Missing or Corrupted Data and the Marčenko-Pastur Theorem

Michael Stolz, University of Münster

Angers, June 2016

Sample covariance matrices

$(X_1^{(1)}, \ldots, X_p^{(1)})^T, \ldots, (X_1^{(n)}, \ldots, X_p^{(n)})^T$ iid random p-vectors.

$$\hat{\Sigma} := \frac{1}{n} X X^T - \overline{X}\,\overline{X}^T,$$

where $\overline{X} = (\overline{X_1}, \ldots, \overline{X_p})^T \in \mathbb{R}^p$ and $X = (X^{(1)}, \ldots, X^{(n)}) \in \mathbb{R}^{p \times n}$.

Classical fixed p asymptotics

Let $X^{(1)} \sim \mathrm{N}(0, \Sigma)$. If p is fixed and $n \to \infty$, then $\hat{\Sigma}$ is a consistent estimator for Σ, and the eigenvalues of $\hat{\Sigma}$ are consistent estimators for the eigenvalues of Σ.

Classical fixed p asymptotics

Let $X^{(1)} \sim \mathrm{N}(0, \Sigma)$. If p is fixed and $n \to \infty$, then $\hat{\Sigma}$ is a consistent estimator for Σ, and the eigenvalues of $\hat{\Sigma}$ are consistent estimators for the eigenvalues of Σ.

So in the case $\Sigma = \mathrm{Id}_p$, one would expect the eigenvalues of $\hat{\Sigma}$ to be close to 1 for reasonably large n.

Classical fixed p asymptotics

Let $X^{(1)} \sim \mathrm{N}(0, \Sigma)$. If p is fixed and $n \to \infty$, then $\hat{\Sigma}$ is a consistent estimator for Σ, and the eigenvalues of $\hat{\Sigma}$ are consistent estimators for the eigenvalues of Σ.

So in the case $\Sigma = \mathrm{Id}_p$, one would expect the eigenvalues of $\hat{\Sigma}$ to be close to 1 for reasonably large n.

But...

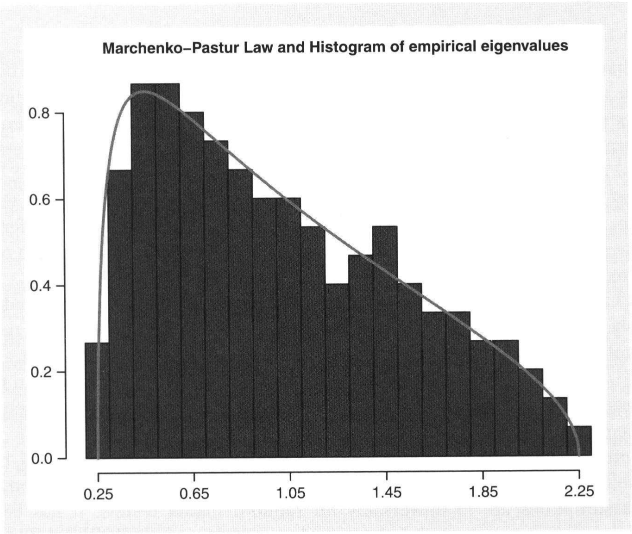

Fig. 28.3 Normalized histogram of eigenvalues and Marčenko–Pastur density (solid line), $n = 600$, $p = 150$, iid Gaussian data, $\mu = 0$, $\Sigma = \mathrm{Id}_p$. The population (or true) eigenvalues are all equal to 1. The 'overspreading' of sample eigenvalues is striking.

Figure: from N. El Karoui, in: Handbook of RMT

High dimensional asymptotics: $p/n \to y > 0$

If $\lambda_1, \ldots, \lambda_{p(n)} \in \mathbb{R}$ denote the eigenvalues of XX^T (with multiplicities), define their empirical measure as

$$\mathsf{L}_n(XX^T) := \frac{1}{p(n)} \sum_{j=1}^{p(n)} \delta_{\lambda_j}.$$

(Marčenko/Pastur): $\mathbb{E}(\mathsf{L}_n(XX^T))$ converges weakly to the Marčenko-Pastur distribution with parameter $y := \lim_{n\to\infty} \frac{p(n)}{n}$.

A physics motivation for symmetries in the data

Toy model for a Dirac operator:

$$\mathbb{M}_n^{\mathrm{CII}} = \left\{ \begin{pmatrix} 0 & X \\ X^* & 0 \end{pmatrix} : \ X \in \mathbb{H}^{s\times t} \right\},$$

where the space $\mathbb{H}^{s\times t}$ of quaternionic matrices is embedded into $\mathbb{C}^{2s\times 2t}$ as

$$\mathbb{H}^{s\times t} = \left\{ \begin{pmatrix} U & V \\ -\overline{V} & \overline{U} \end{pmatrix} : \ U, V \in \mathbb{C}^{s\times t} \right\}.$$

Observe:

$$\operatorname{Tr} \begin{pmatrix} 0 & X_n \\ X_n^* & 0 \end{pmatrix}^k = \begin{cases} 0 & \text{if } k \text{ odd} \\ 2 \operatorname{Tr}((X_n^* X_n)^l) & \text{if } k = 2l \text{ even.} \end{cases}$$

Viewing symmetry as an extreme form of dependence

A flexible framework for incorporating symmetries into the data matrices: Allowing **dependencies of arbitrary type**, but subject to quantitative restrictions.

Viewing symmetry as an extreme form of dependence

A flexible framework for incorporating symmetries into the data matrices: Allowing **dependencies of arbitrary type**, but subject to quantitative restrictions.

Our result: The Marčenko-Pastur theorem remains true in this framework.

Viewing symmetry as an extreme form of dependence

A flexible framework for incorporating symmetries into the data matrices: Allowing **dependencies of arbitrary type**, but subject to quantitative restrictions.

Our result: The Marčenko-Pastur theorem remains true in this framework.

Statistics interpretation: Robustness of the Marčenko-Pastur approximaton w.r.t. certain manipulations of the data.

Scenario I: Missing data

Suppose that for

$$I \subset \{1, \ldots, p\}, \ \#I \leq \log p, \ \ J \subset \{1, \ldots, n\}, \ \#J \leq \log n,$$

observations $X_i^{(j)} \quad (i \in I, j \in J)$ are missing.
Replace them with $X_{i_0}^{(j_0)} \quad (i_0 \notin I, j_0 \notin J \text{ arbitrary})$.

Scenario II: The lazy research assistant

$I \subset \{1, \ldots, p\}$, $\#I \leq \log p$. For $i \in I$, fill the ith row with copies of the shorter sequence $X_i^{(1)}, \ldots, X_i^{(\lfloor n/i \rfloor)}$.

Scenario II: The lazy research assistant

$I \subset \{1, \ldots, p\}$, $\#I \leq \log p$. For $i \in I$, fill the ith row with copies of the shorter sequence $X_i^{(1)}, \ldots, X_i^{(\lfloor n/i \rfloor)}$.

Or perhaps more fanciful schemes that also create dependencies between different rows...

References

- J. Schenker / H. Schulz-Baldes, Semicircle law and freeness for random matrices with symmetries or correlations, *Mathematical Research Letters* 12 (2005), 531–542.
- K. Hofmann-Credner / M. Stolz, Wigner theorems for random matrices with dependent entries: Ensembles associated to symmetric spaces and sample covariance matrices, *Electronic Communications in Probability* 13 (2008), 401–414.
- O. Friesen / M. Löwe / M. Stolz, Gaussian fluctuations for sample covariance matrices with dependent data, *Journal of Multivariate Analysis* 114 (2013), 270–287.