Random Matrix Theory for Modern Machine Learning: New Intuitions, Improved Methods, and Beyond: Part 2

Short Course @ Institut de Mathématiques de Toulouse, France

Zhenyu Liao

School of Electronic Information and Communications
Huazhong University of Science and Technology

July, 2nd, 2024

Outline

1. Four Ways to Characterize Sample Covariance Matrices
   - Traditional analysis of SCM eigenvalues
   - SCM analysis beyond eigenvalues: a modern RMT approach via Deterministic Equivalents for resolvent
   - The Gaussian method alternative approach
2. Some More Random Matrix Models
   - Wigner semicircle law
   - Generalized sample covariance matrix
   - Separable covariance model

Four ways to characterize sample covariance matrices

Definition (Sample Covariance Matrix, SCM)

The SCM $\hat{\mathbf{C}} \in \mathbb{R}^{p \times p}$ of data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ composed of n independent data samples $\mathbf{x}_i \in \mathbb{R}^p$ of zero mean is given by

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} = \frac{1}{n} \mathbf{X}\mathbf{X}^\mathsf{T}. \tag{1}$$

Definition (Classical versus proportional regimes)

For SCM $\hat{\mathbf{C}} \in \mathbb{R}^{p \times p}$ from n samples of dimension p, consider the following two regimes.

1. **Classical regime** with $n \gg p$, this includes both asymptotic ($n \to \infty$ with p fixed) and non-asymptotic characterizations ($n \gg p$ for large but finite n).
2. **Proportional regime** with $n \sim p$, this includes both asymptotic ($n, p \to \infty$ with $p/n \to c \in (0, \infty)$, also known as thermodynamic limit in the statistical physics literature) and non-asymptotic characterizations ($n \sim p \gg 1$ both large but finite).

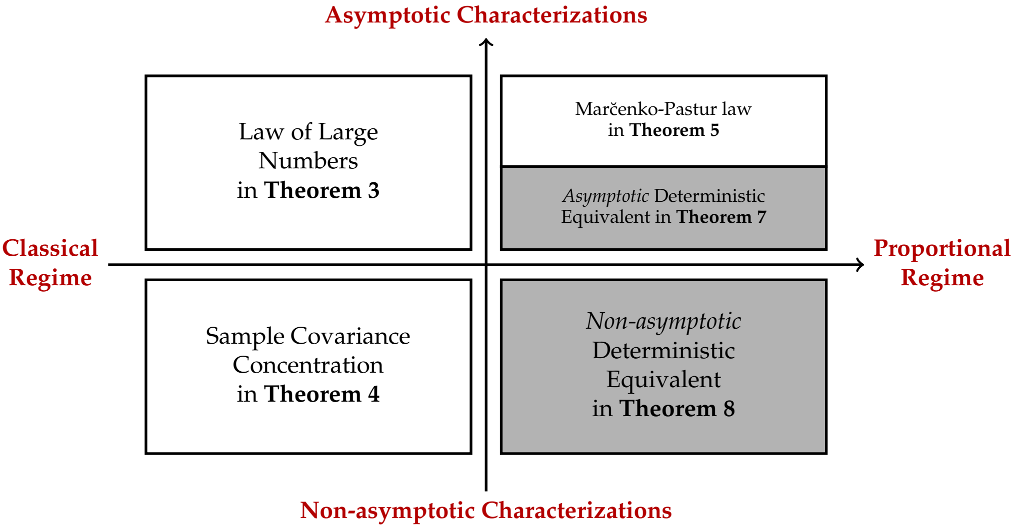


Figure: Taxonomy of four different ways to characterize the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$.

Asymptotically deterministic behavior: from random vectors to random matrices

- **Key object**: characterizations of large random matrices in the **proportional** regime
- e.g., eigenspectral behaviors of $\hat{\mathbf{C}}$ very different in the classical from the proportional regime, **not sure** whether they establish a close-to-deterministic behavior in the proportional $n \sim p \gg 1$ regime
- we have seen concentration of (linear, Lipschitz, quadratic, and nonlinear quadratic) scalar observations of large-dimensional random vectors

$$f(\mathbf{x}) \simeq \mathbb{E}[f(\mathbf{x})] + O(n^{-1/2}). \tag{2}$$

- we expect something similar for random matrices:

(i) similar to vectors, the random matrices themselves do not concentrate (in a spectral norm sense) **in the proportional** $n \sim p \gg 1$ regime, e.g., $\|\hat{\mathbf{C}} - \mathbf{C}\|_2 \to 0$ as $n, p \to \infty$ limit with $p/n \to c \in (0, \infty)$[1]

(ii) large-dimensional close-to-deterministic/concentration behavior for its scalar (e.g., eigenspectral) observations $F(\hat{\mathbf{C}})$ holds for scalar matrix functional $F \colon \mathbb{R}^{p \times p} \to \mathbb{R}$, in the **proportional** $n \sim p \gg 1$ regime.

[1] This is sharp contrast to the **classical** $n \gg p \sim 1$ regime, where $\|\hat{\mathbf{C}} - \mathbf{C}\| \simeq 0$ for any matrix norm.

Asymptotic behavior of SCM in the classical regime via law of large numbers

Theorem (Asymptotic Law of Large Numbers for SCM)

Let p be fixed, and let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random matrix with independent sub-gaussian columns $\mathbf{x}_i \in \mathbb{R}^p$ such that $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. Then one has,

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \to 0, \tag{3}$$

almost surely, as $n \to \infty$.

- LLN is "parameterized" to hold only in the **classical limit**, not the **proportional limit**
- many variants and extensions of the LLN exist, but become vacuous when applied to the **proportional regime** $n, p \to \infty$ and $p/n \to c \in (0, \infty)$, see below for an example

Non-asymptotic behavior of SCM in the classical regime via matrix concentration

Theorem (Non-asymptotic matrix concentration for SCM, [Ver18, Theorem 4.6.1])

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random matrix with independent sub-gaussian columns $\mathbf{x}_i \in \mathbb{R}^p$ such that $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. Then, one has, with probability at least $1 - 2\exp(-t^2)$, for any $t \geq 0$, that

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \leq C_1 \max(\delta, \delta^2), \quad \delta = C_2(\sqrt{p/n} + t/\sqrt{n}), \tag{4}$$

for some constants $C_1, C_2 > 0$, independent of n, p.

Proof: combines Bernstein's concentration inequality with ϵ-net argument, see [Ver18] for details.

- can reproduce the LLN asymptotic result by taking $n \to \infty$ with Borel–Cantelli lemma

(i) **Classical regime.** Here, $n \gg p$, **say that** $n \sim p^2$. Then with high probability, that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(n^{-1/4})$ and conveys a similar intuition to the asymptotic LLN result

(ii) **Proportional regime.** Here, n, p **are both large and** $n \sim p$. Then, with high probability, that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(\sqrt{p/n}) = O(1)$, and qualitatively different LLN with a vacuous $\sim 100\%$ relative error, e.g., as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$.

