MATH5473/CSIC5011 - Topological and Geometric Data Reduction and Visualization (Homework #2)

Li Yakun 209200531

- 1. Phase transition:

  (a) Find λ given SNR $> T$.

  Suppose $t = \alpha u, \quad \alpha \sim N(0, \lambda_0)$. And u is a direction s.t. $u^\top u = 1$.

  $\epsilon \sim N(0, \varsigma^2 I_p)$, $x = t + \varepsilon$ then $x \sim N(0, \sigma^2 I_p + \lambda u u^\top)$, where $\Sigma = \sigma^2 I_p + \lambda u u^\top$ is $p \times p$.

  $x_i \sim N(0, \Sigma) \in \mathbb{R}^p, \quad x = [x_1 | x_2 | \cdots x_n] \in \mathbb{R}^{p \times n}$.

  Assign $\frac{\text{signal of doth}}{\text{signal of noise}} = \frac{\lambda_0}{6^2} = SNR$, $S_n \triangleq \frac{1}{n} \sum_{i=1}^n x_i x_i^\top = \frac{1}{n} x x^\top$.

  Then. the eigenvalue λ and corresponding eigenvector v satisfies $S_n v = \lambda v$.

  Let $y_i = \Sigma^{-\frac{1}{2}} x_i$, $Y = [y_1 | y_2 | \ldots y_n] = \Sigma^{-\frac{1}{2}} X \sim N(0, I_p)$.

  $T_n = \frac{1}{n} \cdot \sum_{i=1}^n y_i y_i{}^\top = \frac{1}{n} \cdot Y Y^\top$ is a Wishart Matrix.

  So the limit distribution of Tn's eigenvalues follow a M_p distribution.

  $T_n = \frac{1}{n} Y Y^\top = \frac{1}{n} \left(\Sigma^{-\frac{1}{2}} X\right) \left(\Sigma^{-\frac{1}{2}} X\right)^\top = \Sigma^{-\frac{1}{2}} S_n \Sigma^{-\frac{1}{2}}$.

  $S_n = \sum^{\frac{1}{2}} T_n \sum^{\frac{1}{2}}$.

  $S_n v = \Sigma^{\frac{1}{2}} T_n \Sigma^{\frac{1}{2}} v = \lambda v$, $\Sigma^{\frac{1}{2}} T_n \left(\Sigma \Sigma^{-\frac{1}{2}}\right) v = \lambda v$, $T_n \Sigma \left(\Sigma^{-\frac{1}{2}} v\right) = \Sigma^{-\frac{1}{2}} \lambda v = \lambda \left(\Sigma^{-\frac{1}{2}} v\right)$.

  So, λ and $\left(\Sigma^{-\frac{1}{2}} v\right)$ is the eigenvalue and comesponding eigenvector of (T_n, Σ).

  Given $SNR > \sqrt{\gamma}$, $\lambda = (1 + \lambda_0)(1 + \frac{\gamma}{\lambda_0})$.

  $$\text{Actually } \lambda_{\max}(S_n) = \begin{cases} (1+\gamma)^2 = \sigma & \sigma x^2 \leqslant \sqrt{r} \\ \left(1 + \sigma x^2\right)\left(1 + \frac{\gamma}{\sigma x^2}\right) & \sigma x^2 > \sqrt{r} \end{cases}$$

  (b) Based on (a), we have the following results.

  If $\lambda_{\max}(S_n) = \sigma$ then we know $SNR \geq \sqrt{\gamma}$.

  If $\lambda_{\max}(S_n) = \left(1 + \sigma x^2\right)\left(1 + \frac{\gamma}{\sigma x^2}\right)$ we know $SNR > \sqrt{\gamma}$.

  (c)

  $\left|u^\top v\right|^2 = \left(\frac{1}{c} u^\top \Sigma^{\frac{1}{2}} v^*\right)^2 = \frac{(1+R)\left(u^\top v^*\right)^2}{R\left(u^\top v\right)^2 + 1} = \frac{1 + R - \frac{r}{R} - \frac{r}{R^2}}{1 + R + r + \frac{r}{R}} = \frac{1 - \frac{r}{R^2}}{1 + \frac{r}{R}}$,

  here $r = \lim_{p \cdot n \to \infty} \frac{p}{n}$, and $R = SNR = \frac{\sigma x^2}{\sigma \varepsilon^2} = \frac{\lambda_0}{\sigma^2}$.

  (d) By the code in $HW2\_T1.py$ attached in the email, all basic conclusions can be verified by the simulation experiments.

- 2. Exploring $S\&P500$ Stock Prices:

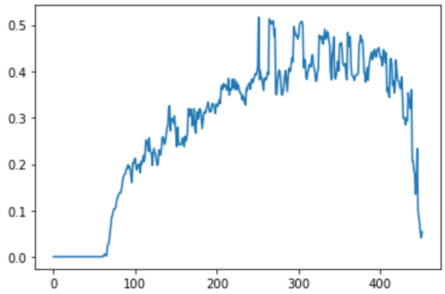

Figure 0.1: P-value for eigenvalues.

We have calculated the p-value for all eigenvalues of S and visualized them. The first eigenvalue that has bigger competitor from Sr's is the 63-th and its value is 3.0 with p-value 0.003996003996003996. Thus, we have evidence to believe PCA can be conducted to this dataset efficiently and effectively.