh\left(-\frac{1}{2}\int_x^\infty q(t)dt\right)[F_2(x)]^{1/2} \end{aligned}$$

  where $q(t)$ satisfies the ODE $q'' = tq + 2q^3$ with the marginal condition $q(t) \sim Ai(t), \quad \text{as } t \to \infty$.
- The proof uses the exact pdf eigenvalues of Gaussian Wigner matrix.

Some References

- K. Johannson (2002) Non-Intersecting Paths, Random Tilings and Random Matrices *Probab. Theory and Related Fields*, **123**, 225–280.
- K. Johansson (2001) Universality of the Local Spacing Distribution in Certain Ensembles of Hermitian Wigner Matrix. *Commun. In Math. Physics.* **215**, 683–705.
- S. Albeverio, L. Pastur, M. Shcherbina (2001) On the $1/n$ expansion for Some Unitary Invariant Ensembles of Random Matrices. *Commun. Mth. Physics*, **224**, 271–305.
- A. Edelman (2005) Random matrix theory, *Acta Numerica*, pp. 233–297
- More than 1000 paper, not listed.

Moment Method

Moment Convergence Theorem

- Step 1. Truncation.
  For statistical purposes, it's better to assume entries having certain finite moments, instead of normality. Thus, truncation is necessary for all results. Generally, truncate variables at $\eta_n \sqrt{n}$ in various cases.
- Step 2. Re-normalization.
  After truncation, we need to centralize and normalize the truncated variables to the standard conditions.
- Step 3. Moment Convergence and Summability of Variances.
  After truncation, we can compute all moments of the ESD, which is $\mu_{nk} = \frac{1}{n}\mathrm{tr}(W_n^k)$. One needs to show that

  $$\mathrm{E}(\mu_{nk}) \to \mu_k \quad \sum_{n=N}^{\infty} \mathrm{Var}(\mu_{nk}) < \infty.$$

  Then, $\nu_{nk} \to \mu_k,\ a.s.$
- Conclude the LSD exists, by Verifying the Carelman Condition $\sum_k \mu_{2k}^{-1/(2k)} = \infty$.

Stieltjes Transformation

- Same first two steps, truncation and normalization.
- Step 3. Convergence of Stieltjes transforms.
- $s_n(z) = \frac{1}{n}\mathbf{tr}(W - z\mathbf{I})^{-1} = \int \frac{1}{x-z} dF_n(x)$,
  $z = u + iv$, $v > 0$, i.e. $z \in \mathbb{C}^+$.
- Show that $s_n(z) \to s(z)$ for each fixed $z \in \mathbb{C}^+$ in some sense, a.s. or in prob..
- Show that $s_n(z) \to s(z)$ for all $z \in \mathbb{C}^+$ in some sense, a.s. or in prob., by using Vitali Lemma.
- If $s(z)$ has an explicit expression, then the density of LSD is
  $p(u) = \frac{1}{\pi} \lim_{v \downarrow 0} s(u + iv)$.

Methodologies

Martingale Decomposition

- $s_n(z) - \mathrm{E}(s_n(z))$ be expressed as

$$\frac{1}{n}\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})\mathrm{tr}(W - zI)^{-1},$$

  by conditional expectations E_k with suitably defined sub-sigma fields.
- Then treat the random part and non-random parts separately.

Moment Condition

LSD, Semicircular Law

- Wigner Matrix, $W = \frac{1}{\sqrt{n}}(x_{ij})$, $n \times n$, Symmetric or Hermitian,
- The iid (off diagonal entries) case, variance 1, any means, any distributions for diagonal elements, limit to semicircular law, i.e.

$$F'(x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \text{ if } |x| < 2.$$

- non-iid case, mean 0 variances 1 and Lindeberg condition:

$$\frac{1}{n^2}\sum_{ij} \mathrm{E}|x_{ij}^2| I(|x_{ij}| > \epsilon\sqrt{n}) \to 0,$$

  for any $\epsilon > 0$.
- Limiting moments $\mu_{2k} = \frac{1}{k+1}\binom{2k}{k}$ and $\mu_{2k+1} = 0$.
- Limiting Stieltjes transform $s(z) = -\frac{1}{2}(z - \sqrt{z^2 - 4})$, where the square roots of a complex number having positive imaginary part.

