Large Dimensional Random Matrices and High Dimensional Statistics

CONFERENCE ON STATISTICS AND STOCHASTIC PROCESSES
IN HONOUR OF
RAJEEVA L. KARANDIKAR
MARCH 29-31, 2022

ARUP BOSE

Indian Statistical Institute, Kolkata

March 29, 2022

Examples: fixed and high dimension

Contingency table.

Random design matrix.

Sample covariance matrix/Wishart matrix (1928).

Wigner matrix (Anderson, Guionnet and Zeitouni book, CUP (2009)).

Random Toeplitz and Hankel matrices (Bose book CRC (2018)).

IID matrix (Bordanave and Chafai survey article, Prob. Surveys (2012)).

β matrices (Forrester book, LMS (2010)).

Autocovariance matrices (Bose and Bhattacharjee book CRC (2018)).

Some others will be introduced later.

Sample Covariance Matrix

Data matrix: $X = ((X_{ij}))_{1 \le i \le p, 1 \le j \le n}$. [$p$ is the dimension of each observation and n is the number of observations.]

The (unadjusted) $p \times p$ *Sample Covariance matrix* S is a key matrix in multivariate analysis.

$$S = n^{-1} X X^*.$$

When the entries are Gaussian (normal), S is called a Wishart matrix. The joint density of the *eigenvalues* can then be calculated and used for statistical inference. [Recall principal component analysis].

When p is fixed and n is large, CLT arguments show that *joint asymptotic normality* of the eigenvalues of S holds.

If p is also large, the behavior of the eigenvalues is very different, and standard statistical tests break down.

Autocovariance Matrix: real-valued case

Suppose $X_1, X_2, \ldots, X_n$ is a real-valued mean zero *stationary stochastic process*. Then the *autocovariance sequence* is defined as

$$\gamma(i) = E(X_t X_{t+i}),\ i = 0, 1, \ldots.$$

One can consider the *dispersion matrix* of $(X_1, \ldots X_n)$:

$$\Sigma_n = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \cdots & \cdots & \gamma_{n-2} & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \cdots & \cdots & \gamma_{n-3} & \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \cdots & \cdots & \gamma_{n-4} & \gamma_{n-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{n-3} & \gamma_{n-4} & \gamma_{n-5} & \cdots & \cdots & \cdots & \gamma_1 & \gamma_2 \\ \gamma_{n-2} & \gamma_{n-3} & \gamma_{n-4} & \cdots & & \cdots & \gamma_0 & \gamma_1 \\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots & & & \gamma_1 & \gamma_0 \end{bmatrix}_{n \times n}$$

Sample Autocovariance Matrix: real-valued case

Replace the unknown γ_i by their estimates, say

$$\hat{\gamma}(i) = n^{-1} \sum_{t=1}^{n-i} X_t X_{t+i}.$$

It is well known that $\{\hat{\gamma}(i)\}$ are fundamental in any kind of statistical analysis of a real-valued stationary time series.

The *sample auto-covariance matrix* is given by,

$$\Gamma_n(X) = ((\hat{\gamma}_{|i-j|}))_{n \times n}$$

It is a symmetric Toeplitz random matrix.

Large dimensional random matrices (LDRM)

• Large dimensional matrices come up naturally in high dimensional models.

$\{X_t\}$ are p-dimensional observations, $1 \leq t \leq n$. Both p and n are assumed to be large. Then the S matrix is large dimensional.

• Usual procedures do not perform well in high dimensions.

• So, understanding the properties of LDRMs is important

–to understand the reasons for failure of the standard procedures, and

–to devise new methods for statistical analysis.

Autocovariance matrix sequence

For a real-valued series, we had a sequence $\{\gamma_l = E(X_t X_{t-l})\}$ of autocovariances. If $X_{t,p}$ is a p-dimensional time series, then for every i, we get a $p \times p$ matrix

$$\Gamma_i = E(X_{t,p} X'_{t-i,p}).$$

The sample autocovariance matrix of order i is the $p \times p$ matrix

$$\hat{\Gamma}_{i,p} := \frac{1}{n} \sum_{t=i+1}^{n} X_{t,p} X'_{(t-i),p}, i = 0, 2, 3, \ldots, (n-1).$$

Except for $i = 0$, all the matrices are *non-symmetric*. **It is notoriously hard to analyse non-symmetric random matrices**. Usually one resorts to symmetrisation such as:

$$\hat{\Gamma}_{i,p} + \hat{\Gamma}^*_{i,p} \quad \text{or} \quad \hat{\Gamma}_{i,p} \hat{\Gamma}^*_{i,p}.$$

What does one study?