Proportional regime: eigenvalues via traditional RMT and the Marčenko-Pastur law

Theorem (Limiting spectral distribution for SCM: Marčenko-Pastur law, [MP67])

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ *be a random matrix with i.i.d. sub-gaussian columns* $\mathbf{x}_i \in \mathbb{R}^p$ *such that* $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ *and* $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. *Then, as* $n, p \to \infty$ *with* $p/n \to c \in (0, \infty)$, *with probability one, the empirical spectral measure (ESD)* $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ *of* $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ *converges weakly to a probability measure* μ *given explicitly by*

$$\boxed{\mu(dx) = (1 - c^{-1})^+ \delta_0(x) + \frac{1}{2\pi c x} \sqrt{(x - E_-)^+ (E_+ - x)^+} \, dx,} \tag{5}$$

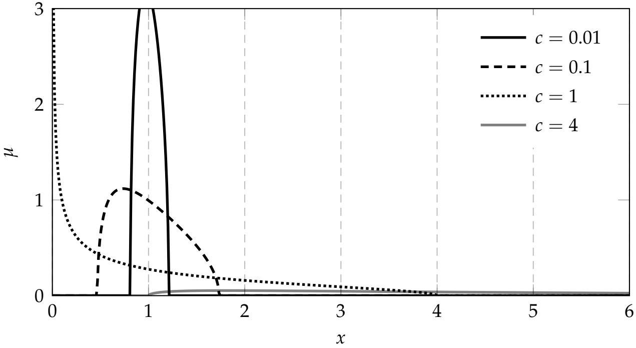
where $E_\pm = (1 \pm \sqrt{c})^2$ *and* $(x)^+ = \max(0, x)$, *which is known as the Marčenko-Pastur distribution.*

- provides a more refined characterization of the eigenspectrum of $\hat{\mathbf{C}}$ (than, e.g., matrix concentration):

(i) **Classical regime.** Here, $n \gg p$ **so that** $c = p/n \to 0$, the Marčenko-Pastur law in Equation (5) shrinks to a Dirac mass, in agreement with $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \sim 0$

(ii) **Proportional regime.** Here, $n \sim p \gg 1$, and by the (true but vacuous) matrix concentration result $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(p/n) = O(1)$, and, depending on the ratio $c = p/n$, the eigenvalues of $\hat{\mathbf{C}}$ can be very different from one, and takes the form of the Marčenko-Pastur law

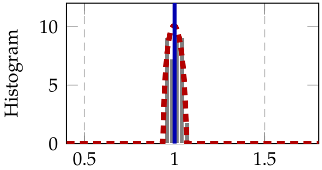
- we have in fact $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \simeq c + 2\sqrt{c}$ as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$



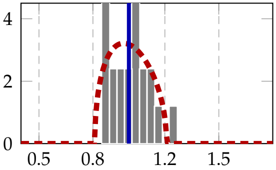
- averaged amount of eigenvalues of $\hat{\mathbf{C}}$ lying within the interval $[1-\delta, 1+\delta]$, for $\delta \ll 1$, as

$$\mu([1-\delta, 1+\delta]) = \int_{1-\delta}^{1+\delta} \frac{1}{2\pi c x} \sqrt{\left(x - (1-\sqrt{c})^2\right)^+ \left((1+\sqrt{c})^2 - x\right)^+} \, dx$$
$$= \frac{1}{2\pi c} \int_{-\delta}^{\delta} \left(\sqrt{4c - c^2} + O(\varepsilon) \right) d\varepsilon = \frac{\sqrt{4c^{-1} - 1}}{\pi} \delta + O(\delta^2).$$

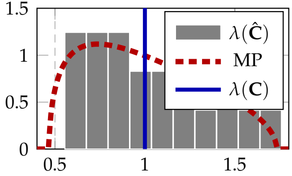
- for $p \approx 4n$ there is asymptotically no eigenvalue of $\hat{\mathbf{C}}$ close to one!
- in accordance with the shape of the limiting Marčenko-Pastur law with $c = 4$ above



(a) $n = 1\,000p$

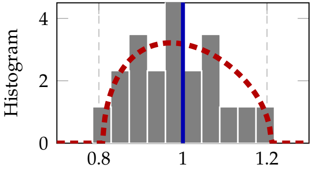


(b) $n = 100p$

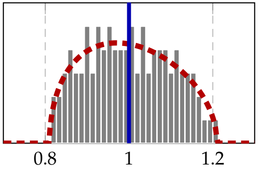


(c) $n = 10p$

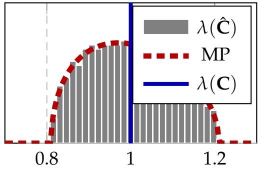
Figure: **Varying n and $c = p/n$ for fixed p.** Histogram of the eigenvalues of $\hat{\mathbf{C}}$ versus the limiting Marčenko-Pastur law in Theorem 5, for $\mathbf{X}$ having standard Gaussian entries with $p = 20$ and different $n = 1\,000p, 100p, 10p$ from left to right.



(a) $p = 20$



(b) $p = 100$



(c) $p = 500$

Figure: **Varying n and p for fixed $c = p/n$.** Histogram of the eigenvalues of $\hat{\mathbf{C}}$ versus the Marčenko-Pastur law, for $\mathbf{X}$ having standard Gaussian entries with $n = 100p$ and different $p = 20, 100, 500$ from left to right.

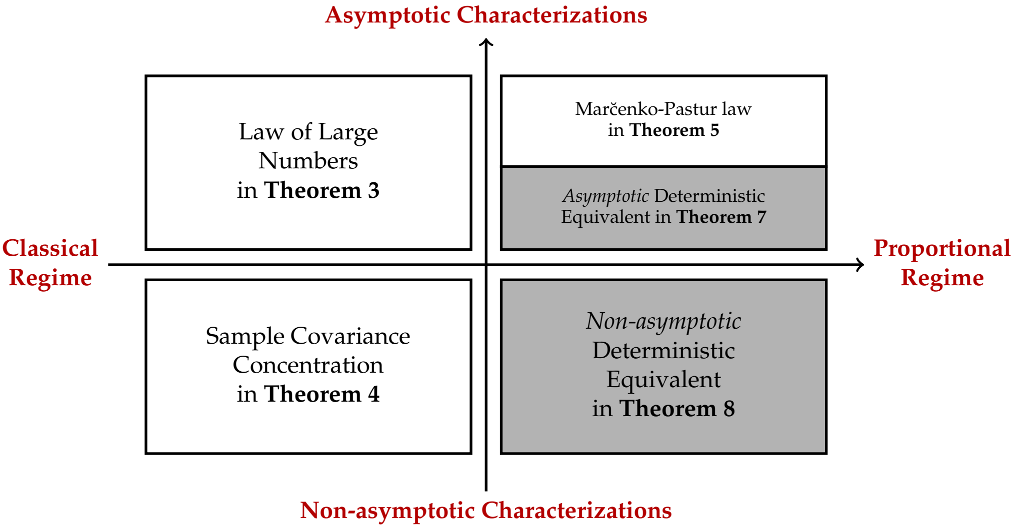


Figure: Taxonomy of four different ways to characterize the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$.

A modern RMT approach via deterministic equivalents for resolvent

- we have seen the **resolvent-based** approach as a **unified** analysis approach to **matrix spectral functionals**
- e.g., interested in the spectral behavior of a random matrix $\mathbf{X} \in \mathbb{R}^{p \times p}$ from n samples, in the proportional $n \sim p \gg 1$ regime, more convenient to work with its resolvent $\mathbf{Q}_{\mathbf{X}}(z) = (\mathbf{X} - z\mathbf{I}_n)^{-1}$
- in particular, scalar observations $F \colon \mathbb{R}^{p \times p} \to \mathbb{R}$ of $\mathbf{X}$ and $\mathbf{Q}_{\mathbf{X}}(z)$ converge/concentrate, and there exists deterministic $\bar{\mathbf{Q}}(z)$ such that

$$F(\mathbf{Q}(z)) - F(\bar{\mathbf{Q}}(z)) \to 0, \tag{6}$$

  as $n, p \to \infty$.
- such $\bar{\mathbf{Q}}(z)$ is a **Deterministic Equivalent** of the random (resolvent) matrix $\mathbf{Q}$.
- so, our general recipe:

$$\text{eigenspectral functional of large random matrix } \mathbf{X}$$
$$\downarrow$$
$$\textbf{more convenient} \text{ to work with } \mathbf{Q}_{\mathbf{X}}(z)$$
$$\downarrow$$
$$\text{find its } \textbf{Deterministic Equivalent}$$

What is actually happening for **Deterministic Equivalent**?

