ELEC 5450: Random Matrix Theory and Applications

Homework Assignment 2

Released on 25th, March, 2019

Department of Electronic and Computer Engineering

The Hong Kong University of Science and Technology

Due Date: 17th, April, 2019

Grade: 20%

Submission Instructions: You should submit a hard copy of your solution to the TA during the lecture (you should also make a copy of your solution before submission), and submit to the TA through email the MATLAB code you would have generated to complete the assignment.

Important Note:

1) In order to get full marks, your solution should be very *clearly explained, including all relevant steps*.
2) **Students should conduct their own work, including their own numerical experiments. Any copying of work, including code among students is considered plagiarism and will be dealt with seriously**.

Portfolio Design: a Real Data Example

In this assignment, we address the problem of portfolio design using a real data set [1].

Given a random vector $\mathbf{x} \in \mathbb{R}^{m\times 1}$ (where m is the number of stocks), with $\mathrm{cov}\left(\mathbf{x}\right) = \mathbf{\Sigma}$, a portfolio $\mathbf{p}$ across these m stocks would take the form $\mathbf{p} = \left(p_1, ..., p_m\right)^T$ with $\sum_{i=1}^m p_i = 1$. The total daily return is then given by

$$\mathrm{tr}\mathbf{p} \triangleq \mathbf{p}^T\mathbf{x} = \sum_{i=1}^{m} p_i \left[\mathbf{x}\right]_i,$$

where $\left[\mathbf{x}\right]_i$ is the log-return (price change) of the $i-\mathrm{th}$ stock.

The variance of the daily return

$$R^2\left(\mathbf{p}\right) = \mathrm{var}\left(\mathrm{tr}\mathbf{p}\right) = \mathbf{p}^T\mathbf{\Sigma}\mathbf{p}$$

then represents the squared investment risk associated with this portfolio.

Assuming the true covariance matrix $\mathbf{\Sigma}$ is known, under a Markowitz optimization framework, the portfolio $\mathbf{p}$ that minimizes the risk $R\left(\mathbf{p}\right)$ takes the form

$$\mathbf{p}_{\mathrm{opt}|\mathbf{\Sigma}} \triangleq \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}},$$

where $\mathbf{1} = \left(1, ..., 1\right)^T$. In practice, $\mathbf{\Sigma}$ is unknown, and we have to estimate it based on the data at hand.

In the following, we will consider different estimators $\hat{\mathbf{\Sigma}}$ of the true covariance matrix, which will give us a corresponding portfolio

$$\mathbf{p}_{\hat{\mathbf{\Sigma}}} = \frac{\hat{\mathbf{\Sigma}}^{-1}\mathbf{1}}{\mathbf{1}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{1}}. \tag{1}$$

We can then compare the portfolios obtained with different estimators based on their associated risks.

I. Data Preprocessing and Analysis

The data - available online - constitute the daily returns of $m = 98$ stocks of the S&P 100 Index (https://en.wikipedia.org/wiki/S%26P_100) over a period of $n = 753$ days. We denote these data as $\mathbf{X}_s = \left[\mathbf{x}_1, ..., \mathbf{x}_n\right] \in \mathbb{R}^{m\times n}$.

Using MATLAB,

1) Compute the sample mean for each stock. Remove the sample mean from each data sample $\mathbf{x}_1, ..., \mathbf{x}_n$. In the following, we will assume the data $\mathbf{X}_s$ has been centered. [2 pts]

[1] Available on course website

2) Compute the sample covariance matrix $\mathbf{\Sigma}_s = \frac{1}{n}\mathbf{X}_s\mathbf{X}_s^T$. Plot the scree-plot[2] of the eigenvalues of the sample covariance matrix $\mathbf{\Sigma}_s$. Which eigenvalues explain most of the variability of the data? [1 pts] How many eigenvalues are sufficient to retain 50%, 60% and 70% of the total variance? [3 pts]

3) We now analyze the sample correlation matrix $\mathbf{C}_S$ of the data. The sample correlation matrix can be computed from $\mathbf{\Sigma}_s$ as

$$\mathbf{C}_S = \mathbf{D}^{-\frac{1}{2}}\mathbf{\Sigma}_s\mathbf{D}^{-\frac{1}{2}}, \tag{2}$$

where $\mathbf{D}$ is the diagonal matrix of $\mathbf{\Sigma}_s$ with $\mathbf{D}(i,i) = \mathbf{\Sigma}_s(i,i)$ for $i = 1, 2, ..., m$.

