Random matrices and covariance estimation

Kaveh S. Nobari

Lectures in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

Motivation

The issue of covariance estimation is intertwined with random matrix theory, since sample covariance is a particular type of random matrix. These slides follow the structure of chapter 6 of Wainwright (2019) to shed light on random matrices in a non-asymptotic setting, with the aim of obtaining explicit deviation inequalities that hold for all sample sizes and matrix dimensions.

In the classical framework of covariance matrix estimation the sample size n tends to infinity while the matrix dimension d is fixed; in this setting the behaviour of sample covariance matrix is characterized by the usual limit theory. In contrast, in high-dimensional settings the data dimension is either comparable to the sample size ($d \asymp n$) or possibly much larger than the sample size $d \gg n$.

Motivation

We begin with the simplest case, namely ensembles of Gaussian random matrices, and we then discuss more general sub-Gaussian ensembles, before moving to milder tail conditions.

First, let us consider rectangular matrices, for instance matrix $A \in \mathbb{R}^{n \times m}$ with $n \geq m$, the ordered singular values are written as follows

$$\sigma_{\max}(A) = \sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_m(A) = \sigma_{\min}(A) \geq 0$$

The maximum and minimum singular values are obtained by maximizing the "blow-up factor"

$$\sigma_{\max}(A) = \max_{\forall x} \frac{\|Ax\|_2}{\|x\|_2}, \quad \sigma_{\min}(A) = \min_{\forall x} \frac{\|Ax\|_2}{\|x\|_2}$$

which is obtained when x is the largest and smallest singular vectors respectively - i.e.

$$\sigma_{\max}(A) = \max_{v \in S^{m-1}} \frac{\|Av\|_2}{\|v\|_2}, \quad \sigma_{\min}(A) = \min_{v \in S^{m-1}} \frac{\|Av\|_2}{\|v\|_2}$$

noting that $\|v\|_2 = 1$, since $S^{d-1} := \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$ is the Euclidean unit sphere in $\mathbb{R}^d$. We may denote

$$|||A|||_2 = \sigma_{\max}(A)$$

However, covariance matrices are square symmetric matrices, thus we must also focus on symmetric matrices in $\mathbb{R}^d$, denoted $S^{d\times d} := \{Q \in \mathbb{R}^{d\times d} \mid Q = Q'\}$, as well as subset of semi-definite matrices given by

$$S_+^{d\times d} := \{Q \in S^{d\times d} \mid Q \geq 0\}.$$

Any matrix $Q \in S^{d\times d}$ is diagonalizable via unitary transformation, and let us denote the vector of eigenvalues of Q by $\gamma(Q) \in \mathbb{R}^d$ ordered as

$$\gamma_{\max}(Q) = \gamma_1(Q) \geq \gamma_2(Q) \geq \cdots \geq \gamma_d(Q) = \gamma_{\min}(Q)$$

Note the matrix Q is semi-positive definite, which may be expressed as $Q \geq 0$, iff $\gamma_{\min}(Q) \geq 0$.

The Rayleigh-Ritz variational characterization of the minimum and maximum eigenvalues

$$\gamma_{\max}(Q) = \max_{v \in S^{d-1}} v'Qv \quad \text{and} \quad \gamma_{\min}(Q) = \min_{v \in S^{d-1}} v'Qv$$

For symmetric matrix Q, the l_2 norm can be expressed as

$$|||Q|||_2 = \max\{\gamma_{\max}(Q), |\gamma_{\min}(Q)|\} := \max_{v \in S^{d-1}} |v'Qv|$$

Finally, suppose we have a rectangular matrix $A \in \mathbb{R}^{n \times m}$, with $n \geq m$. We know that any rectangular matrix can be expressed using singular value decomposition (SVD hereafter), as follows

$$A = U \Sigma V'$$

wher U is an $n \times n$ unitary matrix, Σ is an $n \times m$ rectangular diagonal matrix with non-negative real numbers on the diagonal up and V is an $n \times n$ unitary matrix. Using SVD, we can express $A'A$ where

$$A'A = V\Sigma' U' U \Sigma V'$$

and since U is an orthogonal matrix, we know that $U'U = I$ where I is the identity matrix.

