Spiked Covariance Model

Victor Qian

June 2025

Abstract

This note introduces the Marčenko–Pastur (MP) law for high-dimensional sample covariance matrices, interprets its support and density, and then presents the Baik–Ben Arous–P'ech'e (BBP) spiked-model extension describing phase transitions of outlier eigenvalues and eigenvector consistency.

1 Introduction

In high-dimensional statistics, one often encounters sample covariance matrices of the form $S = \frac{1}{T} X^\top X$, where X is a $T \times N$ data matrix with independent entries of mean zero and variance σ^2. When both N and T grow large with ratio $\gamma = N/T$, classical random matrix theory yields the *Marčenko–Pastur law*, characterizing the limiting spectral distribution of S.

2 The Marčenko–Pastur Law

2.1 Setting and Notation

Let

- $X = (X_{t,i})_{1 \le t \le T, 1 \le i \le N}$ with $\mathbb{E}[X_{t,i}] = 0$, $\mathrm{Var}(X_{t,i}) = \sigma^2$, and finite fourth moments,
- $S = \frac{1}{T} X^\top X$ be the $N \times N$ sample covariance matrix,
- $\gamma_N = N/T \to \gamma \in (0, \infty)$ as $N, T \to \infty$.

2.2 Limiting Density

u

2.3 Interpretation

- *Bulk support:* All noise eigenvalues lie in $[\lambda_-, \lambda_+]$ in the limit.
- *Aspect ratio:* $\gamma < 1$ (skinny) versus $\gamma > 1$ (fat) only affects whether there is an atom at zero of mass $1 - 1/\gamma$.
  1. Data-rich regime ($\gamma \ll 1$): Bulk tightly concentrated around the true variance; spikes are easier to spot and well-estimated.
  2. Balanced regime ($\gamma \approx 1$): Bulk is wide, noise variance is high; only very strong signals (eigenvalues > 2 or more) will separate.
  3. Data-poor regime ($\gamma > 1$): Many zero modes; the non-zero noise bulk still sits in $[\lambda_-, \lambda_+]$, but you must discard the zero-eigenvalue subspace entirely. Signal detection requires even stronger spikes.
- *Edge behavior:* The edges $\lambda_\pm$ set thresholds for separating signal eigenvalues from noise.

3 Spiked Model and BBP Phase Transition

In practice, the population covariance may have a finite number m of "spikes"

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > \sigma^2,$$

with the remaining $N - m$ eigenvalues equal to σ^2. This is the *spiked model*. Baik, Ben Arous, and Péché (BBP) showed:

3.1 Outlier Eigenvalues

Define

$$\gamma = \frac{N}{T}, \qquad \lambda_\pm = \sigma^2 \, (1 \pm \sqrt{\gamma})^2,$$

so that the MP bulk from pure noise lies on $[\lambda_-, \lambda_+]$. For each spike λ_i:

- **Supercritical:** If

$$\lambda_i > \sigma^2 \, (1 + \sqrt{\gamma}),$$

then the sample eigenvalue $\hat{\lambda}_i$ *detaches* from the bulk and converges a.s. to

$$\hat{\lambda}_i \;\longrightarrow\; \mu_i \;=\; \lambda_i\Big(1+\frac{\gamma\,\sigma^2}{\lambda_i-\sigma^2}\Big) \;>\; \lambda_+.$$

- **Subcritical:** If

$$\lambda_i \le \sigma^2\,(1+\sqrt{\gamma}),$$

then $\hat{\lambda}_i \to \lambda_+$ in probability and remains buried in the bulk.

3.2 Eigenvector Consistency

Let u_i be the population eigenvector for spike λ_i, and $\widehat{u}_i$ its sample estimate. Then almost surely

$$\big|\langle u_i, \widehat{u}_i\rangle\big| \;\longrightarrow\; \begin{cases} \big(1+c_i\big)^{-\frac{1}{2}}, & \text{if } \lambda_i > \sigma^2\,(1+\sqrt{\gamma}),\\ 0, & \text{if } \lambda_i \le \sigma^2\,(1+\sqrt{\gamma}), \end{cases} \qquad c_i \;=\; \frac{\gamma\,\sigma^2}{\lambda_i-\sigma^2}.$$

4 Practical Insights and Experimental Implications

Even though the BBP result is derived under the simplest spiked model (one or finitely many spikes), it delivers several intuitions that practitioners can test and extend. Assume the noise variance is 1:

1. **Upward bias of spike estimates.** Sample eigenvalues corresponding to population spikes are always biased upward, by an amount that shrinks as the true spike λ_i increases. When λ_i is close to the bulk (i.e. ≈ 1), the probability that the largest observed eigenvalue actually comes from noise is non-negligible, inducing extra bias.

2. **Detectability threshold.** The critical value $1+\sqrt{\gamma}$ delineates the phase transition: true eigenvalues above this threshold yield observable outliers that

   - separate cleanly from the MP bulk at λ_+,
   - exhibit stronger upward bias (so the outlier gap grows), and

- produce sample eigenvectors with significant alignment to the population directions.

3. **Invisibility below threshold.** True spikes below $1 + \sqrt{\gamma}$, even if strictly greater than 1, merge with the noise bulk. Their empirical eigenvalues stick to the edge λ_+, and their eigenvectors carry no reliable information about the true directions.

4. **Multiple-spike extensions.** Similar phenomena hold when multiple spikes coexist, though one must assume sufficient separation between spikes and more intricate technical conditions (e.g. non-overlapping phases, distinct signal strengths). The same thresholding intuition guides modern covariance denoising and principal component selection in applications.

5. **Experimental verification.** In simulated Gaussian or heavy-tailed data, one can observe:

   - empirical eigenvalue histograms matching the MP density,
   - outlier emergence as λ_i crosses $1 + \sqrt{\gamma}$,
   - decreasing eigenvector misalignment $1 - |\langle u_i, \widehat{u}_i \rangle|$ for stronger signals.

5 Conclusion

The MP law provides the baseline noise spectrum in high dimensions; the BBP extension tells us when low-rank signals generate detectable outliers and how reliably we recover their directions. These insights underpin practical covariance cleaning, PCA trimming, and signal-detection workflows.

References

[1] V.A. Marčenko and L.A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Math. USSR–Sb., 1967.

[2] J. Baik, G. Ben Arous, and S. P'ech'e, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, Ann. Probab., 2005.