MATH 5473 Homework 2 LUO Yuanhui

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.)

Solution: (a) Let the corresponding eigenvector be ν, then

$S_n \nu = \lambda \nu$

Let $\Sigma = \sigma^2 I_{p\times p} + \lambda_0 uu^T$, $y_i \triangleq \Sigma^{-\frac{1}{2}} X_i$, then $Y = \Sigma^{-\frac{1}{2}} X \sim N(0, I_p)$

$T_n = \frac{1}{n} YY^T$ is a Wishart Matrix, then the limit distribution of T_n's eigenvalues follows a MP distribution.

Notice that $T_n = \frac{1}{n} YY^T = \frac{1}{n}(\Sigma^{-\frac{1}{2}} X)(\Sigma^{-\frac{1}{2}} X)^T = \Sigma^{-\frac{1}{2}} S_n \Sigma^{-\frac{1}{2}}$, then

$S_n = \Sigma^{\frac{1}{2}} T_n \Sigma^{\frac{1}{2}}$

$S_n \nu = \Sigma^{\frac{1}{2}} T_n (\Sigma \Sigma^{-\frac{1}{2}}) \nu = \lambda \nu \Leftrightarrow T_n \Sigma (\Sigma^{-\frac{1}{2}} \nu) = \lambda (\Sigma^{-\frac{1}{2}} \nu)$

Then λ, $\Sigma^{-\frac{1}{2}} \nu$ is the eigenvalue and corresponding eigenvector

of $T_n \Sigma$

Let $\nu^* = c(\Sigma^{-\frac{1}{2}} \nu)$ s.t. $\nu^{*T} \nu^* = 1$, we have $c^2 = (R(u^T \sigma)^2 + 1) \sigma^2$

then ν^* is a normalized eigen vector of $T_n \Sigma$

$T_n \Sigma \nu^* = T_n (\sigma^2 I_p + \lambda_0 u u^T) \nu^* = \lambda \nu^* \Leftrightarrow \lambda_0 T_n u u^T \nu^* = (\lambda I_p - T_n \sigma^2 I_p) \nu^*$

$\Leftrightarrow u^T \nu^* = u^T (\lambda I_p - T_n \sigma^2 I_p)^{-1} \lambda_0 T_n u u^T \nu^*$ (*)

Suppose $u^T \nu^* \neq 0$. $T_n = W \Lambda W^T$, $W W^T = I_p$. $\Lambda = \mathrm{diag}\{\lambda_1 \dots \lambda_p\}$, then

$1 = \lambda_0 \sum_{i=1}^{p} u_i^2 \frac{\lambda_i}{\lambda - \sigma^2 \lambda_i} = \lambda_0 \int_a^b \frac{t}{\lambda - \sigma^2 t} d\mu_{MP}$, by Stieltjes transform,

$1 = \frac{\lambda_0}{4r} [2\lambda - (a+b) - 2\sqrt{(\lambda - a)(b - \lambda)}]$ for $\lambda > (1+\sqrt{r})^2 \triangleq b$ and $SNR > \sqrt{r}$

Then given $SNR > \sqrt{r}$, $\lambda = \lambda_0 + \frac{r}{\lambda_0} + 1 + r = (1+\lambda_0)(1+\frac{r}{\lambda_0})$

Therefore, $\lambda_{max}(S_n) = \begin{cases} (1+\sqrt{r})^2 = b, & \sigma_x^2 \leq \sqrt{r} \\ (1+\sigma_x^2)(1+\frac{r}{\sigma_x^2}), & \sigma_x^2 > \sqrt{r} \end{cases}$

(b) For $S_n = \frac{1}{n} X X^T$, $b = (1+\sqrt{r})^2$, we can estimate SNR by $\begin{cases} SNR \le \sqrt{r} & \text{if } \lambda_{max}(S_n) = b \\ SNR > \sqrt{r} & \text{if } \lambda_{max}(S_n) = (1+\sigma_x^2)(1+\frac{r}{\sigma_x^2}) \end{cases}$. Here SNR

$= \sigma_x^2$ by assuming $\sigma_\varepsilon = 1$ WLOG.

(c) By $(*)$ we have $u^T v^* = u^T(\lambda I_p - T_n \sigma^2 I_p)^{-1} \lambda_0 T_n u u^T v^*$,

then $(u^T v^*)^T (u^T v^*) = \lambda_0^2 (u^T v^*)^T u^T T_n (\lambda I_p - T_n \sigma^2 I_p)^{-2} T_n$

$u (u^T v^*)$.

Then $|u^T v^*|^2 \sim \lambda_0^2 \int_a^b \frac{t^2}{(\lambda - \sigma^2 t^2)^2} d\mu_{mp}(t) = \frac{\lambda_0^2}{4r}(-4r + (a+b) +$

$2\sqrt{(\lambda-a)(\lambda-b)} + \frac{\lambda(2\lambda - (a+b))}{\sqrt{(\lambda-a)(\lambda-b)}}) = \frac{1-\frac{r}{R}}{1+\sigma+\frac{2r}{R}}$

Then $|u^T v|^2 = (\frac{1}{c} u^T \Sigma^{\frac{1}{2}} v^*)^2 = \frac{1}{c^2}((\sqrt{1+R}\,u)^T v^*)^2 =$

$$\frac{(1+R)(u^T v^*)^2}{R(u^T v)^2 + 1} = \frac{1+R-\frac{r}{R}-\frac{r}{R^2}}{1+R+r+\frac{r}{R}} = \frac{1-\frac{r}{R^2}}{1+\frac{r}{R}}$$

(d) The code can be seen in the Ex1.ipynb.