Spectral Analysis of Multiscale Sample Covariance Matrices

Joseph Genzer

Tulane University
Department of Mathematics

May 1, 2025

Overview

1. Introduction
2. Fixed scale, low dimensions
3. Fixed scale, high dimensions
4. Multiscale, high dimensions
5. References
6. Appendix

Random Matrices

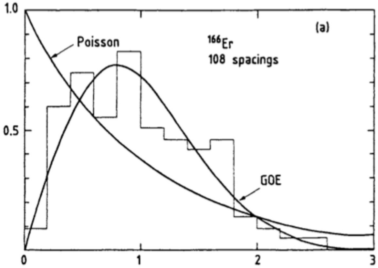
What are Random Matrices?

A *random matrix* is a matrix whose entries are random variables.
Examples include Wishart matrices, Ginibre matrices, and Wigner matrices.

Why are Random Matrices important?

Random matrices are fundamental in probability theory, statistics, and mathematical physics.

- The distributions of nuclear resonances closely match eigenvalue spacing distributions in the *Gaussian Orthogonal Ensemble* (GOE).
- Sample covariance matrices are widely used in statistics and data science.



Fixed Scale, Low Dimensions

Fixed Scale Sample Covariance Matrix

Given a set of random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$, we define

$$\hat{\Sigma}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$$

where n is the sample size and p is the number of variables.

Low-Dimensional Setting

In low dimensions, p is fixed while $n \to \infty$. For example, in the bivariate case $(p = 2)$

$$\hat{\Sigma}_n = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

where $\sigma_1^2 = \mathrm{Var}(X_1)$, $\sigma_2^2 = \mathrm{Var}(X_2)$, and $\sigma_{12} = \mathrm{Cov}(X_1, X_2)$.

Fixed Scale, Low Dimensional Eigenvalue Asymptotics

Low Dimensional Eigenvalue Asymptotics

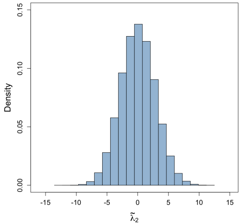
For independent $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$, the largest eigenvalue of $\hat{\Sigma}_n$ is asymptotically normal. That is, if we let $\widetilde{\lambda}_2 = \sqrt{n}\left(\lambda_2(\hat{\Sigma}_n) - 2\right)$, then

$$\widetilde{\lambda}_2 \xrightarrow{d} \mathcal{N}(0,8), \quad n \to \infty.$$

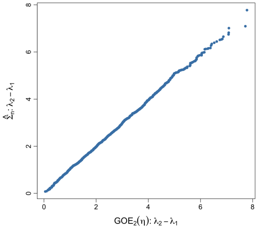
For independent $X_1, X_2 \sim \mathcal{N}(0,1)$, the eigenvalues of $\hat{\Sigma}_n$ are asymptotically GOE-like

$$\sqrt{n}(\hat{\Sigma}_n - I_2) \xrightarrow{d} \mathrm{GOE}_2(\eta), \quad n \to \infty.$$

Distribution of $\hat{\Sigma}_n$ Top Eigenvalue



QQ Plot: $\lambda_2 - \lambda_1$ of $\hat{\Sigma}_n$ vs $\mathrm{GOE}_2(\eta)$



$n = 1000$, $p = 2$, $R = 1000$

Fixed Scale, High Dimensions

High-Dimensional Setting

In high dimensions, both the sample size n and the number of variables p grow to infinity while maintaining a fixed ratio

$$n, p \to \infty, \quad \frac{p}{n} \to c \in (0, \infty)$$

Applications and Motivations

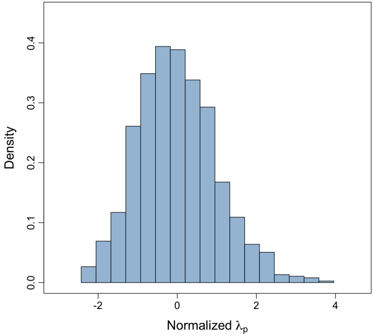
- Genetics: $n =$ sample size, $p =$ number of genes
- Neurology: $n =$ time, $p =$ the number of sensors

Fixed Scale, High Dimensional Eigenvalue Asymptotics

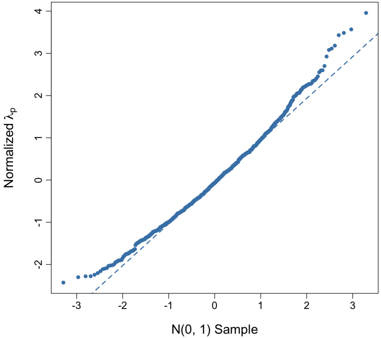
Tracy-Widom Law

Unlike in the low-dimensional case, where the top eigenvalue of $\hat{\Sigma}_n$ is asymptotically normal, in high dimensions the top eigenvalue follows the *Tracy-Widom* law.