a) Perform the eigen-decomposition of the matrix $\mathbf{C}_S = \mathbf{VDV}^T = \sum_{i=1}^{m} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_m$. Plot the histograms of the eigenvalues of $\mathbf{C}_S$. [3 pts]

b) Consider a random correlation matrix $\mathbf{W} = \frac{1}{n}\mathbf{HH}^T$, where $\mathbf{H}$ is an $m \times n$ matrix whose entries are independent, identically distributed Gaussian random variables with zero-mean and unit-variance. Recall the Marcenko-Pastur law that in the limit $m \to \infty$, $n \to \infty$, such that $Q = n/m$ (>1) is fixed, the probability density function $\mathrm{P}_{\mathrm{rcm}}(d)$ of an arbitrary eigenvalue d of $\mathbf{W}$ is given by

$$P_{\mathrm{rcm}}(d) = \frac{Q}{2\pi}\frac{\sqrt{(d_+ - d)(d - d_-)}}{d} \tag{3}$$

for d within the bounds $d_- < d < d_+$, where d_- and d_+ are the minimum and maximum eigenvalues of $\mathbf{W}$, respectively, given by

$$d_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}.$$

Superimposed on the same graph in Section I Question 3) a), plot the curve of $P_{\mathrm{rcm}}(d)$ for $Q = n/m = 753/98$ for our case. [2 pts]

c) Repeat Question 3) a) but remove the largest 4 eigenvalues and only consider the set of eigenvalues $(\lambda_5, \cdots, \lambda_m)$. (You need to normalize the new set of eigenvalues so that the trace is conserved.) [2 pts] Superimpose on the same graph, plot the curve of $P_{\mathrm{rcm}}(d)$ for $Q = n/m = 753/98$ for our case. [1 pts] What do you observe? [1 pts]

d) Plot the histogram of the eigenvalues of $\mathbf{W}$ for $m = 98$ and $n = 753$. [2 pts] Superimposed on the same graph, plot the curve of $P_{\mathrm{rcm}}(d)$ for $Q = n/m = 753/98$. [1 pts] Compare this plot with the one you obtained in Question 3) a) - c) and give an interpretation of your observation. [3 pts]

e) We consider the correlation matrix of the shuffled data in time series. Let $\tilde{\mathbf{X}}_s$ be the time series randomized data of $\mathbf{X}_s$, with each row of $\tilde{\mathbf{X}}_s$ obtained by independently shuffling the

[2]Scree-plot is the plot of the eigenvalues sorted in descending order against the indices of the sorted eigenvalues.

corresponding row of $\mathbf{X}_s$. Compute the sample correlation matrix $\mathbf{C}_{\text{shuffle}}$ from the shuffled data $\tilde{\mathbf{X}}_s$. Plot the histogram of the eigenvalues of $\mathbf{C}_{\text{shuffle}}$. [2 pts] Superimposed on the same graph, plot the curve of $P_{\text{rcm}}(d)$ for $Q = n/m = 753/98$. [1 pts] Compare this plot with the one you obtained in Question 3) a) - d), and give an interpretation of your observation. [2 pts]

4) We now look at the eigenvectors of the eigenvalues of the correlation matrix that lie inside and deviate from the bulk. First recall that the probability density function of a Gaussian random variable with zero mean and unit variance is given by

$$\rho(v) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-v^2}{2}\right) \tag{4}$$

   a) Choose an arbitrary eigenvector $\mathbf{u}_k$ of correlation matrix $\mathbf{W}$, plot the histogram of the elements of the corresponding scaled[3] eigenvector $\sqrt{m}\mathbf{u}_k$. [2 pts] Superimposed on the same graph, plot the probability density function $\rho(v)$ as defined in (4). [1 pts]
   b) Choose a typical eigenvalue λ_k of correlation matrix $\mathbf{C}_S$ in the bulk, plot the histogram of the elements of the scaled eigenvector $\sqrt{m}\mathbf{v}_k$ associated with λ_k. [2 pts] Superimposed on the same graph, plot the probability density function $\rho(v)$ as defined in (4). [1 pts]
   c) Consider the "deviating" eigenvalues λ_1, λ_2 and λ_3 of $\mathbf{C}_S$, plot the histogram of the elements of the scaled eigenvectors $\sqrt{m}\mathbf{v}_3$, $\sqrt{m}\mathbf{v}_3$ and $\sqrt{m}\mathbf{v}_3$ associated with λ_1, λ_2 and λ_3. [3 pts] Superimposed on the same graphs, plot the probability density function $\rho(v)$ as defined in (4). [1 pts]