- while the random matrix $\mathbf{Q} \in \mathbb{R}^{p \times p}$ remains random as the dimension p grows, in fact even "**more**" random due to the growing degrees of freedom;
- scalar observation $F(\mathbf{Q})$ of $\mathbf{Q}$ becomes "more concentrated" as $p \to \infty$;
- the random $F(\mathbf{Q})$, if concentrates, must concentrated around its expectation $\mathbb{E}[F(\mathbf{Q})]$;
- as $p \to \infty$, more randomness in $\mathbf{Q} \Rightarrow \mathrm{Var}[F(\mathbf{Q})] \to 0$ sufficiently fast (in p)
- if the functional $F \colon \mathbb{R}^{p \times p} \to \mathbb{R}$ is linear, then $\mathbb{E}[F(\mathbf{Q})] = F(\mathbb{E}[\mathbf{Q}])$.
- So, to propose a DE, suffices to evaluate $\mathbb{E}[\mathbf{Q}]$:
- **however**, $\mathbb{E}[\mathbf{Q}]$ may be hardly accessible, due to integration and nonlinear matrix inverse $\mathbf{Q}(z) = (\mathbf{X} - z\mathbf{I}_p)^{-1}$
- find a **simple** and **more accessible** deterministic $\bar{\mathbf{Q}}$ with $\bar{\mathbf{X}} \simeq \mathbb{E}[\mathbf{Q}]$ in some sense for p large, e.g., $\|\bar{\mathbf{Q}} - \mathbb{E}[\mathbf{Q}]\|_2 \to 0$ as $p \to \infty$; and
- show variance or higher-order moments of $F(\mathbf{Q})$ decay sufficiently fast as $p \to \infty$.

Deterministic Equivalent: definition

Definition (Deterministic Equivalent)

We say that $\bar{\mathbf{Q}} \in \mathbb{R}^{p \times p}$ is an $(\varepsilon_1, \varepsilon_2, \delta)$-*Deterministic Equivalent* for the symmetric random matrix $\mathbf{Q} \in \mathbb{R}^{p \times p}$ if, for a deterministic matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ of unit norms (spectral and Euclidean, respectively), we have, with probability at least $1 - \delta(p)$ that

$$\left| \frac{1}{p} \operatorname{tr} \mathbf{A}(\mathbf{Q} - \bar{\mathbf{Q}}) \right| \leq \varepsilon_1(p), \quad \left| \mathbf{a}^\mathsf{T} (\mathbf{Q} - \bar{\mathbf{Q}}) \mathbf{b} \right| \leq \varepsilon_2(p), \tag{7}$$

for some non-negative functions $\varepsilon_1(p), \varepsilon_2(p)$ and $\delta(p)$ that decrease to zero as $p \to \infty$. Denote

$$\mathbf{Q} \overset{\varepsilon_1, \varepsilon_2, \delta}{\longleftrightarrow} \bar{\mathbf{Q}}, \text{ or simply } \mathbf{Q} \leftrightarrow \bar{\mathbf{Q}}. \tag{8}$$

An asymptotic Deterministic Equivalent for resolvent

Theorem (An asymptotic Deterministic Equivalent for resolvent, [CL22, Theorem 2.4])

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random matrix having i.i.d. sub-gaussian entries of zero mean and unit variance, and denote $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ the resolvent of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $z \in \mathbb{C}$ not an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$. Then, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, the deterministic matrix $\bar{\mathbf{Q}}(z)$ is a Deterministic Equivalent of the random resolvent matrix $\mathbf{Q}(z)$ with

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p, \tag{9}$$

with $m(z)$ the unique valid Stieltjes transform as solution to

$$czm^2(z) - (1 - c - z)m(z) + 1 = 0. \tag{10}$$

- The equation of $m(z)$ is quadratic and has two solutions defined via the complex square root
- **only one** satisfies the relation $\Im[z] \cdot \Im[m(z)] > 0$ as a “valid” Stieltjes transform
- this leads to the Marčenko-Pastur law

$$\boxed{\mu(dx) = (1 - c^{-1})^+ \delta_0(x) + \frac{1}{2\pi c x}\sqrt{(x - E_-)^+ (E_+ - x)^+}\, dx,} \tag{11}$$

for $E_\pm = (1 \pm \sqrt{c})^2$ and $(x)^+ = \max(0, x)$.

[2] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022

A non-asymptotic Deterministic Equivalent for resolvent

Theorem (A non-asymptotic Deterministic Equivalent for resolvent)

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random matrix having i.i.d. sub-gaussian entries with zero mean and unit variance, and denote $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ the resolvent of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $z < 0$. Then, there exists universal constants $C_1, C_2 > 0$ depending only on the sub-gaussian norm of the entries of $\mathbf{X}$ and $|z|$, such that for any $\varepsilon \in (0,1)$, if $n \geq (C_1 + \varepsilon)p$, one has

$$\|\mathbb{E}[\mathbf{Q}(z)] - \bar{\mathbf{Q}}(z)\|_2 \leq \frac{C_2}{\varepsilon} \cdot n^{-\frac{1}{2}}, \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p, \tag{12}$$

for $m(z)$ the unique positive solution to the Marčenko-Pastur equation $czm^2(z) - (1 - c - z)m(z) + 1 = 0, c = p/n$.

- this is a deterministic characterization of the **expected resolvent**
- to get DE, it remains to show **concentration** results for trace and bilinear forms: more or less standard

Proof via leave-one-out and self-consistent equation

Let $\mathbf{x}_i \in \mathbb{R}^p$ denote the ith column of $\mathbf{X} \in \mathbb{R}^{p\times n}$ (so that $\mathbf{x}_i$ has i.i.d. sub-gaussian entries of zero mean and unit variance), and let $\mathbf{X}_{-i} \in \mathbb{R}^{p\times(n-1)}$ denote the random matrix $\mathbf{X}$ *without* its ith column $\mathbf{x}_i$. Define similarly $\mathbf{Q}_{-i}(z) = \left(\frac{1}{n}\mathbf{X}_{-i}\mathbf{X}_{-i}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}$ so that

$$\mathbf{Q}(z) = \left(\frac{1}{n}\mathbf{X}_{-i}\mathbf{X}_{-i}^\mathsf{T} + \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T} - z\mathbf{I}_p\right)^{-1} = \left(\mathbf{Q}_{-i}^{-1}(z) + \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right)^{-1}. \tag{13}$$

First note that by definition,

$$\bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p = \left(\frac{1}{1+cm(z)} - z\right)^{-1}\mathbf{I}_p, \tag{14}$$

for $c = p/n$, so that for $z < 0$,

$$\frac{1}{1+cm(z)}\|\bar{\mathbf{Q}}\|_2 \leq 1. \tag{15}$$

Similarly, one has

$$\|\mathbf{Q}(z)\|_2 \leq \frac{1}{|z|}, \quad \left\|\mathbf{Q}(z)\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right\|_2 \leq 1, \quad \left\|\mathbf{Q}(z)\frac{1}{\sqrt{n}}\mathbf{X}\right\|_2 = \sqrt{\left\|\mathbf{Q}(z)\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}(z)\right\|_2} \leq \frac{1}{\sqrt{|z|}}. \tag{16}$$

A few useful lemmas

Lemma (Resolvent identity)

For invertible matrices $\mathbf{A}$ *and* $\mathbf{B}$*, we have* $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$.

