Tracy-Widom limit for the largest eigenvalue of high-dimensional covariance matrices in elliptical distributions

Wang Zhou[1] jointly with Jun Wen[1]

[1]Department of Statistics and Applied Probability
National University of Singapore

Outline

Introduction

Eigenvalues of covariance matrices play vital roles in various statistical problems. PCA, Factor model, Spiked covariance model Johnstone (2001), etc. Traditionally, eigenvalues of the sample covariance matrix are used as the estimators of their population counterparts.

Introduction

Let $\mathbf{y}_1, \ldots, \mathbf{y}_N$ be i.i.d. M-variate random vectors with mean 0 and identity population covariance matrix, i.e. $\mathbb{E}\mathbf{y}_1\mathbf{y}_1 = I$. Let Σ be a $M \times M$ symmetric positive definite matrix. In many statistical problems, we observe i.i.d. data $\Sigma^{1/2}\mathbf{y}_1, \ldots, \Sigma^{1/2}\mathbf{y}_N$. The objective is to make inference on the unknown matrix Σ from the sample covariance matrix $\mathcal{W} = N^{-1}\sum_{i=1}^{N}\Sigma^{1/2}\mathbf{y}_i\mathbf{y}_i^*\Sigma^{1/2}$ where $*$ is the transpose of matrices. Let λ_1 be the largest eigenvalue of $\mathcal{W}$. It is known that if each $\mathbf{y}_i$ consists of i.i.d. components with arbitrary finite order moment and $M, N \to \infty$ with $M/N \to \phi > 0$, then the properly normalized λ_1 converges weakly to Tracy-Widom distribution.

Introduction

Question

What if the i.i.d. components assumption on $\mathbf{y}_i$ is violated. For example, the elliptically distributed data $\mathbf{x} = \xi \Sigma^{1/2} \mathbf{u}$ where ξ is a scaler random variable representing the radius of $\mathbf{x}$ and $\mathbf{u}$ is a random vector uniformly distributed on the $M-1$ dimensional unit sphere and independent with ξ.

Result

Suppose we observe i.i.d. data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ where $\mathbf{x}_i = \xi_i \Sigma^{1/2} \mathbf{u}_i$ for $i = 1, \ldots, N$ are elliptically distributed random vectors. Under certain conditions of ξ_i, the normalized largest eigenvalue of the sample covariance matrix $\mathcal{W} = N^{-1} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^*$ still converges weakly to Tracy-Widom distribution.

Notation

Let X_N, Y_N be two sequences of nonnegative random variables. We say that X_N is **stochastically dominated** by Y_N, denoted as $X_N \prec Y_N$ if for any (small) $\varepsilon > 0$ and (large) $D > 0$, there exists $N_0(\varepsilon, D)$ such that $\mathbb{P}(X_N \geq N^{\varepsilon} Y_N) \leq N^{-D}$ for all $N \geq N_0(\varepsilon, D)$. It can be intuitively interpreted as "with high probability, X_N is no greater than Y_N up to a small multiple depending on N".

Green function and Stieltjes transform

For any probability measure ϱ, its Stieltjes transform is defined as

$$m_\varrho(z) = \int \frac{1}{x - z} \varrho(\mathrm{d}x), \qquad \forall z \in \mathbb{C}^+.$$

There is one-to-one correspondence between Stieltjes transform and probability measure. Using the inversion formula, one can recover the probability measure from its Stieltjes transform. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ be the $M \times N$ matrix. Denote $W = N^{-1} X^* X$. Let $\lambda_1, \ldots, \lambda_N$ be the set of eigenvalues of W. One can see that the empirical distribution of $\lambda_1, \ldots, \lambda_N$, $\varrho_N(x) := N^{-1} \sum_{i=1}^N \mathbf{1}\{\lambda_i \leq x\}$, is in fact a (random) probability measure. Our analysis will focus on its Stieltjes transform $m_{\varrho_N}(z)$. We note that the set of eigenvalues of $\mathcal{W}$ and W are the same up to $|N - M|$ zeros.

Green function and Stieltjes transform

For $z \in \mathbb{C}^+$, define $G(z) = (W - zI)^{-1}$ and $\mathcal{G}(z) = (\mathcal{W} - zI)^{-1}$. $G(z)$ and $\mathcal{G}(z)$ are referred to as the Green functions. We note that $m_{\varrho_N}(z) = N^{-1}\mathrm{Tr}G(z)$.