Distribution of Normalized $\lambda_p(\hat{\Sigma}_n)$



QQ Plot: Normalized $\lambda_p(\hat{\Sigma}_n)$ vs $\mathcal{N}(0,1)$



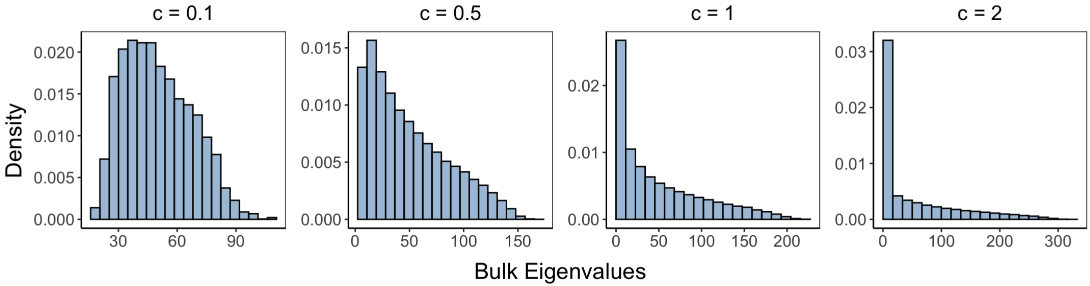
$$n = 5000, p = 2500, R = 1000$$

Fixed Scale, High Dimensional Eigenvalue Asymptotics

Marchenko–Pastur Law

In high dimensions, the bulk eigenvalue distribution of $\hat{\Sigma}_n$ converges to the *Marchenko–Pastur* law, which describes the limiting density as a function of $c = \frac{p}{n}$.

Bulk Eigenvalue Distribution of $\hat{\Sigma}_n$



$$n = 5000, \quad c \in \{0.1, 0.5, 1, 2\}$$

Fractional Brownian Motion

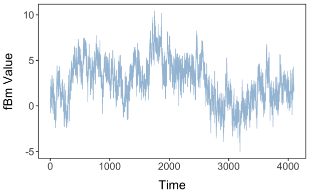
Why consider dependence structures?

Fractional Brownian motion is a universal model for long-range dependencies, and has applications in biophysics and network traffic.

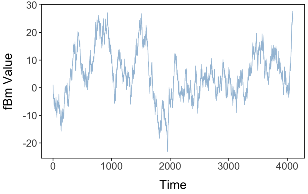
Fractional Brownian Motion (fBm)

A Gaussian process $B_H(t)$ with stationary increments, formed by cumulatively summing fractional Gaussian noise. The Hurst parameter, $H \in (0, 1)$, controls memory:

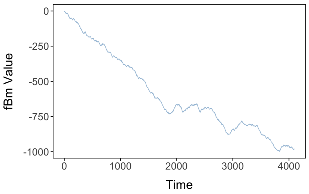
$H < 0.5$: **Anti-persistence**



$H = 0.5$: **Brownian Motion**



$H > 0.5$: **Long-range dependence**



Multi Scale, High Dimensions

Multiscale Sample Covariance Matrix

To analyze dependence across time, we define

$$M(h) = \frac{1}{n-h}\sum_{t=1}^{n-h}(Y(t+h) - Y(t))(Y(t+h) - Y(t))^\top \in \mathbb{R}^{p^2}$$

where h is the lag parameter and $Y \in \mathbb{R}^p$ is a multivariate fBm process.