Moment Conditions

LSD, MP Law

- Sample Covariance Matrix, $\mathbf{S} = \frac{1}{n}(\sum_{k=1}^{n} x_{ik}\bar{x}_{jk})$, $p \times p$, Symmetric or Hermitian, with $p/n \to y > 0$.
- The iid case, variance 1, any means, limit to MP law, i.e.

$$F_y'(x) = \frac{1}{2xy\pi}\sqrt{(b-x)(x-a)}, \text{ if } a < x < b,$$

  where $a, b = (1 \mp \sqrt{y})^2$,
- F_y has a point mass $1 - 1/y$ if $y > 1$.
- non-iid case, mean 0 variances 1 and Lindeberg condition:

$$\frac{1}{np}\sum_{ij} \mathrm{E}|x_{ij}^2| I(|x_{ij}| > \epsilon\sqrt{n}) \to 0,$$

  for any $\epsilon > 0$.
- Limiting moments $\mu_k = \sum_{r=0}^{k-1} \frac{y^r}{r+1}\binom{k}{r}\binom{k-1}{r}$.
- Limiting Stieltjes transform $s(z) = \frac{1-y-z+\sqrt{(1+y-z)^2-4y}}{2yz}$.

Moment Conditions

General Sample Covariance Matrix

- Sample Covariance Matrix, $\mathbf{B} = \mathbf{T}^{1/2}\mathbf{S}\mathbf{T}^{1/2}$, $p \times p$, Symmetric or Hermitian, with $p/n \to y > 0$.
- ESD of $\mathbf{T}$ tends to a limit H.
- The xs are iid case, variance 1, means 0,
- LSD exists and its Stieltjes transform is given by the unique solution of

$$z = -\frac{1}{\underline{s}} + y \int \frac{t dH(t)}{1 - t\underline{s}},$$

  where $\underline{s} = -\frac{1-y}{z} + ys$.
- Jack Silverstein (1995) Strong Convergence of the empirical spectral distribution of large dimensional random matrices, *J. Multivariate Anal.* **5**, 331–339.

Spectrum Separation, Extreme Eigenvalues

N&S Conditions

- Both Wigner and Sample Covariance Matrix, Extreme Eigenvalues tend to boundaries of Spectrum under 4-th finite Moments. N&S Conditions are given.
- Exact Separation. Extension for limits of extreme eigenvalues. Useful for Signal processing.
- References:
  Bai, Z. D. and Yin, Y. Q. (1988) Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of Wigner matrix. *Ann. Probab.*, 16, 4, 1729-1741.
- Bai, Z. D., Yin, Y. Q. (1993) Limit of the smallest eigenvalue of large dimensional covariance matrix. *Ann. Probab.* Vol. 21, No. 3, 1275-1294.
- Z. D. Bai and J. W. Silverstein (1998) No eigenvalues outside the support of the limiting spectral distribution of large dimensional sample covariance matrices. *Ann. Probab.* **Vol. 26**, No. 1, 316-345.
- Bai, Z.D. and Silverstein, Jack W. (1999) Exact separation of eigenvalues of large dimensional sample covariance matrices. *Ann. Probab.* **Vol. 27**, No. 3, 1536-1555.

Convergence Rates, ESD

Wigner Matrix

- Under 6-th moment
$$\|\mathrm{E}F_n - F_{semi}\| \leq C/\sqrt{n}.$$

- Under the same conditions
$$\|F_n - F_{semi}\| = O_p(n^{-2/5}) \quad \|F_n - F_{semi}\| = O_{a.s.}(n^{-2/5+\eta}) \quad \forall \eta > 0.$$

- Z. D. Bai, B. Q. Miao and J. Tsay (2002) Convergence rates of the spectral distributions of large Wigner matrices. *International Mathematical Journal*. **Vol. 1, No. 1**. 65-90.
- Bai, Z. D. (1993) Convergence rate of Expected spectral distributions of large random matrices. Part I. Wigner Matrices. *Ann. Probab.* Vol. 21, No. 2, 625-648.

