HW2 Liu Yonglin

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.)

(a) Suppose $t = \alpha u$. $\alpha \sim N(0, \lambda_0)$

u is a direction s.t. $u^T u = 1$

$x = t + \varepsilon$ $\quad \varepsilon \sim N(0, \sigma^2 I_p)$

Then $x \sim N(0, \sigma^2 I_p + \lambda_0 uu^T)$

Denote $\Sigma = \sigma^2 I_p + \lambda_0 uu^T \in \mathbb{R}^{p\times p}$

$x_i \sim N(0, \Sigma) \in \mathbb{R}^p$

$X = [x_1 | x_2 | \cdots | x_n] \in \mathbb{R}^{p\times n}$

Assign $\dfrac{\text{signal of data}}{\text{signal of noise}} = \dfrac{\lambda_0}{\sigma^2} = \text{SNR}$

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

Then, the eigenvalue λ and corresponding eigenvector v satisfies

$$S_n v = \lambda v$$

In order to use NP distribution

$$y_i \triangleq \Sigma^{-\frac{1}{2}} x_i$$

then $$Y = [y_1 | y_2 | \cdots | y_n] = \Sigma^{-\frac{1}{2}} X \sim N(0, I_p)$$

$$T_n = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T = \frac{1}{n} Y Y^T \text{ is a Wishart Matrix}$$

So the limit distribution of T_n's eigenvalues follow an NP distribution

Connect T_n and S_n:

$$T_n = \frac{1}{n} Y Y^T$$

$$= \frac{1}{n} (\Sigma^{-\frac{1}{2}} X)(\Sigma^{-\frac{1}{2}} X)^T$$

(symmetric) $$= \Sigma^{-\frac{1}{2}} S_n \Sigma^{-\frac{1}{2}}$$

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

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

$$\Sigma^{\frac{1}{2}} T_n (\Sigma \Sigma^{-\frac{1}{2}}) v = \lambda v$$

$$T_n \Sigma (\Sigma^{-\frac{1}{2}} v) = \Sigma^{-\frac{1}{2}} \lambda v = \lambda (\Sigma^{-\frac{1}{2}} v)$$

So λ, $\Sigma^{-\frac{1}{2}} v$ is the eigenvalue & corresponding eigenvector of $T_n \Sigma$

Suppose $v^* = c(\Sigma^{-\frac{1}{2}} v)$ s.t. $v^{*T} v^* = 1$

$$c^2 (\Sigma^{-\frac{1}{2}} v)^T (\Sigma^{-\frac{1}{2}} v) = 1$$

$$c^2 v^T \Sigma^{-1} v = 1$$

$$c^2 v^T \Sigma^{-1} = v^T$$

$$c^2 v^T = v^T \Sigma$$

$$c^2 v = \Sigma v$$

$$c^2 = v^T \Sigma v$$

$$c^2 = \lambda_0 (u^T v)^2 + \sigma^2$$

$$= \sigma^2 \left(\frac{\lambda_0}{\sigma^2} (u^T v)^2 + 1 \right)$$

$$= (R (u^T \sigma)^2 + 1) \sigma^2$$

We use $v^* = c(\Sigma^{-\frac{1}{2}} v)$ a normalized eigenvector of (ΣT_n)

$$T_n \Sigma v^* = \lambda v^*$$

$$T_n(\sigma^2 I_p + \lambda u u^T) v^* = \lambda v^*$$

$$T_n \sigma^2 I_p v^* + \lambda_0 T_n u u^T v^* = \lambda v^*$$

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

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

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

Suppose $u^T v^* \neq 0$

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

Suppose $T_n = W \Lambda W^T \quad W W^T = I_p \quad \Lambda = \mathrm{diag}\{\lambda_1 \cdots \lambda_p\}$

$$1 = \lambda_0 \sum_{i=1}^{p} u_i^2 \frac{\lambda_i}{\lambda - \sigma^2 \lambda_i} \quad (*)$$

For $\sum u_i^2 = 1$. We regard $(a_i)^2$ as a probability measure.

