Statistical Machine Learning – Homework 1 Solution

Credit to Shuaiwen Wang

Problem 1

Proof. 1.

- Write $L(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$;
- Get log-likelihood $\ell(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta)$;
- To find $\max_\theta L(\theta)$, it is equivalent to find $\max_\theta \ell(\theta)$ (since logarithmic transformation is monotone); As a result we look for the solution of the equation $\nabla_\theta \ell(\theta) = \sum_{i=1}^{n} \nabla_\theta p(x_i|\theta) = 0$. Notice that the solution may also be a minimizer, so we need to exclude these cases;
- The result we get finally is the MLE.

2. Since $p(x|\theta) = \left(\frac{\nu}{\mu}\right)^\nu \frac{x^{\nu-1}}{\Gamma(\nu)} e^{-\frac{\nu x}{\mu}}$, we have $\ell(\theta) = \sum_{i=1}^{n} \nu \log \frac{\nu}{\mu} + (\nu - 1)\log x_i - \log\Gamma(\nu) - \frac{\nu x_i}{\mu}$ and then

$$\frac{\partial \ell(\theta)}{\partial \mu} = \sum_{i=1}^{n} -\frac{\nu}{\mu} + \frac{\nu x_i}{\mu^2} = n\left(\frac{\nu \bar{x}}{\mu^2} - \frac{\nu}{\mu}\right) = 0$$

This gives us $\hat{\mu} = \bar{x}$; easy to see this is maximizer.

3. $\frac{\partial \ell(\theta)}{\partial \nu} = \sum_{i=1}^{n} \log \nu + 1 - \log \mu + \log x_i - \phi(\nu) - \frac{x_i}{\mu} = \sum_{i=1}^{n} \log \frac{\nu x_i}{\mu} - \phi(\nu) - \left(\frac{x_i}{\mu} - 1\right) = 0.$ □

Problem 2

Proof. In this problem, the response Y takes values from K categories. As explained in the statement of the problem, the classifier can be viewed as an itegrable function from $\mathbb{R}^d$ to $\{1, 2, \cdots, K\}$ (this is denoted as $[K]$). For any classifier f, given the value of X, the conditional risk is

$$R(f|X = \boldsymbol{x}) = \sum_{y \in [K]} L^{0-1}(y, f(\boldsymbol{x}))P(y|X = \boldsymbol{x}).$$

To understand this expression, let's consider a simple case. Assume that we only have one input variable, denoted as X_1, which is distributed according to a standard normal distribution. Our model is that if $X_1 < 0$, then the response Y takes value 1 with probability 0.8, and 0 with probability 0.2. Otherwise, Y can be either 0 or 1 with equal probability. Our classifier is that we always classify a new observation as 1. Then when $X_1 = -1$, the conditional risk is

$$\begin{aligned} R(f|X_1 = -1) &= L^{0-1}(1,1)P(Y=1|X_1=-1) + L^{0-1}(0,1)P(Y=0|X_1=-1) \\ &= 0 \cdot P(Y=1|X_1=-1) + 1 \cdot P(Y=0|X_1=-1). \end{aligned}$$

In this example, it's easily seen that the proposed classifier can minimize the conditional risk.

To show the result, we have

$$\begin{aligned} R(f|X=\boldsymbol{x}) &= \sum_{y\in[K]} L^{0-1}(y, f(\boldsymbol{x}))P(y|X=\boldsymbol{x}) \\ &\geq (1 - \max_{y\in[K]} P(y|X=\boldsymbol{x})) \\ &= R(f_0|X=\boldsymbol{x}), \end{aligned}$$

the equality holds when the classifer f is the Bayes classifier.

Therefore

$$\begin{aligned} R(f) &= \int_{\mathbb{R}^d} R(f|X=\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\ &\geq \int_{\mathbb{R}^d} R(f_0|X=\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\ &= R(f_0). \end{aligned}$$

This shows that the Bayes classifier is the optimal one under the 0-1 loss. □

Problem 3

Proof. 2. From Figure 1 we can see that

- First from the screeplot, we can see that the variances of the first two PCs are much higher than the rest, which means the first two components can represent the distribution of the data without normalizing the variances.

- Notice that there are several clusters based on the dates of the stocks (sort of depends on seasons). From this we may conclude that the prices within each season are similar, and vary between seasons.

- Boeing is very significant in the first component. Goldman Sachs is very significant in the second component. This means that the stock price of Boeing varies a lot compared with others. Both the first and the second PCs can explain the variance in IBM.