Convergence Rates, ESD

Sample Covariance Matrix

- Under 6-th moment

$$\|\mathrm{E}F_n - F_y\| \le \begin{cases} O(n^{-1/2}a^{-1}) & \text{if } a > n^{1/3}, \\ O(n^{-1/6}) & \textit{otherwise}. \end{cases}$$

- Under the same conditions

$$\|F_n - F_y\| \le \begin{cases} O_p(n^{-2/5}a^{-2/5}) & \text{if } a > n^{1/3}, \\ O_{a.s}(n^{-2/5}a^{-2/5+\eta}) & \text{if } a > n^{1/3}, \\ O_{a.s}(n^{-1/6}) & \textit{otherwise}. \end{cases}$$

- Bai, Z. D., Miao, B. Q. and Yao, J. F. (2003) Convergence rates of spectral distributions of large sample covariance matrices *SIAM J. Matrix Anal. Appl.* **Vol. 25, No. 1,** pp. 105-127
- Bai, Z. D. (1993) Convergence rate of expected spectral distributions of large random matrices. Part II. Sample Covariance Matrices. *Ann. Probab.* Vol. 21, No. 2, 649-672.

Bai and Silverstein's CLT for linear spectral statistics of S_n

Set

- the Empirical spectral distribution:

$$F_n = \frac{1}{p}\sum_{j=1}^{p} \delta_{\lambda_j},$$

  where λ_j's are p eigenvalues of S_n;
- $y_n = \frac{p}{n}$;
- $[a, b] \subset \mathcal{U} \text{ open } \subset \mathbb{C}$.
- for any g analytic on $\mathcal{U}$

$$G_n(g) = p\left[F_n(g) - \mu^{y_n}(g)\right]$$

  where μ^{α} is the MP distribution of index $\alpha \in (0, 1)$.

A CLT for linear spectral statistics

Bai and Silverstein, '04

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* $n, p \to \infty$, $y_n = \dfrac{p}{n} \to y \in (0, 1)$*;*

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* $V = V(g_1, \ldots, g_k)$.

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:
$$S_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_i \mathbf{x}_i^*, \qquad S_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{y}_j \mathbf{y}_j^*.$$
- Fisher matrix: $V_n = S_1 S_2^{-1}$ where $n_2 > p$.

Random Fisher matrices

- 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:
$$\ell(x) = \begin{cases} \dfrac{(1-y_2)\sqrt{(b-x)(x-a)}}{2\pi x(y_1+y_2x)}, & a \le x \le b, \\ 0, & \text{otherwise} \end{cases}$$
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
$$\left[I_{(0,1)}(y_1)\frac{(1-\sqrt{y_1})^2}{(1+\sqrt{y_2})^2}, \quad \frac{(1+\sqrt{y_1})^2}{(1-\sqrt{y_2})^2} \right] \subset \widetilde{\mathcal{U}} \ \text{ open } \quad \subset \mathbb{C},$$
- for an analytic function f on $\widetilde{\mathcal{U}}$, define
$$\widetilde{G_n}(f) = p \cdot \int_{-\infty}^{+\infty} f(x) \left[F_n^{V_n} - F_{y_{n_1}, y_{n_2}} \right] (dx) \ ,$$
where $F_{y_{n_1}, y_{n_2}}$ is the LSD with indexes y_{n_k}, $k = 1, 2$.

CLT for LSS of random Fisher matrices

Zheng, '08

Theorem

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

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

with suitable asymptotic mean and covariance functions m *and* υ.

CLT for LSS of random Fisher matrices

Zheng, '08

Limiting mean function m

$$
\begin{aligned}
m(f_j) &= \lim_{r \to 1_+} [(1) + (2) + (3)] \\
&\frac{1}{4\pi i} \oint_{|\zeta|=1} f_j(z(\zeta)) \left[\frac{1}{\zeta - \frac{1}{r}} + \frac{1}{\zeta + \frac{1}{r}} - \frac{2}{\zeta + \frac{y_2}{hr}} \right] d\zeta \qquad (1) \\
&+ \frac{\beta \cdot y_1(1-y_2)^2}{2\pi i \cdot h^2} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{1}{(\zeta + \frac{y_2}{hr})^3} d\zeta \qquad (2) \\
&+ \frac{\beta \cdot y_2(1-y_2)}{2\pi i \cdot h} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{\zeta + \frac{1}{hr}}{(\zeta + \frac{y_2}{hr})^3} d\zeta, \qquad (3)
\end{aligned}
$$