Lemma (Woodbury)

For $\mathbf{A} \in \mathbb{R}^{p \times p}$, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{p \times n}$*, such that both* $\mathbf{A}$ *and* $\mathbf{A} + \mathbf{U}\mathbf{V}^\top$ *are invertible, we have*

$$(\mathbf{A} + \mathbf{U}\mathbf{V}^\top)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}_n + \mathbf{V}^\top \mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^\top \mathbf{A}^{-1}.$$

In particular, for $n = 1$*, i.e.,* $\mathbf{U}\mathbf{V}^\top = \mathbf{u}\mathbf{v}^\top$ *for* $\mathbf{U} = \mathbf{u} \in \mathbb{R}^p$ *and* $\mathbf{V} = \mathbf{v} \in \mathbb{R}^p$*, the above identity specializes to the following Sherman–Morrison formula,*

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{A}^{-1}}{1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}, \quad \text{and } (\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}.$$

And the matrix $\mathbf{A} + \mathbf{u}\mathbf{v}^\top \in \mathbb{R}^{p \times p}$ *is invertible if and only if* $1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u} \neq 0$.

Letting $\mathbf{A} = \mathbf{M} - z\mathbf{I}_p$, $z \in \mathbb{C}$, and $\mathbf{v} = \tau\mathbf{u}$ for $\tau \in \mathbb{R}$ in Woodbury identity leads to the following rank-one perturbation lemma for the resolvent of $\mathbf{M}$.

Lemma ([SB95, Lemma 2.6])

For $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{p \times p}$ symmetric and nonnegative definite, $\mathbf{u} \in \mathbb{R}^p$, $\tau > 0$ and $z < 0$,

$$\left| \operatorname{tr} \mathbf{A}(\mathbf{M} + \tau \mathbf{u}\mathbf{u}^\mathsf{T} - z\mathbf{I}_p)^{-1} - \operatorname{tr} \mathbf{A}(\mathbf{M} - z\mathbf{I}_p)^{-1} \right| \leq \frac{\|\mathbf{A}\|_2}{|z|}.$$

Proof

It follows from the resolvent identity that

$$\begin{aligned}
\mathbb{E}[\mathbf{Q} - \bar{\mathbf{Q}}] &= \mathbb{E}\left[\mathbf{Q}\left(\frac{\mathbf{I}_p}{1+cm(z)} - \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right)\right]\bar{\mathbf{Q}} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}} - \frac{1}{n}\mathbb{E}[\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}]\bar{\mathbf{Q}} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}} - \sum_{i=1}^n \frac{1}{n}\mathbb{E}[\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}]\bar{\mathbf{Q}} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}} - \sum_{i=1}^n \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}\right]\bar{\mathbf{Q}}, \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}} - \sum_{i=1}^n \frac{\mathbb{E}\left[\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\bar{\mathbf{Q}}}{1+cm(z)} + \sum_{i=1}^n \frac{\mathbb{E}\left[\mathbf{Q}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}d_i\right]\bar{\mathbf{Q}}}{1+cm(z)} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}} - \sum_{i=1}^n \frac{\mathbb{E}\left[\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\bar{\mathbf{Q}}}{1+cm(z)} + \frac{\mathbb{E}\left[d_i\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\bar{\mathbf{Q}}}{1+cm(z)}
\end{aligned}$$

with $\boxed{d_i = \mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i/n - cm(z)}$, so that $\mathbb{E}[\mathbf{Q} - \bar{\mathbf{Q}}] = (\mathbb{E}[\mathbf{Q} - \mathbf{Q}_{-i}])\frac{\bar{\mathbf{Q}}}{1+cm(z)} + \frac{\mathbb{E}[d_i\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}]\bar{\mathbf{Q}}}{1+cm(z)}$.

Let

$$T_1 = \|\mathbb{E}[\mathbf{Q} - \mathbf{Q}_{-i}]\|_2, \quad T_2 = \left\|\mathbb{E}\left[d_i \mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\right\|_2, \tag{17}$$

we then have $\|\mathbb{E}[\mathbf{Q} - \bar{\mathbf{Q}}]\| \leq T_1 + T_2$.

For the first term T_1, it follows from Sherman–Morrison that

$$0 \preceq \mathbb{E}[\mathbf{Q}_{-i} - \mathbf{Q}] = \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}}{1 + \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}\right] \preceq \frac{1}{n}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}] = \frac{1}{n}\mathbb{E}\left[\mathbf{Q}_{-i}^2\right] \tag{18}$$

so

$$T_1 = \|\mathbb{E}[\mathbf{Q} - \mathbf{Q}_{-i}]\|_2 = O(n^{-1}). \tag{19}$$

For T_2,

$$\begin{aligned}
T_2 &= \left\|\mathbb{E}\left[d_i \mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\right\|_2 \\
&= \sup_{\|\mathbf{u}\|=1, \|\mathbf{v}\|=1} \mathbb{E}\left[d_i \mathbf{u}^\mathsf{T}\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{v}\right] \\
&\leq \sqrt{\mathbb{E}[d_i^2]} \cdot \sup_{\|\mathbf{u}\|=1, \|\mathbf{v}\|=1} \sqrt{\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{v})^2]} \\
&\leq \underbrace{\sqrt{\mathbb{E}[d_i^2]}}_{T_{2,1}} \cdot \underbrace{\sup_{\|\mathbf{u}\|=1} \sqrt[4]{\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}\mathbf{x}_i)^4]}}_{T_{2,2}} \cdot \underbrace{\sup_{\|\mathbf{v}\|=1} \sqrt[4]{\mathbb{E}[(\mathbf{x}_i^\mathsf{T}\mathbf{v})^4]}}_{T_{2,3}}.
\end{aligned}$$

For the term $T_{2,2}$. Note that

$$\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}\mathbf{x}_i)^4] = \mathbb{E}\left[\frac{(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4}{(1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4}\right] \leq \mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4] = \mathbb{E}[(\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{u}\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^2],$$

with

$$\|\mathbf{Q}_{-i}\mathbf{u}\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\|_2 = \mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}^2\mathbf{u} \leq |z|^{-2}, \tag{20}$$

for $\|\mathbf{u}\| = 1$.

By Hanson–Wright inequality (concentration of quadratic form), there exists $C, C' > 0$ such that

$$\begin{aligned}\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4] = \mathbb{E}\left[\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4|\mathbf{Q}_{-i}]\right] &\leq \mathbb{E}_{\mathbf{Q}_{-i}}\left[\int_0^\infty 2t\cdot\mathbb{P}\left(\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{u}\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i \geq t\right)\,dt\right] \\ &\leq 2C'\cdot\mathbb{E}_{\mathbf{Q}_{-i}}\left[\int_0^\infty t\exp\left(-Ct/(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}^2\mathbf{u})\right)\,dt\right] \\ &= 2C'\mathbb{E}\left[\frac{(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}^2\mathbf{u})^2}{C^2}\right] \leq (Cz^2)^{-2}.\end{aligned}$$

This allows us to conclude that $T_{2,2} = O(1)$, and analogously that $T_{2,3} = O(1)$.
We thus have

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq T_1 + T_2 \leq T_1 + T_{2,1}\cdot T_{2,2}\cdot T_{2,3} \leq C_1 n^{-1} + C_2\sqrt{\mathbb{E}[d_i^2]}, \tag{21}$$

for some universal constants C_1, C_2 and recall $\boxed{d_i \equiv \mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i/n - cm(z)}$.