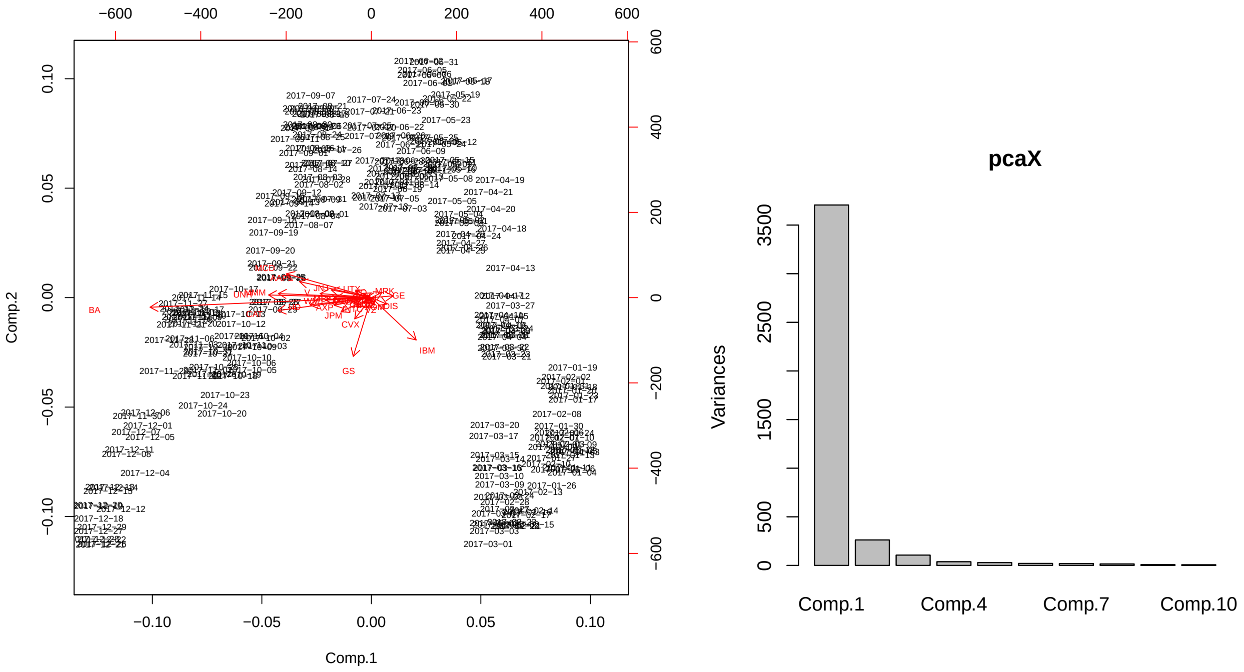

Figure 1: `biplot` and `screeplot` when `cor=F`.

3. From Figure 2 we can see that

- The first PCs are all significant. Thus only use the first two may lose some of the features in the data.

- The cluster depending seasons still exists.

- All stocks are equally important in the first two principal components. Some of them are more important in the first one. Others are more important in the second. Besides, some stocks from the same industry are more correlated (since the vectors are closer between them), like McDonald's and

Coca-Cola, Apple and Microsoft, American Express and Visa, Cisco and Intel, the Home Depot and Walmart, and 3M and United Technologies.

Seems three to six PCs can be reasonable, depends on subjective judgment.