      Note: You should remove sign ambiguity to choose $\mathbf{v}_i$ such that first non-zero entry is positive.
   d) Looking at the plots generated in Question 4) a) - c), comment on how these can be used to identify eigenmodes representing informative correlations. [3 pts]

5) We define the Inverse Participation Ratio (IPR.) of a vector $\mathbf{u} = (u_1, ..., u_m)$ for $i = 1, ..., m$ as

$$I(\mathbf{u}) = \sum_{i=1}^{m} u_k^4.$$

   a) What is the IPR. of a vector $\mathbf{p} = (1/\sqrt{m}, ..., 1/\sqrt{m})^T$? [1 pts] What is the IPR. of a vector $\mathbf{q} = (1, 0, 0, ..., 0)^T$? [1 pts] What do you think I.P.R represents? [1 pts]
   b) For each eigenvector $\mathbf{u}_i$ of the random correlation matrix $\mathbf{W}$, compute the IPR. of $I(\mathbf{u}_i)$. Plot the log value of the IPR. of the vectors $\mathbf{u}_i$ versus the log value of their associated eigenvalues d_i. [3 pts]

      Hint: the command loglog in MATLAB is useful for plotting in logarithm scale.

[3] In MATLAB, the command eig computes the eigenvector with unit norm, i.e. $\|\mathbf{u}_k\| = 1$. Since the variance of the elements of the eigenvector $\mathbf{u}_k$ will vary with m, we typically need to rescale the elements of the eigenvector by a constant $\sqrt{m}$, so that $E\left\{\|\sqrt{m}\mathbf{u}_k\|^2\right\} = m$

c) Repeat question 5) b) with the correlation matrix $\mathbf{C}_S$ instead of $\mathbf{W}$. Superimpose your figure on the one you obtained in Question 5) b). [3 pts] Repeat the same procedure but only consider the set of eigenvalues $(\lambda_2, \cdots, \lambda_m)$ and the eigenvectors $(\mathbf{v}_2, ..., \mathbf{v}_m)$ of $\mathbf{C}_S$. Superimpose the figure you obtained on the one you obtained in Question 5) b). (You need to normalize the new set of eigenvalues so that the trace is conserved.) [3 pts] What can be inferred from your plots regarding the information conveyed by the eigenvectors of $\mathbf{C}_S$? [2 pts]

6) In Question 3) e), we considered shuffling the data in time series. What if we shuffle the data in the other dimension? To investigate this case, let $\hat{\mathbf{X}}_s$ be the stock domain randomized data of $\mathbf{X}_s$, with each column of $\hat{\mathbf{X}}_s$ obtained by independently shuffling the corresponding column of $\mathbf{X}_s$. Compute the sample correlation matrix $\hat{\mathbf{C}}_{\text{shuffle}}$ from the shuffled data $\hat{\mathbf{X}}_s$. Plot the histogram of the eigenvalues of $\hat{\mathbf{C}}_{\text{shuffle}}$. [3 pts] Superimposed on the same graph, plot the curve of $P_{\text{rcm}}(d)$ for $Q = n/m = 753/98$. [1 pts] What is your observation? [1 pts] Give a reasonable explanation to what you observed. [2 pts]