A matrix is a linear transformation. Its features are captured by its *eigenvalues*.

• Bulk behaviour of eigenvalues (limit spectral distributions such as the semi-circle and the Marchenko-Pastur laws).

• Extreme eigenvalues (Tracy-Widom and Gumbel laws).

• Gaps between eigenvalues (determinantal point processes).

• Joint behaviour of random matrices (free independence).

–Use the knowledge of these aspects to develop statistical procedures to analyse high-dimensional data (**need more tools and techniques**).

Bulk: Empirical Spectral Distribution (ESD)

R_p: a $p \times p$ random matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$.

Empirical Spectral Distribution (ESD) of R_p is the *random probability distribution* that puts probability $1/p$ at each eigenvalue:

$$P(X = \lambda_i) = 1/p, \; i = 1, \ldots, p.$$

Its cumulative form (when all eigenvalues are real) is given by

$$ECDF(x) = \frac{\text{Number of eigenvalues} \leq x}{p}.$$

We can also consider the *histogram* of these eigenvalues.

Bulk: Limiting Spectral Distribution (LSD)

Limiting spectral distribution (LSD): If as $p \to \infty$, this ESD converges weakly (almost surely or in probability) to a probability distribution, then the limit is called the LSD. Often the limit is random. If the expectation of the ESD converges, that is called the convergence of the EESD and this limit is also called the LSD.

Loosely speaking, if the histogram stablizes as $p \to \infty$, then the limit curve defines the LSD.

This is *convergence in the bulk* since any finite number of individual eigenvalues do not matter for the limit.

Some LSD results

Wigner (1955) [semi-circle law],

Sample covariance (1967) when $p/n \to y$, $0 < y < \infty$. [Marchenko-Pastur law with parameter y], MP_y.

Patterned matrices [Toeplitz, Hankel, circulant..] (2005...)

IID (2010) [uniform law on the unit disc],

Sample autocovariance matrix for real-valued time series (2014) [LSD is a function (but not 1-1) of the spectral density]

Symmetric polynomials of sample autocovariance matrices (2016) [LSD as functions of free semi-circle and other variables]

LSD results are easier to establish for hermitian matrices.

Results on the LSD of non-hermitian matrices are much less in number and there are many such interesting matrices for which LSD results are not available.

Spectral statistics

If $\{\lambda_k\}$ are the eigenvalues of a matrix M, then $\sum f(\lambda_k)$ where f is a suitable function, are called *linear spectral statistics* (LSS). These serve as summaries for the eigenvalues.

A special case is $f(x) = x^k$ and then the spectral statistics is nothing but the trace of M^k. Often the trace is asymptotic normal. These traces have been traditionally used in dependent models for testing of hypothesis.

(Functional) convergence of spectral statistics have been proved only under restrictive assumptions on the function f.

Extreme eigenvalue

The maximum eigenvalue of real symmetric matrices, especially the sample covariance matrix has been quite useful in statistical tests to test for the so-called “spiked models”.

The limit distribution of the maximum for the S matrix as well as the Wigner matrix are (two of) the Tracy-Widom laws.

Very limited results are known for the maximum (and other order statistics) in other random matrix models.

Recent extensions of the S matrix in high dimensional statistics

The S matrix uses the *product-moment* as a measure of dependence. Other measures of dependence have recently been used leading to other random matrices. These include:

Spearman's rank correlation matrix.

Kendall's τ matrix.

The *Separable Covariance Matrix* and its generalisations also extend the S matrix.

Spearman's rank correlation matrix

Recall the data matrix $X = ((X_{ij}))_{1 \le i \le p, 1 \le j \le n}$. Spearman's rank correlation matrix: $R = ((r_{kl}))_{1 \le k,l \le p}$ where

$$r_{kl} = \frac{\sum_{j=1}^{n} (Q_{kj} - \bar{Q}_k)(Q_{lj} - \bar{Q}_l)}{\left(\sum_{j=1}^{n} (Q_{kj} - \bar{Q}_k)^2 \right)^{1/2} \left(\sum_{j=1}^{n} (Q_{kj} - \bar{Q}_k)^2 \right)^{1/2}}$$

where Q_{kj} is the rank of X_{kj} among $\{X_{ki} : 1 \le i \le n\}$ and $\bar{Q}_k = (n-1)/2$.