When $p, n \to \infty \quad \lim_{p,n \to \infty} \frac{p}{n} = \gamma$

We have $\lambda_i \sim MP$ distribution

For $(*)$

$$1 = \lambda_0 \sum_{i=1}^{p} u_i^2 \frac{\lambda_i}{\lambda - \sigma^2 \lambda_i} = \lambda_0 \int_a^b \frac{t}{\lambda - \sigma^2 t} \, d\mu^{MP}(t)$$

According to Stieltjes transform

$$1 = \frac{\lambda_0}{4\gamma} \left[2\lambda - (a+b) - 2\sqrt{|\lambda - a||b - \lambda|} \right]$$

for $\lambda > (1+\sqrt{\gamma})^2 \triangleq b$ and $SNR > \sqrt{\gamma}$

Suppose $\sigma_\varepsilon^2 = 1$

for $SNR > \sqrt{\gamma}$ and $\sigma_x^2 > \sqrt{\gamma}$

$$\lambda = \lambda_0 + \frac{\gamma}{\lambda_0} + 1 + \gamma = (1+\lambda_0)(1+\frac{\gamma}{\lambda_0})$$

So given $SNR > \sqrt{\gamma}$ $\quad \lambda = \lambda_0 + \frac{\gamma}{\lambda_0} + 1 + \gamma = (1+\lambda_0)(1+\frac{\gamma}{\lambda_0})$

Actually $\quad \lambda_{max}(S_n) = \begin{cases} (1+\sqrt{\gamma})^2 = \sigma & \sigma_x^2 \leq \sqrt{\gamma} \\ (1+\sigma_x^2)(1+\frac{\gamma}{\sigma_x^2}) & \sigma_x^2 > \sqrt{\gamma} \end{cases}$

(b) We can estimate $SNR = \frac{\sigma_X^2}{\sigma_\varepsilon^2}$ WLOG $\sigma_\varepsilon^2 = 1$ by comparing

$\lambda_{max}(S_n)$ $(S_n = \frac{1}{n}XX^T)$ and $b \overset{\Delta}{=} (1+\sqrt{\gamma})^2$

If $\lambda_{max}(S_n) = b$, $SNR \le \sqrt{\gamma}$

If $\lambda_{max}(S_n) = (1+b_X^2)(1+\frac{\gamma}{\sigma_X^2})$, $SNR > \sqrt{\gamma}$

(c) According to $(**)$

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

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

$$1 = v^{*T}v^* = v^{*T} u u^T v^* = (u^T v^*)^T u^T v^*$$

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

Thus

$$|u^T v^*|^{-2} = \lambda_0 \left[u^T T_n(\lambda I_p - \sigma^2 T_n)^{-2} T_n u\right.$$

$$\sim \lambda_0^2 \int_a^b \frac{t^2}{(\lambda-\sigma^2 t)^2} d\mu^{MP}(t)$$

$$= \frac{\lambda_0}{4\gamma}\left(-4\gamma + (a+b) + 2\sqrt{(\lambda-a)(\lambda-b)} + \frac{\lambda(2\lambda-(a+b))}{\sqrt{(\lambda-a)(\lambda-b)}}\right)$$

Since $R = SNR = \frac{\sigma_X^2}{\sigma_\varepsilon^2} = \frac{\lambda_0}{\sigma^2} > b = (1+\sqrt{\gamma})^2$

We proved that $\hat{\lambda} = \lambda_{max} \to (1+R)(1+\frac{\gamma}{R})$

Thus, $|u^T v^*|^2 = \dfrac{1-\frac{\gamma}{R}}{1+\gamma+\frac{2\gamma}{R}}$

$$|u^T v|^2 = \left(\frac{1}{c}u^T \Sigma^{\frac{1}{2}} v^*\right)^2$$

$$= \frac{1}{c}\left(\left((Ruu^T + I_p)^{\frac{1}{2}}u\right)^T v^*\right)^2$$

$$= \frac{1}{c^2}\left(\left(\sqrt{(1+R)}u\right)^T v^*\right)^2$$

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

$$= \frac{1+R-\frac{\gamma}{R}-\frac{\gamma}{R^2}}{1+R+\gamma+\frac{\gamma}{R}}$$

$$= \frac{1 - \frac{\gamma}{R^2}}{1 + \frac{\gamma}{R}}$$

where $\gamma = \lim_{p,n \to \infty} \frac{p}{n}$

$R = SNR = \frac{\sigma_x^2}{\sigma_\varepsilon^2} = \frac{\lambda_0}{\sigma^2}$