II. Portfolio Design

1) Let $\mathbf{\Sigma}_w^d$ be the covariance matrix computed from $\left[\mathbf{x}_{d-(w-1)}, ..., \mathbf{x}_d\right]$ where w denotes the window size. Compute the sample covariance matrix $\mathbf{\Sigma}_{w=200}^{200}$ based on the first 200 days. Based on this estimator, compute the associated portfolio weight $\mathbf{p}_{\mathbf{\Sigma}_{w=200}^{200}}$. What is the value of the return $\mathrm{tr}\mathbf{p}_{\mathbf{\Sigma}_{w=200}^{200}} = \mathbf{p}_{\mathbf{\Sigma}_{w=200}^{200}}^T \mathbf{x}_{201}$ you would get if you were to apply this portfolio weight on the 201-st day? [3 pts]
2) Repeat the same procedure for the whole period of time with $w = 200$, i.e., use $\left[\mathbf{x}_{i-(w-1)}, ..., \mathbf{x}_i\right]$ as training data to compute $\mathbf{p}_{\mathbf{\Sigma}_{w=200}^i}$, and compute the return on the $(1+i)\,\mathrm{th}$ day for $i = 200, ..., 752$. What is the total return $\sum_{i=200}^{752} \mathrm{tr}\mathbf{p}_{\mathbf{\Sigma}_{w=200}^i}$? [2 pts] What is the variance of the daily return? [2 pts]
3) Repeat Section II Question 2) by using $w = 100, 120, 150, 180, 200, 300, 400$ and 600 in the portfolio period from the $(w+1)$-th day to 753th day. Compute the average total returns $\frac{\sum_{i=w}^{752} \mathrm{tr}\mathbf{p}_{\mathbf{\Sigma}_w^i}}{753-w}$ and variance of the daily return respectively. Plot the variance of the daily return in logarithm scale versus the window size w. [2 pts] Plot also the average of the total return versus the window size w. [2 pts] How does the risk and average return change with the window size? [1 pts] Explain what your observed from your plots. [2 pts]
4) In the following, we only consider window size $w = 200$ and simplify our notation $\mathbf{\Sigma}_{w=200}^d$ by $\mathbf{\Sigma}_S^d$. We will now consider a more "robust" covariance estimator by "clipping" the covariance matrix which keeps the top K eigenvalues and shrinks the others such that the trace is preserved. In order to determine the value of K, we look instead at the eigen-decomposition of the sample correlation matrix, $\mathbf{C}_S^d$, which is computed from sample covariance matrix $\mathbf{\Sigma}_S^d$ by (2). The clipped correlation

matrix, $\mathbf{C}^d_{clip}$, is estimated from the top K eigenvalues while the remaining eigenvalues are shrunk to a constant $\bar{\eta}$ such that the trace is preserved, i.e., $\mathrm{Tr}\left(\mathbf{C}^d_{clip}\right) = \mathrm{Tr}\left(\mathbf{C}^d_S\right)$. Specifically,

$$\mathbf{C}^d_{clip}(K) \triangleq \sum_{i=1}^{K} \hat{\eta}_i \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T + \bar{\eta} \sum_{i=K+1}^{m} \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T$$

where $\hat{\eta}_i$ and $\hat{\mathbf{v}}_i$ are the eigenvalues and corresponding eigenvectors of $\mathbf{C}^d_S$. Then, the "clipped" covariance matrix $\boldsymbol{\Sigma}^d_{clip}(K)$ is obtained by transforming the $\mathbf{C}^d_{clip}(K)$ back to covariance matrix using equation (2).

a) Based on Section I Question 3), 4) and 5), explain in which sense $\boldsymbol{\Sigma}^d_{clip}(K)$ is more robust than the true covariance matrix $\boldsymbol{\Sigma}^d_S$. [2 pts]

b) Repeat Section II Question 2) by replacing $\boldsymbol{\Sigma}^d_S$ by $\boldsymbol{\Sigma}^d_{clip}(K)$ for $K = 1, 2, ..., 40, 60, 98$. Plot the variance of daily return versus K for $K = 1, 2, ..., 40, 60, 98$. [3 pts] What is the reasonable range of K you would choose to obtain a low variance of the daily return? [1 pts] For $K > 30$, how does the variance of the daily return changes as K increases? [1 pts] Explain your observation. [2 pts]

c) Construct another clipped correlation matrix but regard the largest eigenvalue and corresponding eigenvector as noise, i.e.,