This is a non-parametric correlation matrix and is used in testing of independence when the observations are heavy tailed random vectors.

Results for R, $p/n \to y$

- ESD of R converges weakly to MP_y almost surely.

Suppose $\{X_{ij}\}$ are identically distributed.

- $\{\text{Trace}(R^k) - \text{E Trace}(R^k)\}_{k \geq 2}$ converges weakly to a mean zero Gaussian process with covariance function characterized by y.

- Appropriately centered and scaled largest eigenvalue of R converges weakly to the Tracy-Widom law of type 1.

Kendall's τ matrix

$\mathcal{T} = ((\tau_{kl}))_{1 \le k,l \le p}$, where

$$\tau_{k,l} = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \text{Sign}(X_{ki} - X_{kj})\text{Sign}(X_{li} - X_{lj})$$

where $\{X_{ij}\}$ are identically distributed.

- ESD of $\mathcal{T}$ converges weakly to the law of $\frac{2}{3}Y + \frac{1}{3}$ in probability as $p/n \to y > 0$ where Y is an MP_y variable.
- Analytic linear spectral statistic (LSS) converges weakly to a mean zero Gaussian process with a covariance function that depends on y.
- Largest eigenvalue of $\mathcal{T}$ converges weakly to the Tracy-Widom law of type 1.

Separable covariance matrices, $p/n \to y \neq 0$

Data: $X = ((\, X_{ij} \,))_{1 \leq i \leq p, 1 \leq j \leq n}$. Consider the matrix:

$$B_n = \frac{1}{n} T_{2n}^{1/2} X_n T_{1n} X_n' T_{2n}^{1/2},$$

where $T_{1n} \in \mathbb{R}^{n \times n}$ and $T_{2n} \in \mathbb{R}^{p \times p}$ are positive definite.

These are known as *Separable Sample Covariance Matrices*. Pre/post multiplication helps to model dependence between columns/rows, starting usually with independent entries in X.

Under suitable assumptions (for example the LSD of T_{1n} and T_{2n} exist) the LSD of B_n exists. The limit can be expressed in terms of free variables and its Stieltjes transform can be written down in terms of some functional equations.

The LSS of B_n are known to be asymptotically normal.

Matrix polynomial generalisations, $p/n \to y \geq 0$

We can have several independent X matrices and T matrices (of orders $n \times n$ and $p \times p$) and form a polynomial in these matrices, always making sure that each X occurs in pairs. For example $X_1 A X_1' B B' X_2 X_2' X_1 A' X_1'$.

- Under suitable assumptions, the LSD is known to exist for any symmetric matrix polynomial. Further any trace is asymptotically normal.
- Extensions to the case where $p, n = n(p) \to \infty, p/n \to 0$ are available.
- These results can be used to study the LSD of symmetric autocovariance matrix polynomials in high dimensional linear time series models.

LSS and their functions: testing correlations

Write $X = \begin{pmatrix} \mathbb{X}_1 \\ \mathbb{X}_2 \end{pmatrix}$ where $\mathbb{X}_1 \in \mathbb{R}^{p_1 \times n}$ and $\mathbb{X}_2 \in \mathbb{R}^{(p-p_1)\times n}$.

$$A_{x_1x_1} = \frac{1}{n}\mathbb{X}_1\mathbb{X}_1', \quad A_{x_2x_2} = \frac{1}{n}\mathbb{X}_2\mathbb{X}_2', \quad A_{x_1x_2} = A_{x_2x_1}' = \frac{1}{n}\mathbb{X}_1\mathbb{X}_2'. \quad (1)$$

Suppose $p/n \to y > 0$, $p_1/p \to c \in (0,1)$

Consider the linear spectral statistics

$$\mathbb{A} = A_{x_1x_1}^{-1/2} A_{x_1x_2} A_{x_2x_2}^{-1} A_{x_2x_1}' A_{x_1x_1}^{-1/2}$$

This statistic is asymptotically normal and can be used for testing the independence of $\mathbb{X}_1$ and $\mathbb{X}_2$.

LSS and their functions, $p/n \to y \neq 0$

Wilk's Λ *statistic*: $-\log\left(\frac{\text{Det}(T)}{\text{Det}(T+W)}\right)$.

Lawley-Hotelling trace statistic: $\text{Trace}(WT^{-1})$.