Now, note that

$$
\begin{aligned}
d_i^2 &= \left(\frac{1}{n} \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{-i} \mathbf{x}_i - cm(z) \right)^2 \\
&= \left(\frac{1}{n} \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{-i} \mathbf{x}_i - \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] + \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] - cm(z) \right)^2 \\
&\leq 2 \left(\frac{1}{n} \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{-i} \mathbf{x}_i - \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] \right)^2 + 2 \left(\frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] - cm(z) \right)^2 \\
&= 2 \left(\frac{1}{n} \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{-i} \mathbf{x}_i - \frac{1}{n} \operatorname{tr} \mathbf{Q}_{-i} + \frac{1}{n} \operatorname{tr} \mathbf{Q}_{-i} - \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] \right)^2 + 2 \left(\frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] - cm(z) \right)^2,
\end{aligned}
$$

so that

$$
\frac{1}{2} \mathbb{E}[d_i^2] \leq \underbrace{\mathbb{E} \left(\frac{1}{n} \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{-i} \mathbf{x}_i - \frac{1}{n} \operatorname{tr} \mathbf{Q}_{-i} \right)^2}_{D_1} + \underbrace{\mathbb{E} \left(\frac{1}{n} \operatorname{tr} \mathbf{Q}_{-i} - \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] \right)^2}_{D_2} + \left(\frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] - cm(z) \right)^2.
$$

- $D_1 \leq Cn^{-2}$ by the same line of arguments as the term $T_{2,2}$
- D_2 that characterizes the **concentration property** of the resolvent trace $\operatorname{tr} \mathbf{Q}_{-i}$, using a martingale difference argument via **Burkholder inequality**.

Lemma

Under the notations and settings above, we have

$$\mathbb{E}\left[\left(\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\mathbb{E}\mathbf{Q})\right)^2\right] \leq Cn^{-1} \text{ and } \mathbb{E}\left[\left(\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\mathbb{E}\mathbf{Q})\right)^4\right] \leq Cn^{-2}, \tag{22}$$

for any $\mathbf{A} \in \mathbb{R}^{p\times p}$ of unit norm and some constant $C > 0$, and thus in particular for $\mathbf{A} = \mathbf{I}_p$.

Thus,

$$\mathbb{E}[d_i^2] \leq 2(D_1 + D_2) + 2\left(\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}] - cm(z)\right)^2 \leq Cn^{-1} + 2\left(\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}] - cm(z)\right)^2, \tag{23}$$

for some universal constant $C > 0$. Putting together and by the trace rank-one update result,

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq C_1 n^{-\frac{1}{2}} + C_2\left|\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}] - cm(z)\right|. \tag{24}$$

Finishing the proof

We "close the loop" by noting that by definition $\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}} = \frac{p}{n}m(z) = cm(z)$, so that

$$\left|\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}] - cm(z)\right| \leq \frac{p}{n}\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq \frac{p}{n}\left(C_1 n^{-\frac{1}{2}} + C_2\left|\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}] - cm(z)\right|\right), \tag{25}$$

and therefore for any $\epsilon > 0$ and $n > (C_2 + \varepsilon)p$, one has

$$\left|\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}] - cm(z)\right| \leq \frac{C_1}{\varepsilon}\cdot n^{-\frac{1}{2}}, \tag{26}$$

and thus

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq \frac{C}{\varepsilon}\cdot n^{-\frac{1}{2}}, \tag{27}$$

for some universal constant $C > 0$. This concludes the proof.

Remark: extension to $z = 0$

- assume above $z < 0$ so that the bound on the random resolvent $\|\mathbf{Q}_{\hat{\mathbf{C}}}(z)\|_2 \leq 1/|z|$
- this, however, does not exploit the information in the random sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} \in \mathbb{R}^{p \times n}$ on, e.g., how it concentrates around its population counterpart $\mathbf{C} = \mathbb{E}[\hat{\mathbf{C}}]$
- to extend the result above to, say, an inverse SCM of the type $\mathbf{Q}(z=0) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}$ with $z = 0$, first needs to ensure the inverse is **well-defined** for sub-gaussian $\mathbf{X}$ and for a specific choice of p, n
- can be obtained, e.g., per concentration of SCM $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ around its expectation.
- it follows from standard SCM concentration (Theorem 4) that there exists universal constant $C > 0$ such that for $n \geq C(p + \ln(1/\delta))$, one has, with probability at least $1 - \delta$, $\delta \in (0, 1/2]$ that

$$\left\| \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - \mathbf{I}_p \right\|_2 \leq \frac{\mathbf{I}_p}{2}, \tag{28}$$

  and therefore $\|\mathbf{Q}(z)\|_2 \leq \frac{1}{1/2 - z} \leq 2$ for any $z \leq 0$
- allows for a control of the spectral norm $\|\mathbf{Q}(z)\|_2 \leq 2$ independent of $z \leq 0$ and holds with probability at least $1 - \delta$
- do everything else **conditioned on this high-probability event**, to get a bound on the **conditional expectation** $\mathbb{E}[\mathbf{Q} \,|\mathcal{E}]$, with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$

Remark: as extensions to results in the classical regime

(i) In the "easy" **classical regime**, with $n \gg p$ (and thus $p/n \to c = 0$), one has that $\hat{\mathbf{C}} \equiv \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} \to \mathbb{E}[\hat{\mathbf{C}}] = \mathbf{I}_p$ as $n \to \infty$, so that

$$(\hat{\mathbf{C}} - z\mathbf{I}_p)^{-1} \simeq (\mathbb{E}[\hat{\mathbf{C}}] - z\mathbf{I}_p)^{-1} = (1-z)^{-1}\mathbf{I}_p = \bar{\mathbf{Q}}(z). \tag{29}$$

(ii) In the "harder" and more general **proportional regime**, for $n \sim p$ with $p/n \to c \in (0, \infty)$, one has instead

$$\bar{\mathbf{Q}}(z) \simeq \mathbb{E}[\mathbf{Q}(z)] \equiv \mathbb{E}[(\hat{\mathbf{C}} - z\mathbf{I}_p)^{-1}] \not\simeq (\mathbb{E}[\hat{\mathbf{C}}] - z\mathbf{I}_p)^{-1}. \tag{30}$$

In this case, a Deterministic Equivalent $\bar{\mathbf{Q}}(z)$ can be very different from $(\mathbb{E}[\hat{\mathbf{C}}] - z\mathbf{I}_p)^{-1}$.

- this is not surprising, consider the scalar case where $\mathbb{E}[1/x] \neq 1/\mathbb{E}[x]$ in general, unless $x \simeq C$ for some constant C

Remark: Deterministic Equivalents for Gaussian inverse SCM

- consider the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z}$ and positive definite $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $\mathbf{Z} \in \mathbb{R}^{p\times n}$ having i.i.d. standard Gaussian entries
- the inverse $\hat{\mathbf{C}}^{-1}$ is known to follow the inverse-Wishart distribution [MKB79] with p degrees of freedom and scale matrix $\mathbf{C}^{-1}$, such that
$$\mathbb{E}[\hat{\mathbf{C}}^{-1}] = \frac{n}{n-p-1}\mathbf{C}^{-1} \tag{31}$$
for $n \geq p+2$.
- On the other hand, it follows from our non-asymptotic result above by taking $z = 0$ that
$$\mathbb{E}[\mathbf{Q}(z)] \leftrightarrow \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p = \frac{n}{n-p}\mathbf{I}_p \tag{32}$$
with $m(z) = \frac{1}{1-c} = \frac{n}{n-p}$.
- **note**: Deterministic Equivalents **are not unique**: could replace the "-1" in denominator by any constant $C' \ll n, p$ to propose another equally correct Deterministic Equivalent.

[3] Kanti Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. 1st ed. Probability and Mathematical Statistics. Academic Press, Dec. 1979

Some thoughts on the "leave-one-out" proof

- **in essence**: propose $\bar{\mathbf{Q}}(z) \simeq \mathbb{E}[\mathbf{Q}(z)]$ (in spectral norm sense), but simple to evaluate (via a quadratic equation)
- leave-one-out analysis of large-scale system: $\mathbf{Q}(z) \simeq \mathbf{Q}_{-i}(z)$ for n, p large.
- low complexity analysis of large random system: joint behavior of p eigenvalues $\overset{\text{RMT}}{\to}$ a single deterministic (quadratic) equation
- **Side Remark**: another (as well) systematic and convenient RMT proof approach: **Gaussian method**, as the combination of

(1) Stein's lemma (Gaussian integration by parts)

(2) Nash–Poincaré inequality (a bound on the variance of smooth scalar observation of multivariate Gaussian random vector)

(3) interpolation from Gaussian to non-Gaussian, see [CL22, Section 2.2.2] for details.

