Classical vs High Dimensional Theory
Exercises

1 Covariance estimation in high-dimensions

Reference:
Wainwright (2019)
High-Dimensional Statistics: A Non-Asymptotic Viewpoint
Cambridge University Press
Chapter 1.2.2

Suppose $x_1, \dots, x_n$ are i.i.d. $N_p(0, \Sigma)$. A natural estimator for Σ is the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\mathsf{T}$$

(you can also consider the usual estimator `cov()`)

A classical analysis considers the behavior of the sample covariance matrix Σ as the sample size n increases while the dimension p stays fixed. The sample covariance $\hat{\Sigma}$ is a consistent estimate of Σ in the classical setting. Is this type of consistency preserved if we also allow the dimension p to tend to infinity?

Using n samples $x_1, \dots, x_n$ i.i.d. $N_p(0, I_p)$, obtain $\hat{\Sigma}$ and then compute its vector of eigenvalues $\lambda(\hat{\Sigma})$ arranged in non-increasing order:

$$\lambda_1(\hat{\Sigma}) \geq \lambda_2(\hat{\Sigma}) \geq \ldots \geq \lambda_p(\hat{\Sigma})$$

If the sample covariance matrix $\hat{\Sigma}$ were converging to the identity matrix $\Sigma = I_p$, then the vector of eigenvalues should converge to the all-ones vector:

$$\lambda_1(I_p) = \lambda_2(I_p) = \ldots = \lambda_p(I_p) = 1$$

Perform a simulation with $(p, n) = (10, 2000)$ and $(p, n) = (1000, 2000)$: repeat the estimation of $\lambda(\hat{\Sigma})$ many times, and plot the histogram of the estimated vector of eigenvalues. Comments on the results.