Bartlett-Nanda-Pillai trace statistic:
$\text{Trace}(WT^{-1}(1+WT^{-1})^{-1})$.

Here

$$W = A_{x_2x_1}A_{x_1x_1}^{-1}A_{x_1x_2} \text{ and } T = (A_{x_2x_2} - W)^{-1}.$$

They are all asymptotic normal.

Spiked covariance models, $p/n \to y \neq 0$

Data matrix: $X = ((\, X_{ij} \,))_{1 \le i \le p, 1 \le j \le n}$.

Spiked separable covariance matrix.

$$B_n = \frac{1}{n} T_{2n}^{1/2} X_n T_{1n} X_n' T_{2n}^{1/2}.$$

T_{1n} and T_{2n} are positive definite matrices and their first few leading eigenvalues are far away from the bulk spectrum.

- Suppose T_{1n} is the identity matrix. Then the spiked eigenvalues are asymptotically normal. Moreover, they are asymptotically independent of the LSS of the bulk (remainder) of the spectrum.

- Tracy-Wisdom law for the largest non-spiked eigenvalue of B_n is known.
- Application: high-dimensional PCA

Selected journal references-I

Sample Autocovariance Matrix: Basak, Bose and Sen, Bernoulli (2014). [LSD for $p = 1$.]

Sample Autocovariance Matrices, $p, n = n(p) \to \infty$, $p/n \to y$, $0 \leq y < \infty$: Bhattacharjee and Bose, AoS (2019), IJPAM (2017), AoS (2016), RMTA (2016). [LSD of symmetric polynomials and asymptotic normality of traces.]

Sample Autocovariance Matrices, $p, n = n(p) \to \infty$, $p/n \to y$, $0 < y < \infty$ Bose and Hachem, arxiv (2021), JMA (2020). [LSD of Γ_1 in the white noise and the MA(1) case (under Gaussianity).]

Spearman's R: Bao, Preprint (2019); Bao Lin, Pan and Zhou, AoS (2015); Bao and Zhu, Statistica Sinica (2008). [LSD, Trace, Tracy-Widom.]

Kendall's τ: Li, Wang, and Li, AoS (2021); Bao, AoS (2019); Bandeira, Afonso, Lodhia, and Rigollet, ECP (2017). [LSD, Spectral statistics, Tracy-Widom.]

Selected journal references-II

LSS:

Bai, Li, and Pan. Bernoulli (2019). [CLT for LSS of separable sample covariance matrices.];

Yang and Pan. AoS (2015). [Test of independence based on the asymptotic normality of an LSS].

Bodnar, Dette, and Parolya. AoS (2019).[Asymptotic normality of Wilk's Λ, Lawley-Hotelling trace and Bartlett-Nanda-Pillai trace statistics.]

Selected journal references-III

Spiked covariance matrix:

Jiang and Zhidong Bai. Bernoulli (2021). [Asymptotic normality of spiked eigenvalues when T_{1n} is identity.]

Zhang, Zheng, Pan and Zhong. arxiv (2020). [Asymptotic independence of spiked eigenvalues and LSS when T_{1n} is identity.]

Cai, Han, and Pan. AoS (2020). [Asymptotic normality of spiked eigenvalues and Tracy-Wisdom law for the largest non-spiked eigenvalue of B_n.]

Selected books

Anderson, G.; Guionnet, A. and Zeitouni, O. (2010). *An Introduction to Random Matrices*. Cambridge Univ. Press.

Bai, Z.D. and Silverstein, J. (2009). *Spectral Analysis of Large Dimensional Random Matrices*. Springer.

Bose, A. (2018). *Patterned Random Matrices*. Chapman & Hall.

Bose, A. (2021). *Random Matrices and Non-commutative Probability*. Chapman & Hall.

Bose. A. and Bhattacharjee, M. (2018). *Large Covariance and Autocovariance matrices*. Chapman & Hall.

Bose, A. and Saha, K. (2018). *Random Circulant Matrices*. Chapman & Hall.

Couillet, R. and Debbah, M. (2011). *Random Matrix Methods for Wireless Communications*. Cambridge Univ. Press.

Nica, A. and Speicher, R. (2006). *Lectures on the Combinatorics of Free Probability*. London Math. Soc.

Pastur, L. and Shcherbina, M. (2018). *Eigenvalue Distribution of Large Random Matrices*. Amer. Math. Soc.

Tulino, A.M. and Verdu, S. (2004). *Random Matrix Theory and Wireless Communications*. Now Publishers.