Theorem (Stein's Lemma)

Let $x \sim \mathcal{N}(0,1)$ and $f : \mathbb{R} \to \mathbb{R}$ a continuously differentiable function having at most polynomial growth and such that $\mathbb{E}[f'(x)] < \infty$. Then,

$$\mathbb{E}[xf(x)] = \mathbb{E}[f'(x)]. \tag{33}$$

In particular, for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C} \in \mathbb{R}^{p \times p}$ and $f : \mathbb{R}^p \to \mathbb{R}$ a continuously differentiable function with derivatives having at most polynomial growth with respect to p,

$$\mathbb{E}[[\mathbf{x}]_i f(\mathbf{x})] = \sum_{j=1}^{p} [\mathbf{C}]_{ij} \mathbb{E}\left[\frac{\partial f(\mathbf{x})}{\partial [\mathbf{x}]_j}\right], \tag{34}$$

where $\partial / \partial [\mathbf{x}]_i$ indicates differentiation with respect to the i-th entry of $\mathbf{x}$; or, in vector form $\mathbb{E}[\mathbf{x} f(\mathbf{x})] = \mathbf{C}\mathbb{E}[\nabla f(\mathbf{x})]$, with $\nabla f(\mathbf{x})$ the gradient of $f(\mathbf{x})$ with respect to $\mathbf{x}$.

Proof of MP law with Gaussian method

First observe that $\mathbf{Q} = \frac{1}{z}\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q} - \frac{1}{z}\mathbf{I}_p$, so that $\mathbb{E}[\mathbf{Q}_{ij}] = \frac{1}{zn}\sum_{k=1}^n \mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] - \frac{1}{z}\delta_{ij}$, in which $\mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] = \mathbb{E}[xf(x)]$ for $x = \mathbf{X}_{ik}$ and $f(x) = [\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}$.
Therefore, from Stein's lemma and the fact that $\partial\mathbf{Q} = -\frac{1}{n}\mathbf{Q}\partial(\mathbf{X}\mathbf{X}^\mathsf{T})\mathbf{Q}$,[2]

$$\begin{aligned}\mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] &= \mathbb{E}\left[\frac{\partial[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}}{\partial\mathbf{X}_{ik}}\right] = \mathbb{E}[\mathbf{E}_{ik}^\mathsf{T}\mathbf{Q}]_{kj} - \mathbb{E}\left[\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{Q}(\mathbf{E}_{ik}\mathbf{X}^\mathsf{T} + \mathbf{X}\mathbf{E}_{ik}^\mathsf{T})\mathbf{Q}\right]_{kj} \\ &= \mathbb{E}[\mathbf{Q}_{ij}] - \mathbb{E}\left[\frac{1}{n}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{ki}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}\right] - \mathbb{E}\left[\frac{1}{n}[\mathbf{X}^\mathsf{T}\mathbf{Q}\mathbf{X}]_{kk}\mathbf{Q}_{ij}\right]\end{aligned}$$

for $\mathbf{E}_{ij}$ the indicator matrix with entry $[\mathbf{E}_{ij}]_{lm} = \delta_{il}\delta_{jm}$, so that, summing over k,

$$\frac{1}{z}\frac{1}{n}\sum_{k=1}^n \mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] = \frac{1}{z}\mathbb{E}[\mathbf{Q}_{ij}] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}_{ij}\operatorname{tr}(\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T})] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}]_{ij}.$$

[2]This is the matrix version of $d(1/x) = -dx/x^2$.

We have

$$\frac{1}{z}\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] = \frac{1}{z}\mathbb{E}[\mathbf{Q}_{ij}] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}_{ij}\operatorname{tr}(\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T})] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}]_{ij}.$$

The term in the second line has vanishing operator norm (of order $O(n^{-1})$) as $n, p \to \infty$. Also, $\operatorname{tr}(\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}) = np + zn \operatorname{tr}\mathbf{Q}$. As a result, matrix-wise, we obtain

$$\mathbb{E}[\mathbf{Q}] + \frac{1}{z}\mathbf{I}_p = \mathbb{E}[\mathbf{X}_{\cdot k}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{k\cdot}] = \frac{1}{z}\mathbb{E}[\mathbf{Q}] - \frac{1}{z}\frac{1}{n}\mathbb{E}[\mathbf{Q}(p + z\operatorname{tr}\mathbf{Q})] + o_{\|\cdot\|}(1),$$

where $\mathbf{X}_{\cdot k}$ and $\mathbf{X}_{k\cdot}$ is the k-th column and row of $\mathbf{X}$, respectively.
As the random $\frac{1}{p}\operatorname{tr}\mathbf{Q} \to m(z)$ as $n, p \to \infty$, "take it out of the expectation" in the limit and

$$\mathbb{E}[\mathbf{Q}](1 - p/n - z - p/n \cdot zm(z)) = \mathbf{I}_p + o_{\|\cdot\|}(1),$$

which, taking the trace to identify $m(z)$, concludes the proof.

Nash–Poincaré inequality and Interpolation trick

Theorem (Nash–Poincaré inequality)

For $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C} \in \mathbb{R}^{p \times p}$ and $f : \mathbb{R}^p \to \mathbb{R}$ continuously differentiable with derivatives having at most polynomial growth with respect to p,

$$\operatorname{Var}[f(\mathbf{x})] \leq \sum_{i,j=1}^{p} [\mathbf{C}]_{ij} \mathbb{E}\left[\frac{\partial f(\mathbf{x})}{\partial [\mathbf{x}]_i} \frac{\partial f(\mathbf{x})}{\partial [\mathbf{x}]_j} \right] = \mathbb{E}\left[(\nabla f(\mathbf{x}))^{\mathsf{T}} \mathbf{C} \nabla f(\mathbf{x}) \right],$$

where we denote $\nabla f(\mathbf{x})$ the gradient of $f(\mathbf{x})$ with respect to $\mathbf{x}$.

Theorem (Interpolation trick)

For $x \in \mathbb{R}$ a random variable with zero mean and unit variance, $y \sim \mathcal{N}(0,1)$, and f a $(k+2)$-times differentiable function with bounded derivatives,

$$\mathbb{E}[f(x)] - \mathbb{E}[f(y)] = \sum_{\ell=2}^{k} \frac{\kappa_{\ell+1}}{2\ell!} \int_0^1 \mathbb{E}[f^{(\ell+1)} x(t)] t^{(\ell-1)/2} dt + \epsilon_k,$$

where κ_ℓ is the ℓ^{th} cumulant of x, $x(t) = \sqrt{t}x + (1-\sqrt{t})y$, and $|\epsilon_k| \leq C_k \mathbb{E}[|x|^{k+2}] \cdot \sup_t |f^{(k+2)}(t)|$ for some constant C_k only dependent on k.