where

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

CLT for LSS of random Fisher matrices

Zheng, '08

Limiting covariance function υ

$$\upsilon(f_j, f_\ell) = \lim_{1<r_1<r_2\to 1_+} [(4)+(5))]$$

$$-\frac{1}{2\pi^2}\oint_{|\zeta_2|=1}\oint_{|\zeta_1|=1}\frac{f_j(z(r_1\zeta_1))f_\ell(z(r_2\zeta_2))r_1r_2}{(r_2\zeta_2-r_1\zeta_1)^2}d\zeta_1 d\zeta_2, \tag{4}$$

$$-\frac{\beta\cdot(y_1+y_2)(1-y_2)^2}{4\pi^2h^2}\oint_{|\zeta_1|=1}\frac{f_j\left(z(\zeta_1)\right)}{(\zeta_1+\frac{y_2}{hr_1})^2}d\zeta_1\oint_{|\zeta_2|=1}\frac{f_\ell\left(z(\zeta_2)\right)}{(\zeta_2+\frac{y_2}{hr_2})^2}d\zeta_2 \tag{5}$$

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

One-sample test on covariance matrices

- a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid.}{\sim} \mathcal{N}_p(\mu, \Sigma)$
- want to test $H_0 : \ \Sigma = I_p$
- LR statistic:
$$T_n = n\,[tr S_n - \log|S_n| - p]$$
with the sample covariance matrix
$$S_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

Classical LRT:

- Data dimension p is fixed, and when $n \to \infty$, $T_n \Longrightarrow \chi^2_{p(p+1)/2}$.
- Will see: rapidly deficient when p is not "small".

RMT Corrected LRT:

Bai, Jiang, Y and Zheng '09

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*

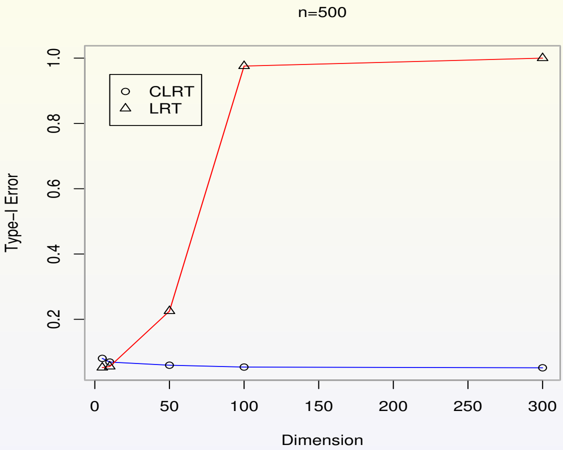
$$\begin{aligned} m(g) &= -\frac{\log(1-y)}{2}, \\ \upsilon(g) &= -2\log(1-y) - 2y. \end{aligned}$$

Comparison of LRT and CLRT 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
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
Dimension

Two-samples test on covariance matrices

- two samples
  $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1} \sim \mathcal{N}_p(\mu_1, \Sigma_1)$, $\quad \mathbf{y}_1, \ldots, \mathbf{y}_{n_2} \sim \mathcal{N}_p(\mu_2, \Sigma_2)$
- want to test $H_0 : \ \Sigma_1 = \Sigma_2$
- The associated sample covariance matrices are

$$S_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})', \quad S_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})',$$

- Let the LR statistic

$$L_1 = \frac{\left| S_1 S_2^{-1} \right|^{\frac{n_1}{2}}}{\left| c_1 S_1 S_2^{-1} + c_2 I_p \right|^{\frac{n}{2}}},$$

  where $n = n_1 + n_2$ and $c_k = \frac{n_k}{n}, k = 1, 2.$

Two-samples test on covariance matrices

Classical LRT:

- Data dimension p is fixed, and when $n_1, n_2 \to \infty$ and under H_0 ,

$$T_n = -2 \log L_1 \Rightarrow \chi^2_{p(p+1)/2} \ .$$

- Will see: rapidly deficient when p is not “small”.