Let $\sigma_1, \ldots, \sigma_M$ be the eigenvalues of Σ. Suppose the empirical distribution π_N of $\sigma_1, \ldots, \sigma_M$ converges weakly to a probability distribution π as as $M, N \to \infty$ with $M/N \to \phi > 0$. A well-known result in random matrix theory is that for any fixed $z \in \mathbb{C}^+$, $m_{\varrho_N}(z)$ converges to a limit $m(z)$ The limit $m(z)$ is the Stieltjes transform of a probability distribution, say ϱ. Correspondingly, ϱ_N converges weakly to ϱ. Moreover, the quantitative relationship between ϱ and π is described by the following equation involving the Stieltjes transform

$$z = -\frac{1}{m(z)} + \phi \int \frac{x\pi(\mathrm{d}x)}{1 + xm(z)} \qquad \forall z \in \mathbb{C}^+. \tag{1}$$

From global law to local law

The above equation establishes convergence of $m_{\varrho_N}(z)$ given any fixed $z \in \mathbb{C}^+$. To investigate the convergence of the largest eigenvalue λ_1, it is necessary to consider the convergence of $m_{\varrho_N}(z)$ together with z approaching from $\mathbb{C}^+$ to the real line at certain rate as $N \to \infty$. To be specific, to carry out the analysis, we assume $z = E + \iota\eta$ such that $N^{-1+\tau} \leq \eta \leq \tau^{-1}$ for some fixed small constant $\tau > 0$ and E is restricted into a small neighborhood of λ_+. Here λ_+ is the right endpoint of ϱ.

Assumptions

Assumptions

1. $N, M \to \infty$ with $M/N \to \phi > 0$ and the spectral norm $||\Sigma||$ of Σ is bounded uniformly for all large M, N.
2. $\mathbf{u}_1, \ldots, \mathbf{u}_N$ are i.i.d. random vectors uniformly distributed on the $M-1$ unit sphere.
3. $\xi_1, \ldots, \xi_N$ are independent real-valued random variables such that $\mathbb{E}\xi_i^2 = MN^{-1}$, $\mathbb{E}|\xi_i|^p < \infty$ for all $p \in \mathbb{Z}_+$ and $\xi_i^2 - \mathbb{E}\xi_i^2 \prec N^{-1/2}$ uniformly for all $i \in \{1, \ldots, M\}$.
4. Subcriticality assumption on the eigenvalues of Σ. (The eigenvalues of Σ satisfy that the corresponding limiting sample eigenvalues are not separated from the bulk)

Assumptions

Assumption 3 excludes some elliptical distributions, such as multivariate student-t distributions and normal scale mixtures. But there are still a wide range of distributions satisfying Assumption 3, including

$$
\begin{aligned}
f(\mathbf{x}) &\sim |\Sigma|^{-1/2}\exp(-\mathbf{x}^\top\Sigma^{-1}\mathbf{x}/2), && \text{normal,}\\
f(\mathbf{x}) &\sim |\Sigma|^{-1/2}(1-\mathbf{x}^\top\Sigma^{-1}\mathbf{x})^{\beta/2-1}, && \text{Pearson type II,}\\
f(\mathbf{x}) &\sim |\Sigma|^{-1/2}(\mathbf{x}^\top\Sigma^{-1}\mathbf{x})^{k-1}\exp(-\beta[\mathbf{x}^\top\Sigma^{-1}\mathbf{x}]^{s}), && \text{Kotz-type.}
\end{aligned}
$$

see the examples and Table 1 of Hu et al. (2019).

Local law

Theorem (Strong local law)

Let δ_{ij} *be the Kronecker delta. Denote* $\Lambda(z) = \max_{i,j} |G_{ij}(z) - \delta_{ij} m(z)|$. *Then*

$$\Lambda(z) \prec \sqrt{\frac{\Im m(z)}{N\eta}} + \frac{1}{N\eta},$$

and

$$|m_{\varrho_N}(z) - m(z)| \prec \frac{1}{N\eta}.$$

This theorem shows that as $N, M \to \infty$ with $\eta \to 0$ at the rate $N^{-1+\tau}$, with high probability, $m_{\varrho_N}(z)$ and $m(z)$ are close to each other in the order $N^{-\tau}$.

Eigenvalue rigidity

Theorem (Eigenvalue rigidity)

$$|\lambda_1 - \lambda_+| \prec N^{-2/3}, \tag{2}$$

and for any real numbers E_1, E_2 *in a neighborhood of* λ_+

$$\Big|\varrho_N([E_1, E_2]) - \varrho([E_1, E_2])\Big| \prec N^{-1}. \tag{3}$$

This theorem shows that the largest sample eigenvalue λ_1 is close to λ_+ in the order of $N^{-2/3}$ which suggests that the scaler of λ_1 to obtain the weak convergence is $N^{2/3}$. This theorem also shows that the probability measure between ϱ_N and ϱ of a small neighborhood of λ_+ is of order N^{-1} with high probability.

Edge universality

Theorem (Edge universality)

Let $\tilde{X} = (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_N)$ such that $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_N$ are i.i.d. M-variate normal random vectors with mean 0 and covariance matrix Σ. Denote $\tilde{\lambda}_1$ be the largest eigenvalue of the sample covariance matrix of $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_N$. Then for any s and $\varepsilon, \delta > 0$, as N large

$$\mathbb{P}(N^{2/3}(\tilde{\lambda}_1 - \lambda_+) \leq s - N^{-\varepsilon}) - N^{-\delta} \leq \mathbb{P}(N^{2/3}(\lambda_1 - \lambda_+) \leq s) \leq \mathbb{P}(N^{2/3}(\tilde{\lambda}_1 - \lambda_+) \leq s + N^{-\varepsilon}) + N^{-\delta}.$$

This theorem shows that the limiting distribution of λ_1 is the same as that of $\tilde{\lambda}_1$ which is already well-known to be the Tracy-Widom distribution. Thus our final result follows.