Take-away of this section

- p-by-p SCM $\hat{\mathbf{C}}$ from n samples have different behavior in the **classical** ($n \gg p$) versus **proportional** ($n \sim p$) regime
- four ways to characterize SCM, asymptotic and non-asymptotic fashion
- "**old school**" results: (1) LLN and (2) matrix concentration in the classical regime, and (3) asymptotic Marčenko-Pastur law on SCM eigenvalues in the proportional regime
- **modern** approach of **deterministic equivalent for SCM resolvent**, both (4) asymptotic and (5) non-asymptotic
- proof via "leave-one-out" and self-consistent equation
- alternative proof via Gaussian method

Wigner semicircle law

Theorem (Wigner semicircle law)

Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ *be symmetric and such that the* $\mathbf{X}_{ij} \in \mathbb{R}$, $j \geq i$, *are independent zero mean and unit variance random variables. Then, for* $\mathbf{Q}(z) = (\mathbf{X}/\sqrt{n} - z\mathbf{I}_n)^{-1}$, *as* $n \to \infty$,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_n, \tag{35}$$

with $m(z)$ *the unique Stieltjes transform solution to*

$$m^2(z) + zm(z) + 1 = 0. \tag{36}$$

The function $m(z)$ *is the Stieltjes transform of the probability measure*

$$\mu(dx) = \frac{1}{2\pi}\sqrt{(4 - x^2)^+}\,dx, \tag{37}$$

known as the Wigner semicircle law.

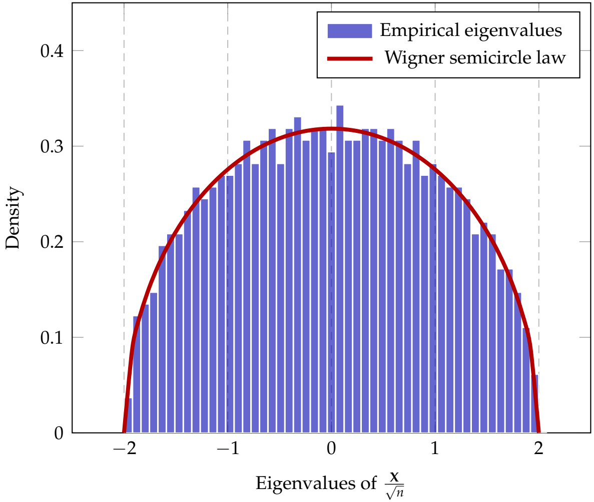


Figure: Histogram of the eigenvalues of $\mathbf{X}/\sqrt{n}$ versus Wigner semicircle law, for standard Gaussian $\mathbf{X}$ and $n = 1\,000$.

Theorem (General sample covariance matrix)

Let $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z} \in \mathbb{R}^{p\times n}$ *with nonnegative definite* $\mathbf{C} \in \mathbb{R}^{p\times p}$, $\mathbf{Z} \in \mathbb{R}^{p\times n}$ *having independent zero mean and unit variance entries. Then, as* $n, p \to \infty$ *with* $p/n \to c \in (0, \infty)$, *for* $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ *and* $\tilde{\mathbf{Q}}(z) = (\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} - z\mathbf{I}_n)^{-1}$,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = -\frac{1}{z}\left(\mathbf{I}_p + \tilde{m}_p(z)\mathbf{C}\right)^{-1}, \quad \tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\tilde{\mathbf{Q}}}(z) = \tilde{m}_p(z)\mathbf{I}_n,$$

with $\tilde{m}_p(z)$ *unique solution to*

$$\tilde{m}_p(z) = \left(-z + \frac{1}{n}\operatorname{tr}\mathbf{C}\left(\mathbf{I}_p + \tilde{m}_p(z)\mathbf{C}\right)^{-1}\right)^{-1}. \tag{38}$$

Moreover, if the empirical spectral measure of $\mathbf{C}$ *converges* $\mu_{\mathbf{C}} \to \nu$ *as* $p \to \infty$, *then* $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}} \to \mu$, $\mu_{\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}} \to \tilde{\mu}$ *where* $\mu, \tilde{\mu}$ *admitting Stieltjes transforms* $m(z)$ *and* $\tilde{m}(z)$ *such that*

$$m(z) = \frac{1}{c}\tilde{m}(z) + \frac{1-c}{cz}, \quad \tilde{m}(z) = \left(-z + c\int \frac{t\nu(dt)}{1+\tilde{m}(z)t}\right)^{-1}. \tag{39}$$

A few remarks on the generalized MP law

- different from the explicit MP law, the generalized MP is in general implicit
- we have explicitness in essence due to with $\mathbf{C} = \mathbf{I}_p$, the implicit equation boils down to a quadratic equation that has explicit solution
- if $\mathbf{C}$ has discrete eigenvalues, e.g., $\mu_{\mathbf{C}} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_5)$, then becomes a (possibly higher-order) polynomial equation, which may admit explicit solution (up to fourth order) using radicals
- the uniqueness of (Stieltjes transform) solution is ensured within a certain region on the complex plane, there may exist solutions $\tilde{m}(z)$ with imaginary parts of wrong sign
- **numerical evaluation of $\tilde{m}(z)$**: note that the equation

$$\tilde{m}_p(z) = \left(-z + \frac{1}{n} \operatorname{tr} \mathbf{C} \left(\mathbf{I}_p + \tilde{m}_p(z) \mathbf{C} \right)^{-1} \right)^{-1} \tag{40}$$

naturally defines a fixed-point equation.

Matlab code

```
clear i % make sure i stands for the imaginary unit
y = 1e-5;
zs = edges_mu+y*1i;
mu = zeros(length(zs),1);

tilde_m=0;
for j=1:length(zs)
    z = zs(j);

    tilde_m_tmp=-1;
    while abs(tilde_m-tilde_m_tmp)>1e-6
        tilde_m_tmp=tilde_m;
        tilde_m = 1/( -z + 1/n*sum(eigs_C./(1+tilde_m*eigs_C)) );
    end

    m = tilde_m/c+(1-c)/(c*z);
    mu(j)=imag(m)/pi;
end
```

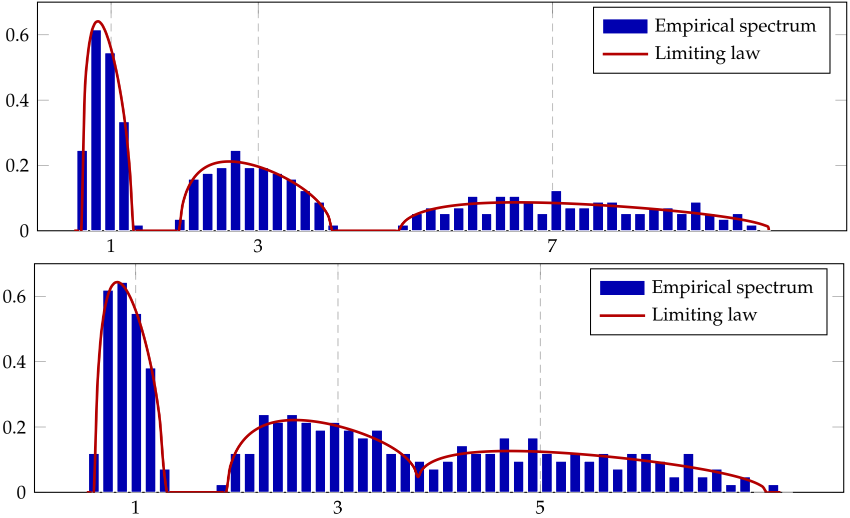


Figure: Histogram of the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}$, $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z} \in \mathbb{R}^{p \times n}$, $[\mathbf{Z}]_{ij} \sim \mathcal{N}(0,1)$, $n = 3\,000$; for $p = 300$ and $\mathbf{C}$ having spectral measure $\mu_{\mathbf{C}} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_7)$ **(top)** and $\mu_{\mathbf{C}} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_5)$ **(bottle)**.

Further comments on generalized SCM

- we know a lot more for the generalized SCM model: precise characterization of the support of its (limiting) eigenspectrum
- applications in **statistical inference**: given $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ SCM of the population covariance $\mathbf{C}$, infer eigenspectral functions of $\mathbf{C}$ using those of $\hat{\mathbf{C}}$ and wisely-chosen contour integration, etc.