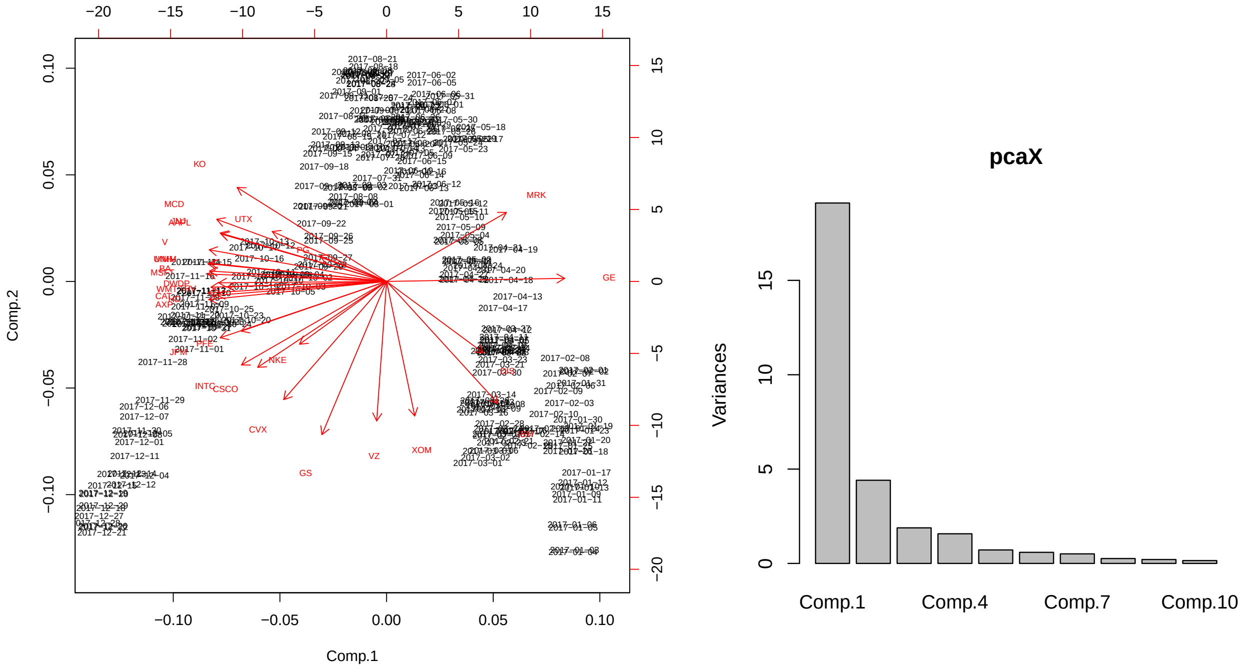

Figure 2: `biplot` and `screeplot` when `cor=T`.

4. From Figure 3 we can see that

- The first PC is has a much larger variance than other PCs; This means the data tends to distribute along a certain line.
- The return seems to mix along the entire year. (No obvious seasonal change)
- The returns of the stocks are correlated with each other (Some of them are highly correlated). Some vectors are still clustered by industries.

The screeplot should look flat if all the stocks are independent and randomly changed with each other.

□

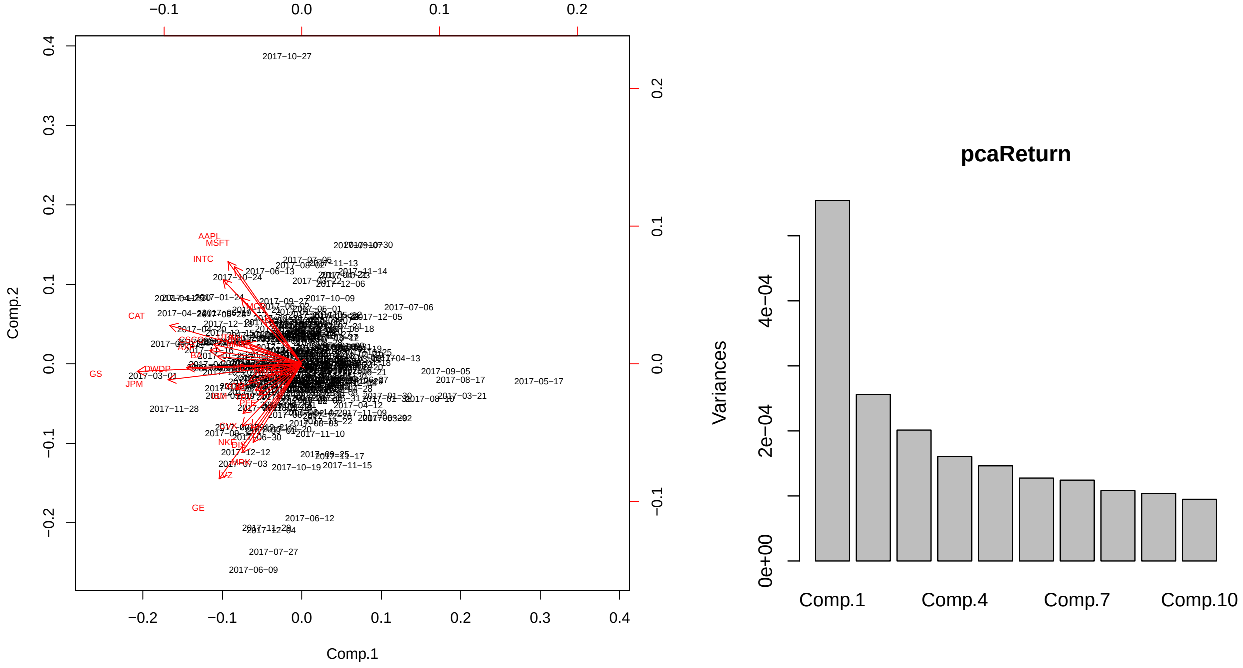

Figure 3: `biplot` and `screeplot` of return when `cor=T`.