RMT Corrected LRT:

Bai, Jiang, Yao and Zheng '09

Theorem

Assuming that the conditions of CLT for LSS of Fisher matrices hold and let

$$f(x) = \log(y_1 + y_2 x) - \frac{y_2}{y_1 + y_2} \log x - \log(y_1 + y_2).$$

Then under H_0 and as $n_1 \wedge n_2 \to \infty$,

$$\left[-\frac{2 \log L_1}{n} - p \cdot F_{y_{n_1}, y_{n_2}}(f) \right] \Rightarrow \mathcal{N}\left(m(f), \upsilon(f)\right),$$

with

$$m(f) = \frac{1}{2} \left[\log\left(\frac{y_1 + y_2 - y_1 y_2}{y_1 + y_2} \right) - \frac{y_1}{y_1 + y_2} \log(1 - y_2) - \frac{y_2}{y_1 + y_2} \log(1 - y_1) \right],$$

$$\upsilon(f) = -\frac{2y_2^2}{(y_1 + y_2)^2} \log(1 - y_1) - \frac{2y_1^2}{(y_1 + y_2)^2} \log(1 - y_2) - 2 \log \frac{y_1 + y_2}{y_1 + y_2 - y_1 y_2}.$$

Comparison of LRT and CLRT by simulation

- nominal test level $\alpha = 0.05$;
- for each (p, n_1, n_2), 10,000 independent replications with real Gaussian variables.
- Powers are estimated under the alternative H_1: $\Sigma_1\Sigma_2^{-1} = \mathrm{diag}(3, 1, 1, \cdots,)$.

Comparison of LRT and CLRT by simulation

with $(y_1, y_2) = (0.05, 0.05)$:

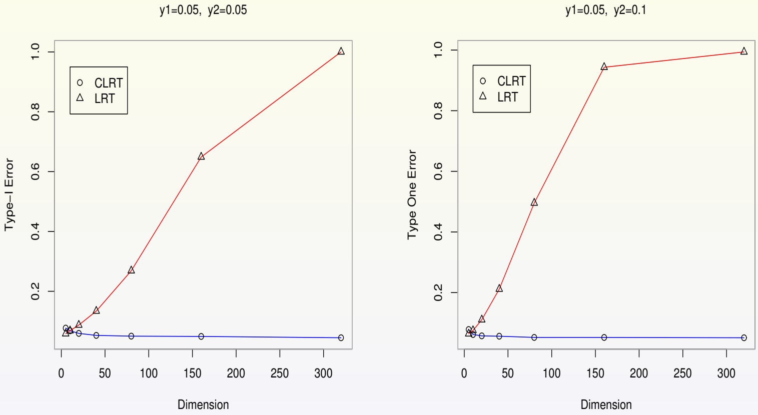
	CLRT			LRT	
(p, n_1, n_2)	Size	Difference with 5%	Power	Size	Power
(5, 100, 100)	0.0770	0.0270	1	0.0582	1
(10, 200, 200)	0.0680	0.0180	1	0.0684	1
(20, 400, 400)	0.0593	0.0093	1	0.0872	1
(40, 800, 800)	0.0526	0.0026	1	0.1339	1
(80, 1600, 1600)	0.0501	0.0001	1	0.2687	1
(160, 3200, 3200)	0.0491	-0.0009	1	0.6488	1
(320, 6400, 6400)	0.0447	-0.0053	0.9671	1	1

Comparison of LRT and CLRT by simulation

with $(y_1, y_2) = (0.05, 0.1)$:

	CLRT			LRT	
(p, n_1, n_2)	Size	Difference with 5%	Power	Size	Power
(5, 100, 50)	0.0781	0.0281	0.9925	0.0640	0.9849
(10, 200, 100)	0.0617	0.0117	0.9847	0.0752	0.9904
(20, 400, 200)	0.0573	0.0073	0.9775	0.1104	0.9938
(40, 800, 400)	0.0561	0.0061	0.9765	0.2115	0.9975
(80, 1600, 800)	0.0521	0.0021	0.9702	0.4954	0.9998
(160, 3200, 1600)	0.0520	0.0020	0.9702	0.9433	1
(320, 6400, 3200)	0.0510	0.0010	1	0.9939	1