Multi Scale, High Dimensions

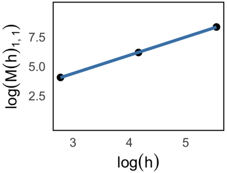
Scaling Laws ($p = 2$)

The growth of the diagonal entries and eigenvalues of $M(h)$ follows a power law in h

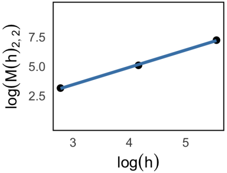
Diagonal Elements

$$M(h)_{1,1}, M(h)_{2,2} \overset{\mathbb{P}}{\sim} |h|^{2H_2}$$

$\alpha_{1,1} = 1.517$



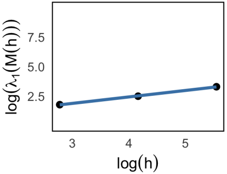
$\alpha_{2,2} = 1.492$



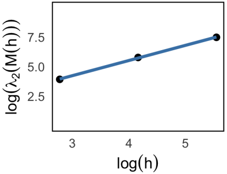
Eigenvalues

$$\lambda_1(M(h)) \overset{\mathbb{P}}{\sim} |h|^{2H_1}, \quad \lambda_2(M(h)) \overset{\mathbb{P}}{\sim} |h|^{2H_2}$$

$\alpha_{\lambda_1} = 0.514$



$\alpha_{\lambda_2} = 1.533$



$$n = 2^{16}, p = 2, h \in \{2^4, 2^6, 2^8\}, H_1 = 0.25, H_2 = 0.75$$

Large Scale, High Dimensions

Consider the high-dimensional model

$$Y(t) = PX(t) + Z(t),$$

where

- $P \in \mathbb{R}^{p \times 2}$ contains a 2×2 orthogonal matrix $O_2 \in \mathrm{O}(2)$ in its top two rows and zeros elsewhere,
- $X(t) \in \mathbb{R}^2$ is a bivariate fBm with independent components $B_{H_1}(t)$ and $B_{H_2}(t)$,
- $Z(t) \in \mathbb{R}^p$ is Gaussian noise.

Large Scale, High Dimensional Eigenvalue Asymptotics

Consider the limit $n, p, h \to \infty$ with $\frac{ph}{n} = O(1)$, known as the large-scale, high dimensional setting. In this case, the asymptotic fluctuations of the top two eigenvalues are given by

$$\sqrt{\frac{n}{h}} (\log \lambda_\ell(M(h)) - \log \lambda_\ell(\mathbb{E}M(h))) \xrightarrow{d} \widetilde{\lambda}_\ell, \quad \ell = p-1, p$$

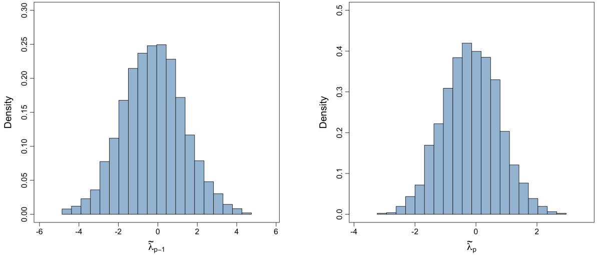
Large Scale, High Dimensional Eigenvalue Asymptotics

For $0 < H_1 < H_2 < 0.75$

The top two eigenvalues are asymptotically **jointly Gaussian** and **independent**.

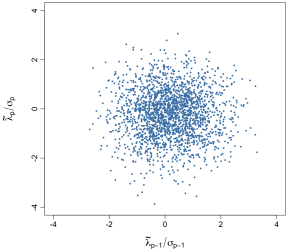
Distribution of Top Two Eigenvalues of M(h)

K-S Test: $p = 0.72, 0.89$



Normalized Top Two Eigenvalues of M(h)

Correlation $= 0.01$



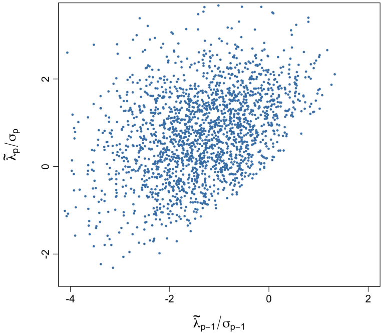
$n = 2^{16}, p = 50, h = 2^{12}, H_1 = 0.3, H_2 = 0.7, R = 1000$