$$A'A = V(\Sigma'\Sigma)V'$$

Therefore, as the diagonal matrix Σ contains the eigenvalues of matrix A, hence, $\Sigma'\Sigma$ contains the eigenvalues of $A'A$ and it can be thus concluded

$$\gamma_j(A'A) = (\sigma_j(A))^2, \quad j = 1, \cdots, m$$

Let $\{x_1, \cdots, x_n\}$ be a collection of n i.i.d samples from a distribution in $\mathbb{R}^d$ with zero mean and the covariance matrix Σ. A standard estimator of sample covariance matrix is

$$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} x_i x_i'.$$

Since, each x_i for $i = 1, \cdots, n$ has zero mean, it is guaranteed that

$$\mathbb{E}[x_i x_i'] = \Sigma$$

and the random matrix $\hat{\Sigma}$ is an unbiased estimator of the population covariance Σ. Consequently the error matrix $\hat{\Sigma} - \Sigma$ has mean zero, and goal is to obtain bounds on the error measures in l_2-norm. We are essentially seeking a band of the form

$$\left|\left|\left| \hat{\Sigma} - \Sigma \right|\right|\right|_2 \leq \varepsilon,$$

where,

$$\begin{aligned}
\left|\left|\left|\hat{\Sigma} - \Sigma\right|\right|\right|_2 &= \max_{v \in S^{d-1}} \left| v' \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i x_i' - \Sigma \right\} v \right| \\
&= \max_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} v' x_i x_i' v - v' \Sigma v \right| \\
&= \max_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \langle x_i, v_i \rangle^2 - v' \Sigma v \right| \leq \varepsilon
\end{aligned}$$

which suggests that controlling the deviation $\left|\left|\left|\hat{\Sigma} - \Sigma\right|\right|\right|_2$ is equivalent to establishing a ULLN for the class of functions $x \to \langle x, v \rangle^2$, indexed by vectors $v \in S^{d-1}$.

Definition (Weyl's Inequality)

(I) Given any real symmetric matrices A, B,

$$\gamma_1(A+B) \geq \gamma_1(A) + \gamma_1(B)$$
$$\gamma_n(A+B) \leq \gamma_n(A) + \gamma_n(B)$$

(II) Given any real symmetric matrices A, B,

$$|\gamma_k(A) - \gamma_k(B)| \leq |||(A-B)|||_2$$

(see DasGupta (2008)).

,

Control in the operator norm further guarantees that the eigenvalues of $\hat{\Sigma}$ are uniformly close to those of Σ. Furthermore, given Weyl's inequality II above, we have

$$\max_{j=1,\cdots,d} |\gamma_j(\hat{\Sigma}) - \gamma_j(\Sigma)| \leq \left\lvert\!\left\lvert\!\left\lvert \hat{\Sigma} - \Sigma \right\rvert\!\right\rvert\!\right\rvert_2$$

Note that the random matrix $X \in \mathbb{R}^{n \times d}$ has the vectors x_i' on its i^{th} row and singular values denotes by $\{\sigma_j(X)\}_{j=1}^{\min n,d}$. Thus,

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i' = \frac{1}{n} X' X$$

and hence, the eigenvalues of $\hat{\Sigma}$ are the squares of the singular values of $X/\sqrt{n}$.

Definition (Gaussian ensembles and Wishart distribution)

Suppose that each sample x_i of a matrix $X \in \mathbb{R}^{n\times d}$ is drawn from an i.i.d multivariate $N(0, \Sigma)$ distribution. In this case we say that the associated matrix $X \in n \times d$, with x_i' and its i^{th} row, is drawn from the Σ-Gaussian ensemble. The associated sample covariance $\hat{\Sigma} = \frac{1}{n}X'X$ is said to follow a multivariate Wishart distribution.

Following Wainwright (2019), we present deviation inequalities for Σ-Gaussian ensembles and present a few examples before proving said inequalities.

Theorem

Let $X \in \mathbb{R}^{n\times d}$ be drawn according to the Σ-Gaussian ensemble. Then for $\delta > 0$, the maximum singular value $\sigma_{\max}(X)$ satisfies the upper deviation inequality

$$P\left[\frac{\sigma_{\max}(X)}{\sqrt{n}} \geq \gamma_{\max}(\sqrt{\Sigma})(1+\delta) + \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}\right] \leq \exp\left(-\frac{n\delta^2}{2}\right)$$

Furthermore, for $n \geq d$, the minimum singular value $\sigma_{\min}(X)$ satisfies the lower deviation inequality

$$P\left[\frac{\sigma_{\min}(X)}{\sqrt{n}} \leq \gamma_{\min}(\sqrt{\Sigma})(1-\delta) - \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}\right] \leq \exp\left(-\frac{n\delta^2}{2}\right)$$

Example (Norm bounds for standard Gaussian ensemble): Consider $W \in \mathbb{R}^{n\times d}$ generated with i.i.d $N(0,1)$ entries, which leads to the I_d-Gaussian ensemble. Given the above Theorem, it can be concluded that for $n \geq d$

$$\frac{\sigma_{\max}(W)}{\sqrt{n}} \leq 1+\delta+\sqrt{\frac{d}{n}} \quad \text{and} \quad \frac{\sigma_{\min}(W)}{\sqrt{n}} \geq 1-\delta-\sqrt{\frac{d}{n}}$$