Example: estimation of population eigenvalues of large multiplicity

Consider the following SCM inference,

$$\nu_\mathbf{C} = \frac{1}{p}\sum_{i=1}^K p_i \delta_{\ell_i} \to \sum_{i=1}^K c_i \delta_{\ell_i}$$

for $\ell_1 > \ldots > \ell_K > 0$, K fixed/small with respect to n, p, and $p_i/p \to c_i > 0$ as $p \to \infty$, i.e., each eigenvalue has a large multiplicity of order $O(p)$.

- **native** estimator: $\hat{\ell}_a = \frac{1}{p_a}\sum_{i=p_1+\ldots+p_{a-1}+1}^{p_1+\ldots+p_a} \lambda_i$
- **RMT-improved** estimator: $\hat{\ell}_a = \frac{n}{p_a}\sum_{i=p_1+\ldots+p_{a-1}+1}^{p_1+\ldots+p_a} (\lambda_i - \eta_i)$, with λ_i eigenvalues of $\hat{\mathbf{C}}$ and η_i eigenvalues of $\mathbf{\Lambda} - \frac{1}{n}\sqrt{\boldsymbol{\lambda}}\sqrt{\boldsymbol{\lambda}}^\mathsf{T}$, $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_i\}_{i=1}^p$ and $\sqrt{\boldsymbol{\lambda}} \in \mathbb{R}^p$ the vector of $\sqrt{\lambda_i}$s.

- see [CL22, Sections 2.3 and 2.4] for detailed derivations and discussions

Numerical results

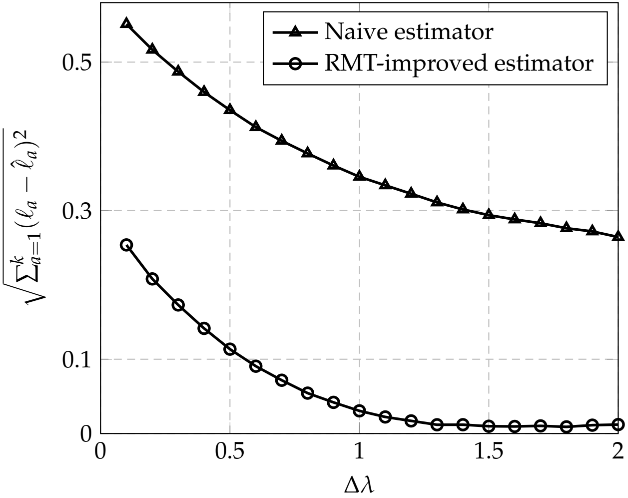


Figure: Eigenvalue estimation errors with naive and RMT-improved approach, as a function of $\Delta\lambda$, for $\ell_1 = 1$, $\ell_2 = 1 + \Delta\lambda$, $p = 256$ and $n = 1\,024$. Results averaged over 30 runs.

Separable covariance model: motivation

- data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ arise from a time series, each data vector is weighted by a coefficient
- SCM can be generalized to the so-called **bi-correlated** (or **separable covariance**) model

$$\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} = \frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\tilde{\mathbf{C}}\mathbf{Z}^\mathsf{T}\mathbf{C}^{\frac{1}{2}} \tag{41}$$

  for $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $\tilde{\mathbf{C}} \in \mathbb{R}^{n\times n}$ two nonnegative definite matrices and $[\mathbf{Z}]_{ij}$ i.i.d. random variables with zero mean and unit variance.
- in particular, for $\mathbf{Z}$ Gaussian and $\tilde{\mathbf{C}}^{\frac{1}{2}}$ Toeplitz (i.e., such that $[\tilde{\mathbf{C}}^{\frac{1}{2}}]_{ij} = \alpha_{|i-j|}$ for some sequence $\alpha_0, \ldots, \alpha_{n-1}$), the columns of $\mathbf{Z}\tilde{\mathbf{C}}^{\frac{1}{2}}$ model a first order auto-regressive process

Separable covariance model

Theorem (Bi-correlated model, separable covariance model, [PS09])

Let $\mathbf{Z} \in \mathbb{R}^{p \times n}$ be a random matrix with i.i.d. zero mean, unit variance and light tail entries, and $\mathbf{C} \in \mathbb{R}^{p \times p}$, $\tilde{\mathbf{C}} \in \mathbb{R}^{n \times n}$ be symmetric nonnegative definite matrices with bounded operator norm. Then, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, letting $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\tilde{\mathbf{C}}\mathbf{Z}^\mathsf{T}\mathbf{C}^{\frac{1}{2}} - z\mathbf{I}_p)^{-1}$ and $\tilde{\mathbf{Q}}(z) = (\frac{1}{n}\tilde{\mathbf{C}}^{\frac{1}{2}}\mathbf{Z}^\mathsf{T}\mathbf{C}\mathbf{Z}\tilde{\mathbf{C}}^{\frac{1}{2}} - z\mathbf{I}_n)^{-1}$, we have

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = -\frac{1}{z}\left(\mathbf{I}_p + \tilde{\delta}_p(z)\mathbf{C}\right)^{-1}, \quad \tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\tilde{\mathbf{Q}}}(z) = -\frac{1}{z}\left(\mathbf{I}_n + \delta_p(z)\tilde{\mathbf{C}}\right)^{-1}$$

with $(z, \delta_p(z)), (z, \tilde{\delta}_p(z)) \in \mathcal{Z}(\mathbb{C} \setminus \mathbb{R}^+)$ unique solutions to

$$\delta_p(z) = \frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}(z), \quad \tilde{\delta}_p(z) = \frac{1}{n}\operatorname{tr}\tilde{\mathbf{C}}\bar{\tilde{\mathbf{Q}}}(z).$$

In particular, if $\mu_{\mathbf{C}} \to \nu$ and $\mu_{\tilde{\mathbf{C}}} \to \tilde{\nu}$, then $\mu_{\frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\tilde{\mathbf{C}}\mathbf{Z}^\mathsf{T}\mathbf{C}^{\frac{1}{2}}} \xrightarrow{a.s.} \mu$, $\mu_{\frac{1}{n}\tilde{\mathbf{C}}^{\frac{1}{2}}\mathbf{Z}^\mathsf{T}\mathbf{C}\mathbf{Z}\tilde{\mathbf{C}}^{\frac{1}{2}}} \xrightarrow{a.s.} \tilde{\mu}$,, where $\mu, \tilde{\mu}$ are defined by their Stieltjes transforms $m(z)$ and $\tilde{m}(z)$ given by

$$m(z) = -\frac{1}{z}\int\frac{\nu(dt)}{1+\tilde{\delta}(z)t}, \quad \tilde{m}(z) = -\frac{1}{z}\int\frac{\tilde{\nu}(dt)}{1+\delta(z)t}, \quad \delta(z) = -\frac{c}{z}\int\frac{t\nu(dt)}{1+\tilde{\delta}(z)t}, \quad \tilde{\delta}(z) = -\frac{1}{z}\int\frac{t\tilde{\nu}(dt)}{1+\delta(z)t}$$

[4] Debashis Paul and Jack W. Silverstein. "No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix". In: Journal of Multivariate Analysis 100.1 (2009), pp. 37–57.

Take-away messages of this section

Asymptotic Deterministic Equivalent for resolvent results for

- symmetric $\mathbf{X}/\sqrt{n} \in \mathbb{R}^{n\times n}$: **Wigner semicircle law**, quadratic equation (again)
- **generalized SCM model** $\frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\mathbf{Z}^\mathsf{T}\mathbf{C}^{\frac{1}{2}}$: one self-consistent but integral equation
- application to **inference** of SCM eigenspectral functionals
- **bi-correlated model** or **separable covariance model** $\frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\tilde{\mathbf{C}}\mathbf{Z}^\mathsf{T}\mathbf{C}^{\frac{1}{2}}$: two coupled self-consistent integral equations

Thank you! Q & A?