Large Scale, High Dimensional Eigenvalue Asymptotics

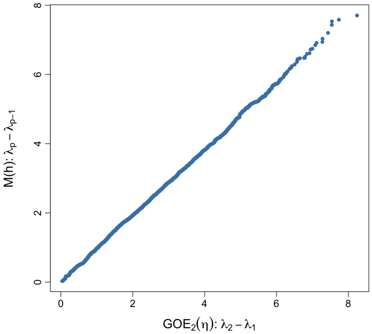
For $0 < H_1 = H_2 < 0.75$

The top two eigenvalues are asymptotically **non-Gaussian** with **GOE-like repulsion**

Scatter Plot: Top Two Eigenvalues of M(h)
Correlation = 0.46



QQ Plot: $\lambda_p - \lambda_{p-1}$ of M(h) vs GOE
K-S Test: $p = 0.78$



$n = 2^{16}$, $p = 50$, $h = 2^{12}$, $H_1 = H_2 = 0.6$, $R = 1000$

Large Scale, High Dimensional Eigenvalue Asymptotics

Phase Transition

There exists a phase transition at the critical value $H = 0.75$ that controls the eigenvalue asymptotics of $M(h)$, emerging from a squared summability constraint on the autocovariance function.

$(p = 1)$ Since $\gamma_H(k) \sim Ck^{2H-2}$, it follows that

$$|\gamma_H(k)|^2 \sim \frac{C^2}{k^{2(2-2H)}}$$

Applying the summability constraint, we have

$$\sum_{k=1}^{\infty} |\gamma_H(k)|^2 < \infty \quad \Longleftrightarrow \quad 2(2-2H) > 1 \quad \Longleftrightarrow \quad H < 0.75$$

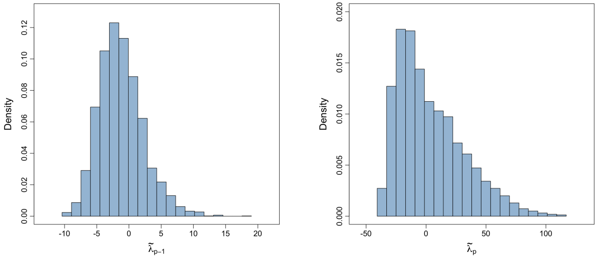
Large Scale, High Dimensional Eigenvalue Asymptotics

For $0.75 < H_1 < H_2 < 1$

The top two eigenvalues are asymptotically **non-Gaussian** and **independent**

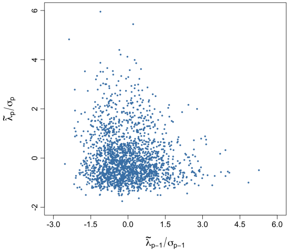
Distribution of Top Two Eigenvalues of M(h)

Both K-S Tests: $p << 0.05$



Normalized Top Two Eigenvalues of M(h)

Correlation $= 0.03$



$$n = 2^{16}, p = 2, h = 2^{12}, H_1 = 0.8, H_2 = 0.9, R = 1000$$

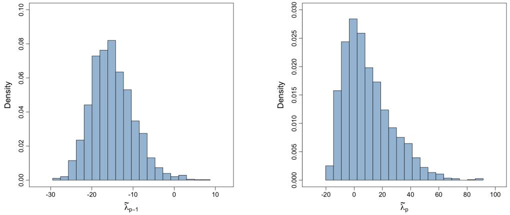
Large Scale, High Dimensional Eigenvalue Asymptotics

For $0.75 < H_1 = H_2 < 1$

The top two eigenvalues are asymptotically **non-Gaussian** with **non-GOE-like repulsion**

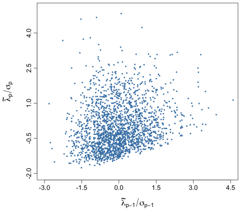
Distribution of Top Two Eigenvalues of M(h)

Both K-S Tests: $p << 0.05$



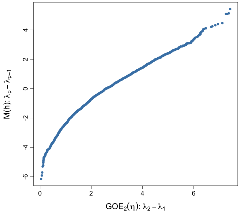
Scatter Plot: Top Two Eigenvalues of M(h)

Correlation $= 0.35$



QQ Plot: $\lambda_p - \lambda_{p-1}$ of M(h) vs GOE

K-S Test: $p << 0.05$



$$n = 2^{16}, p = 2, h = 2^{12}, H_1 = H_2 = 0.9, R = 1000$$

Post-Phase Transition β Estimation

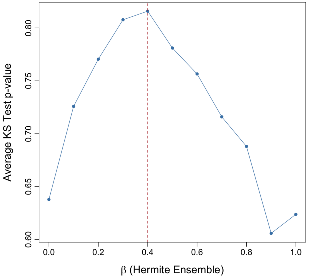
Best-Fit β Estimation

For $0.75 < H_1 = H_2 < 1$, $M(h)$ exhibits eigenvalue dependence without GOE-like repulsion.