Local law

First we establish large deviation bounds of spherical uniform distributed random vectors. Specifically, for any $M \times M$ matrix A,

$$|\mathbf{u}^* A \mathbf{u} - M^{-1} \mathrm{Tr} A| \prec M^{-1} ||A||_F. \tag{4}$$

Our starting point is the following resolvent identity (from elementary linear algebra)

$$\frac{1}{G_{ii}(z)} = -z - z\mathbf{x}_i^* \mathcal{G}^{(i)}(z)\mathbf{x}_i, \qquad \forall i = 1, \ldots, N.$$

where $\mathcal{G}^{(i)}(z)$ is the Green function formed from $\mathbf{x}_1, \ldots, \mathbf{x}_N$ excluding $\mathbf{x}_i$. Next, by applying (4) and a bootstrapping procedures (see e.g. Bao et al. (2013); Knowles and Yin (2017)), we obtain a weaker version of the local law which is

$$\Lambda(z) \prec (N\eta)^{-1/4}.$$

Local law

Then, the bound $|m_{\varrho_N}(z) - m(z)| \prec (N\eta)^{-1}$ can be obtained from the weak local law and the fluctuation averaging technique (see e.g. Benaych-Georges and Knowles (2016)). The key mechanism behind fluctuation averaging is that similar to law of large number, the average $N^{-1}\sum_{i=1}^{N}\{(G_{ii}(z))^{-1} - \mathbb{E}_i(G_{ii}(z))^{-1}\}$ reduces the magnitude of each $\{(G_{ii}(z))^{-1} - \mathbb{E}_i(G_{ii}(z))^{-1}\}$. But the order is $(N\eta)^{-1/2}$ instead of $N^{-1/2}$, where $\mathbb{E}_i$ is the conditional expectation given all $\xi_1, \ldots, \xi_N, \mathbf{u}_1, \ldots, \mathbf{u}_N$ except $\xi_i, \mathbf{u}_i$.

Eigenvalue rigidity

(2) is obtained from showing that with high probability, there is no eigenvalue in the interval $[\lambda_+ + N^{-2/3+\varepsilon}, C]$. The proof is based on standard arguments in literature about local law and (3) can be obtained from Helffer-Sjöstrand arguments (see e.g. Benaych-Georges and Knowles (2016)).

Edge universality

It is not difficult to verify that given an interval $[E_1, E_2]$, the number of eigenvalues of sample covariance matrix can be approximated by the quantity

$$\int_{E_1}^{E_2} N \Im m_{\varrho_N}(y + \iota\eta) \mathrm{d}y.$$

Therefore, the problem of evaluating the probability that λ_1 falls in a neighborhood of λ_+ converts to the problem of comparing the Stieltjes transform $m_{\varrho_N}(z)$ with $m(z)$.

Green function comparison

The proof of the edge universality result relies on a Green function comparison theorem which states that for a four times differentiable functions F satisfying some set of regularity conditions, the difference

$$\left| F\left(\int_{E_1}^{E_2} N \Im m_{\varrho_N}(y + \imath\eta) \mathrm{d}y \right) - F\left(\int_{E_1}^{E_2} N \Im \tilde{m}_{\varrho_N}(y + \imath\eta) \mathrm{d}y \right) \right| \quad (5)$$

is bounded by a negative power of N.

Green function comparison

The key idea behind its proof is the Lindberg type argument which replaces a whole column $\mathbf{x}_i$ by $\tilde{\mathbf{x}}_i$ each time. In fact, it is enough to replace the radius variable ξ_i, by $\tilde{\xi}_i$ $i = 1, \dots, N$ where $\tilde{\xi}_i$ are i.i.d. random variables following chi-square distribution with M degrees of freedom and are independent with $\mathbf{u}_1, \dots, \mathbf{u}_N$. It can be verified that $\tilde{\xi}_i \mathbf{u}_i$ follows standard multivariate normal distribution. Since the limiting distribution of the normalized largest eigenvalue given multivariate normally distributed data has been known to be Tracy-Widom (Johnstone (2001)), our problem converts to bound the difference (5) between general elliptically distributed data and multivariate normally distributed data.

Reference I

BAO, Z. G., PAN, G. M., and ZHOU, W. (2013). Local density of the spectrum on the edge for sample covariance matrices with general population.
Preprint.

BENAYCH-GEORGES, F. and KNOWLES, A. (2016).
Lectures on the local semicircle law for Wigner matrices.
ArXiv e-prints.

HU, J., LI, W., LIU, Z., and ZHOU, W. (2019).
High-dimensional covariance matrices in elliptical distributions with application to spherical test.
Ann. Statist., 47(1):527–555.

Reference II

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

KNOWLES, A. and YIN, J. (2017).
Anisotropic local laws for random matrices.
Probab. Theory Related Fields, 169(1):257–352.

Thank You

Thank you!