Comparisons of LRT and CLRT



Some references

- Bai, Z. D. and Saranadasa, H. (1996). Effect of high dimension comparison of significance tests for a high dimensional two sample problem. *Statistica Sinica.* **6**, 311-329.
- Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann.Probab.* **32**, 553-605.
- Z. D. Bai, D. Jiang, J. Yao and S. Zheng, 2009. Corrections to LRT on Large Dimensional Covariance Matrix by RMT. *Annals of Statistics* **37**, 38223840
- Z. D. Bai, D. Jiang, J. Yao and S. Zheng, 2009. On Wilk's test in MANOVA with high-dimensional data. *submitted*
- Dempster, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Statist.* **29**, 995-1010.
- Zheng, S. (2008). Central Limit Theorem for Linear Spectral Statistics of Large Dimensional F Matrix. *Preprint, Northern-Est Normal University*

Eigenvector Matrix

- Hard to formulate, because p increases.
- Few References, only 5 by Silverstein and 1 by Bai, Miao and Pan (2007)
- Haar Conjecture:
- One description:
- $\mathbf{x} \in \mathbb{R}^p$, $\|\mathbf{y}\| = 1$. Define

$$\mathbf{y} = \mathbf{U}\mathbf{x} = (y_1, ..., y_p)$$

and

$$X_n(t) = \sqrt{p} \sum_{k=1}^{[pt]} (|y_k^2| - \frac{1}{p})$$

- $X_n(t) \to BB(t)$? (Brownian bridge)
- CLT: $\int f(x) dX_n(F_n(x)) \to$ Normal distribution.
- Towards Haar conjecture.

Some references

- Bai, Z. D., Miao, B. Q., and Pan, G. M. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab.* **35**(4), 1532–1572.
- Silverstein, J. W. (1990). Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. *Ann. Probab.* **18**, 1174–1194.
- Silverstein, J. W. (1989). On the eigenvectors of large dimensional sample covariance matrices. *J. Multivariate Anal.* **30**, 1–16.
- Silverstein, J. W. (1984). Some limit theorems on the eigenvectors of large dimensional sample covariance matrices. *J. Multivariate Anal.* **15**, 295–324.
- Silverstein, J. W. (1981). Describing the behavior of eigenvectors of random matrices using sequences of measures on orthogonal groups. *SIAM J. Math. Anal.* **12**, 174–281.
- Silverstein, J. W. (1979). On the randomness of eigenvectors generated from networks with random topologies *SIAM J. Appl. Math.* **37**, 235–245.

Spiked Eigenvalues

- a few eigenvalues out of the Bulk eigenvalues.
- limits and CLT.
- TW law, universality.
- Spiked eigenvectors.
- Estimation of true Spikes (Not yet).

Bai, Zhidong; Yao, Jian-feng (2008) Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincarè Probab. Stat.* **44, No. 3**, 447–474.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**(2), 295–327.

Baik, J., Ben Arous, G., and Péché, S. (2005). Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. *Ann. Probab.* **33**, 1643–1697.

Corrections to LRT statistics

- Testing equality of Mean vectors.
- Testing equality of Covariance Matrices.
- Testing on PCA, Factor Analysis.
- Clustering.
- Any Statistics involving $\mathbf{S}^{-1}$.

Bai, Z. D., Jiang, Dandan, Yao, and Zheng, Shurong (2009) Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* **Vol. 37, No. 6B**, 3822–3840.

Bai, Z. D., Jiang, Dandan, Yao, and Zheng, Shurong (2009) Large Regression Analysis. Submitted.

Corrections to Plug-in rules

- Plug-in rule causes intolerable over-prediction.
- Corrections by MRT and Bootstrap.
- Corrections by new estimation of population Eigenvalues.

Bai, Z. D., Liu, H. X., and Wong, W. K. (2009) Enhancement of the applicability of markowitz's portfolio optimization by utilizing random matrix theory. *Mathematical Finance*, **Vol. 19, No. 4**, 639–667.

Bai, Z. D., Liu, Huixia and Wong, Wing-Keung (2009) On the Markowitz men-variance analysis of self-financing portfolio. *Risk and Decision Analysis* **1**, 35-42.

Thanks!