A Mathematical Introduction to Data Science **Feb 19, 2021**

Homework 2. Random Matrix Theory and PCA

Instructor: Yuan Yao *Due: 1 week*

The problem below marked by * is optional with bonus credits. For the experimental problem, include the source codes which are runnable under standard settings. Since there is NO grader assigned for this class, homework will not be graded. But if you would like to submit your exercise, please send your homework to the address (datascience.hw@gmail.com) with a title "CSIC5011: Homework #". I'll read them and give you bonus credits.

1. *Phase transition in PCA "spike" model:* Consider a finite sample of n i.i.d vectors $x_1, x_2, \ldots, x_n$ drawn from the p-dimensional Gaussian distribution $\mathcal{N}(0, \sigma^2 I_{p\times p} + \lambda_0 uu^T)$, where λ_0/σ^2 is the signal-to-noise ratio (SNR) and $u \in \mathbb{R}^p$. In class we showed that the largest eigenvalue λ of the sample covariance matrix S_n

$$S_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

pops outside the support of the Marcenko-Pastur distribution if

$$\frac{\lambda_0}{\sigma^2} > \sqrt{\gamma},$$

or equivalently, if

$$\text{SNR} > \sqrt{\frac{p}{n}}.$$

(Notice that $\sqrt{\gamma} < (1+\sqrt{\gamma})^2$, that is, λ_0 can be "buried" well inside the support Marcenko-Pastur distribution and still the largest eigenvalue pops outside its support). All the following questions refer to the limit $n \to \infty$ and to almost surely values:

   (a) Find λ given SNR $> \sqrt{\gamma}$.

   (b) Use your previous answer to explain how the SNR can be estimated from the eigenvalues of the sample covariance matrix.

   (c) Find the squared correlation between the eigenvector v of the sample covariance matrix (corresponding to the largest eigenvalue λ) and the "true" signal component u, as a function of the SNR, p and n. That is, find $|\langle u, v\rangle|^2$.

   (d) Confirm your result using MATLAB, Python, or R simulations (e.g. set $u = e$; and choose $\sigma = 1$ and λ_0 in different levels. Compute the largest eigenvalue and its associated eigenvector, with a comparison to the true ones.)

2. *Exploring S&P500 Stock Prices:* Take the Standard & Poor's 500 data: `https://github.com/yao-lab/yao-lab.github.io/blob/master/data/snp452-data.mat` which contains the data matrix $X \in \mathbb{R}^{p\times n}$ of $n = 1258$ consecutive observation days and $p = 452$ daily closing stock prices, and the cell variable "stock" collects the names, codes, and the affiliated industrial sectors of the 452 stocks. Use Matlab, Python, or R for the following exploration.

   (a) Take the logarithmic prices $Y = \log X$;

   (b) For each observation time $t \in \{1, \ldots, 1257\}$, calculate logarithmic price jumps

   $$\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}, \quad i \in \{1, \ldots, 452\};$$

   (c) Construct the realized covariance matrix $\hat{\Sigma} \in \mathbb{R}^{452\times 452}$ by,

   $$\hat{\Sigma}_{i,j} = \frac{1}{1257}\sum_{\tau=1}^{1257} \Delta Y_{i,\tau}\Delta Y_{j,\tau};$$

   (d) Compute the eigenvalues (and eigenvectors) of $\hat{\Sigma}$ and store them in a descending order by $\{\hat{\lambda}_k, k = 1, \ldots, p\}$.

   (e) *Horn's Parallel Analysis*: the following procedure describes a so-called Parallel Analysis of PCA using random permutations on data. Given the matrix $[\Delta Y_{i,t}]$, apply random permutations $\pi_i : \{1, \ldots, t\} \to \{1, \ldots, t\}$ on each of its rows: $\Delta\tilde{Y}_{i,\pi_i(j)}$ such that

   $$[\Delta\tilde{Y}_{\pi(i),t}] = \begin{bmatrix} \Delta Y_{1,1} & \Delta Y_{1,2} & \Delta Y_{1,3} & \ldots & \Delta Y_{1,t} \\ \Delta Y_{2,\pi_2(1)} & \Delta Y_{2,\pi_2(2)} & \Delta Y_{2,\pi_2(3)} & \ldots & \Delta Y_{2,\pi_2(t)} \\ \Delta Y_{3,\pi_3(1)} & \Delta Y_{3,\pi_3(2)} & \Delta Y_{3,\pi_3(3)} & \ldots & \Delta Y_{3,\pi_3(t)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \Delta Y_{n,\pi_n(1)} & \Delta Y_{n,\pi_n(2)} & \Delta Y_{n,\pi_n(3)} & \ldots & \Delta Y_{n,\pi_n(t)} \end{bmatrix}.$$

   Define $\tilde{\Sigma} = \frac{1}{t}\Delta\tilde{Y} \cdot \Delta\tilde{Y}^T$ as the null covariance matrix. Repeat this for R times and compute the eigenvalues of $\tilde{\Sigma}_r$ for each $1 \le r \le R$. Evaluate the p-value for each estimated eigenvalue $\hat{\lambda}_k$ by $(N_k{+}1)/(R{+}1)$ where N_k is the counts that $\hat{\lambda}_k$ is less than the k-th largest eigenvalue of $\tilde{\Sigma}_r$ over $1 \le r \le R$. Eigenvalues with small p-values indicate that they are less likely arising from the spectrum of a randomly permuted matrix and thus considered to be signal. Draw your own conclusion with your observations and analysis on this data. A reference is: Buja and Eyuboglu, "Remarks on Parallel Analysis", Multivariate Behavioral Research, 27(4): 509-540, 1992.

3. *Finite rank perturbations of random symmetric matrices:* Wigner's semi-circle law (proved by Eugene Wigner in 1951) concerns the limiting distribution of the eigenvalues of random symmetric matrices. It states, for example, that the limiting eigenvalue distribution of $n \times n$ symmetric matrices whose entries w_{ij} on and above the diagonal ($i \le j$) are i.i.d Gaussians $\mathcal{N}(0, \frac{1}{4n})$ (and the entries below the diagonal are determined by symmetrization, i.e., $w_{ji} = w_{ij}$) is the semi-circle:

   $$p(t) = \frac{2}{\pi}\sqrt{1-t^2}, \quad -1 \le t \le 1,$$

   where the distribution is supported in the interval $[-1, 1]$.

(a) Confirm Wigner's semi-circle law using MATLAB, Python, or R simulations (take, e.g., $n = 400$).

(b) Find the largest eigenvalue of a rank-1 perturbation of a Wigner matrix. That is, find the largest eigenvalue of the matrix

$$W + \lambda_0 u u^T,$$

where W is an $n \times n$ random symmetric matrix as above, and u is some deterministic unit-norm vector. Determine the value of λ_0 for which a phase transition occurs. What is the correlation between the top eigenvector of $W + \lambda_0 u u^T$ and the vector u as a function of λ_0? Use techniques similar to the ones we used in class for analyzing finite rank perturbations of sample covariance matrices.

[Some Hints about homework] For Wigner Matrix $W = [w_{ij}]_{n\times n}, w_{ij} = w_{ji}, w_{ij} \sim \mathcal{N}(0, \frac{\sigma}{\sqrt{n}})$, the answer is

$$\begin{aligned} \text{eigenvalue is} \quad & \lambda = R + \tfrac{1}{R} \\ \text{eigenvector satisfies} \quad & (u^T \hat{v})^2 = 1 - \tfrac{1}{R^2} \end{aligned}$$