Since the GOE corresponds to the β-Hermite ensemble with $\beta = 1$, we estime the best-fit β to model eigenvalue spacing of $M(h)$ by maximizing the average K-S Test p value.

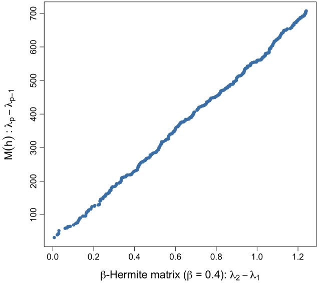
Maximizing p value in terms of β

Average over 50 runs



QQ Plot: $\lambda_p - \lambda_{p-1}$ of M(h) vs $\beta = 0.4$ Matrix

K-S Test: $p = 0.83$



$$n = 2^{16}, p = 2, h = 2^{12}, H_1 = H_2 = 0.9, R = 1000$$

Key Ideas

Fixed Scale Analysis (no dependence)

- In **low dimensions**, eigenvalue distributions are asymptotically Gaussian or GOE-like.
- In **high dimensions**, we see the emergence of the Tracy-Widom and Marchenko-Pastur laws.

Multiscale Analysis

- In **low dimensions**, while the entries of $M(h)$ alone cannot recover the underlying dependence structure, the eigendomain is able to do so.
- In **high dimensions**, eigenvalue asymptotics revert to having Gaussian or GOE-like limits, up until the phase transition.

Thank you for your attention!

Acknowledgments

Thank you Dr. Gustavo Didier and Dr. Ken McLaughlin for serving as my readers and for their guidance and mentorship throughout the course of this thesis. Their encouragement, high standards, and deep engagement with the material have left a lasting impact, for which I am deeply grateful.

References

Garrett, J. D., German, J. R., and Espino, J. M. (1995).
Nuclear level repulsion, order vs. chaos and conserved quantum numbers.
In *Lecture Notes in Physics*, volume 441, pages 59–72. Springer, Berlin, Heidelberg.

Mandelbrot, B. B. and Ness, J. W. V. (1968).
Fractional brownian motions, fractional noises and applications.
SIAM Review, 10(4):422–437.

Appendix

Tracy-Widom Law

As $n, p \to \infty$ with $p/n \to c$, the top eigenvalue distribution satisfies

$$n^{\frac{2}{3}} \left(\frac{\lambda_p(\hat{\Sigma}_n) - \mu_{n,p}}{\sigma_{n,p}} \right) \xrightarrow{d} F_1, \quad \mu_{n,p} \to (1+\sqrt{c})^2, \quad \sigma_{n,p} \to (1+\sqrt{c}) \left(1 + \sqrt{\frac{1}{c}} \right)^{\frac{1}{3}}$$

$F_1(s)$ is given by:

$$F_1(s) = \exp\left(-\frac{1}{2} \int_s^\infty (q(x) + (x-s)q^2(x))dx \right)$$

where $q(x)$ satisfies the Painlevé II equation:

$$q''(x) = xq(x) + 2q^3(x), \quad q(x) \sim \mathrm{Ai}(x) \text{ as } x \to \infty$$

Appendix

Marchenko-Pastur Law

As $n, p \to \infty$ with $p/n \to c$, the bulk eigenvalue distribution satisfies the density function given by

$$\rho_c(\lambda) = \begin{cases} (c-1)\delta(\lambda) + \frac{1}{2\pi c\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)} & \text{if } c > 1, \\ \frac{1}{2\pi c\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)} & \text{if } 0 < c \leq 1, \end{cases}$$

for $\lambda \in [\lambda_-, \lambda_+]$, where $\lambda_\pm = (1 \pm \sqrt{c})^2$, and $\delta(\lambda)$ is the Dirac delta function centered at zero.