Now it is evident that

$$1-P\left[\frac{\sigma_{\max}(W)}{\sqrt{n}} \geq 1+\delta+\sqrt{\frac{d}{n}}\right] = P\left[\frac{\sigma_{\max}(W)}{\sqrt{n}} \leq 1+\delta+\sqrt{\frac{d}{n}}\right]$$

thus according to the earlier Theorem,

$$P\left[\frac{\sigma_{\max}(W)}{\sqrt{n}} \leq 1+\delta+\sqrt{\frac{d}{n}}\right] \geq 1-\exp\left(-\frac{n\delta^2}{2}\right)$$

and similarly

$$P\left[\frac{\sigma_{\min}(W)}{\sqrt{n}} \geq 1-\delta-\sqrt{\frac{d}{n}}\right] \geq 1-\exp\left(-\frac{n\delta^2}{2}\right)$$

Thus, it can easily be seen that both bounds hold with probability greater than $1-2\exp\left(-\frac{n\delta^2}{2}\right)$. As we recall, the eigenvalues of the symmetric covariance matrix $\hat{\Sigma}$ is the square of the singular values $W/\sqrt{n}$. Furthermore,

$$\begin{aligned}
\left|\left|\left|\hat{\Sigma}-\Sigma\right|\right|\right|_2 &= \max_{v\in S^{d-1}}\left|v'\left\{\frac{1}{n}W'W-I_d\right\}v\right| \\
&= \max_{v\in S^{d-1}}\left|\frac{1}{n}v'(W'W)v-v'I_dv\right|
\end{aligned}$$

Note that $v' I_d v = \|v\|_2^2 = 1$. Thus,

$$\left\|\left\|\hat{\Sigma} - \Sigma\right\|\right\|_2 = \left\|\left\|\frac{1}{n} W' W - I_d\right\|\right\|_2$$
$$= \max_{v \in S^{d-1}} \left|\frac{1}{n} v'(W'W)v - 1\right|$$

Moreover, we have

$$\frac{\sigma_{\max}(W)}{\sqrt{n}} \leq 1 + \delta + \sqrt{\frac{d}{n}}$$

or

$$\frac{(\sigma_{\max}(W))^2}{n} \leq 1 + 2\left(\underbrace{\delta + \sqrt{\frac{d}{n}}}_{\varepsilon}\right) + \left(\underbrace{\delta + \sqrt{\frac{d}{n}}}_{\varepsilon}\right)^2$$

$$\left\{\frac{(\sigma_{\max}(W))^2}{n} - 1\right\} \leq 2\varepsilon + \varepsilon^2$$

thus,

$$\left|\!\left|\!\left| \frac{1}{n} W'W - I_d \right|\!\right|\!\right|_2 \leq 2\varepsilon + \varepsilon^2$$

Note that $\frac{d}{n} \to 0$, thus, the sample covariance matrix $\hat{\Sigma}$ is a consistent estimate of the identity matrix I_d.

Example (Gaussian covariance estimation):

Let $X \in \mathbb{R}^{n \times d}$ be a random matrix from the Σ-Gaussian ensemble. Noting that a if $X \sim N(0, \Sigma)$ it can equivalently be written as $X \sim \sqrt{\Sigma} N(0, I_d)$. So assuming that $W \sim N(0, I_d)$, we may express X as $X = W\sqrt{\Sigma}$. Moreover,

$$\begin{aligned} \left|\!\left|\!\left| \frac{1}{n} X'X - \Sigma \right|\!\right|\!\right|_2 &= \left|\!\left|\!\left| \sqrt{\Sigma} \left(\frac{1}{n} W'W - I_d \right) \sqrt{\Sigma} \right|\!\right|\!\right|_2 \\ &\leq |\!|\!| \Sigma |\!|\!|_2 \left|\!\left|\!\left| \frac{1}{n} W'W - I_d \right|\!\right|\!\right|_2 \end{aligned}$$

Thus, given the earlier example we know that

$$\left|\!\left|\!\left| \frac{1}{n} W'W - I_d \right|\!\right|\!\right|_2 \leq 2\varepsilon + \varepsilon^2,$$

where $\varepsilon = \delta + \sqrt{\frac{d}{n}}$. Therefore,

$$\frac{\left|\!\left|\!\left| \hat{\Sigma} - \Sigma \right|\!\right|\!\right|_2}{|\!|\!| \Sigma |\!|\!|_2} \leq 2\varepsilon + \varepsilon^2$$

Therefore, the relative error above converges to zero, so long as $d/n \to 0$.

References

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer Science & Business Media.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.