$$\tilde{\mathbf{C}}^d_{clip}(K) \triangleq \begin{cases} \tilde{\eta} \sum_{i=1}^{m} \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T & K = 1 \\ \sum_{i=2}^{K} \hat{\eta}_i \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T + \tilde{\eta} \hat{\mathbf{v}}_1 \hat{\mathbf{v}}_1^T + \tilde{\eta} \sum_{i=K+1}^{m} \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T & K \geq 2 \end{cases}$$

where $\tilde{\eta}$ is such that $\mathrm{Tr}\left(\tilde{\mathbf{C}}^d_{clip}(K)\right) = \mathrm{Tr}\mathbf{C}^d_S$. Reconstruct the corresponding "clipped" covariance matrix $\tilde{\boldsymbol{\Sigma}}^d_{clip}(K)$ by transforming the $\tilde{\mathbf{C}}^d_{clip}(K)$ back to covariance matrix using equation (2). Repeat Section II Question 4) b) by replacing $\boldsymbol{\Sigma}^d_{clip}(K)$ with $\tilde{\boldsymbol{\Sigma}}^d_{clip}(K)$. Superimposed on the same figure in Question 4) b), plot the variance of daily return versus K for $K = 1, 2, ..., 40, 60, 98$. [3 pts] What is the effect of regarding the largest eigenvalue and corresponding eigenvector as noise? [1 pts]

5) Compare the two estimators of covariance matrix we have introduced: the sample covariance matrix $\boldsymbol{\Sigma}^d_S$ and the clipped sample covariance matrix $\boldsymbol{\Sigma}^d_{clip}(K)$. Which one do you prefer in terms of the variance and daily return and total return? [1 pts] Give your reasons for your preference. [1 pts]

6) In practice, the portfolio does not have to be updated every day. For example, the clipped sample covariance matrix $\boldsymbol{\Sigma}^d_{clip}(K)$ computed from $\left[\mathbf{x}_{i-(w-1)}, ..., \mathbf{x}_i\right]$ can be used to compute the returns obtained on the t following days (based on the data $[\mathbf{x}_{i+1}, ..., \mathbf{x}_{i+t}]$) before being updated on the $i+t+1$-th day. Compute the total return and variance of the daily return for the portfolio period from

the 201th day to 753th day using the covariance matrix $\mathbf{\Sigma}^d_{clip}(K^\star)$ when the portfolio is updated every $t = 1$ day, 5 days, 10days, 15days, ..., 60 days. Here $K^\star$ is an arbitrary value you can choose in the reasonable range you obtained in Question 4) b). Plot the curve of variance of daily return versus update period as specified. [3 pts] How does the frequency of the portfolio update affect the overall risk? [1 pts]

7) In Section II Question 4), we "clipped" the covariance matrix by first clipping the correlation matrix, and we call this method indirect clipping. Now we consider directly clipping the covariance matrix. Assume the sample covariance matrix has the eigenvalue decomposition

$$\mathbf{\Sigma}^d_S = \sum_{i=1}^{m} \hat{\mu}_i \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^T,$$

where $\hat{\mu}_1 > \hat{\mu}_2 > \cdots > \hat{\mu}_m$ are the ordered eigenvalues and $\hat{\mathbf{w}}_i$ are the corresponding eigenvectors. The direct clipped covariance matrix can be constructed by

$$\hat{\mathbf{\Sigma}}^d_S(K) = \sum_{i=1}^{K} \hat{\mu}_i \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^T + \bar{\mu} \sum_{i=K+1}^{m} \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^T,$$

where $\bar{\mu}$ is such that $\mathrm{Tr}(\mathbf{\Sigma}_S) = \mathrm{Tr}\left(\hat{\mathbf{\Sigma}}_S(K)\right)$.

Repeat Section II Question 4) b) by replacing $\mathbf{\Sigma}^d_{clip}(K)$ with $\hat{\mathbf{\Sigma}}^d_S(K)$. Superimpose on the graph you obtained in Question 4) b), plot the variance of daily return versus K for $K = 1, 2, ..., 40, 60, 98$. [3 pts] Which clipping method is better in terms of variance of daily returns, the direct clipping or the indirect one? [1 pts] Give a reasonable explanation to why it is better. [3 pts] (You may want to do similar analysis to the sample covariance matrix $\mathbf{\Sigma}_s$ as you did to the correlation matrix $\mathbf{C}_S$ in Section I Question 3) - 5) and make any interesting interpretation.)