Statistical Machine Learning – Homework 1 Solution

[redacted]

[redacted]

Problem [redacted]

Proof. The error we would like to minimize is $\mathbb{P}(Y \neq f(X)) = \mathbb{E}\mathbb{1}_{\{Y \neq f(X)\}} = \mathbb{E}_X \mathbb{E}_{Y|X} \mathbb{1}_{\{Y \neq f(X)\}} = \mathbb{E}_X \mathbb{P}(Y \neq f(X)|X) = \int \mathbb{P}(Y \neq f(X)|X = x) p_X(x) dx$. Thus in order to minimize $\mathbb{P}(Y \neq f(X))$, it is sufficient to instead minimize $\mathbb{P}(Y \neq f(X)|X = x)$ for each x. Notice that

$$\mathbb{P}(Y \neq f(X)|X = x) = \mathbb{P}(Y \neq f(x)|X = x) = 1 - \mathbb{P}(Y = f(x)|X = x) \geq 1 - \max_j \mathbb{P}(Y = j|X = x)$$

Thus we know $\hat{f}(x) = \operatorname{argmax}_j \mathbb{P}(Y = j|X = x)$. □

Problem ■

Proof. 2. From Figure 3 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.
- Apple is very significant in the first component. Goldman Sachs is very significant in the second component. This means that the stock price of Apple varies a lot compared with others.

3. From Figure 4 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.

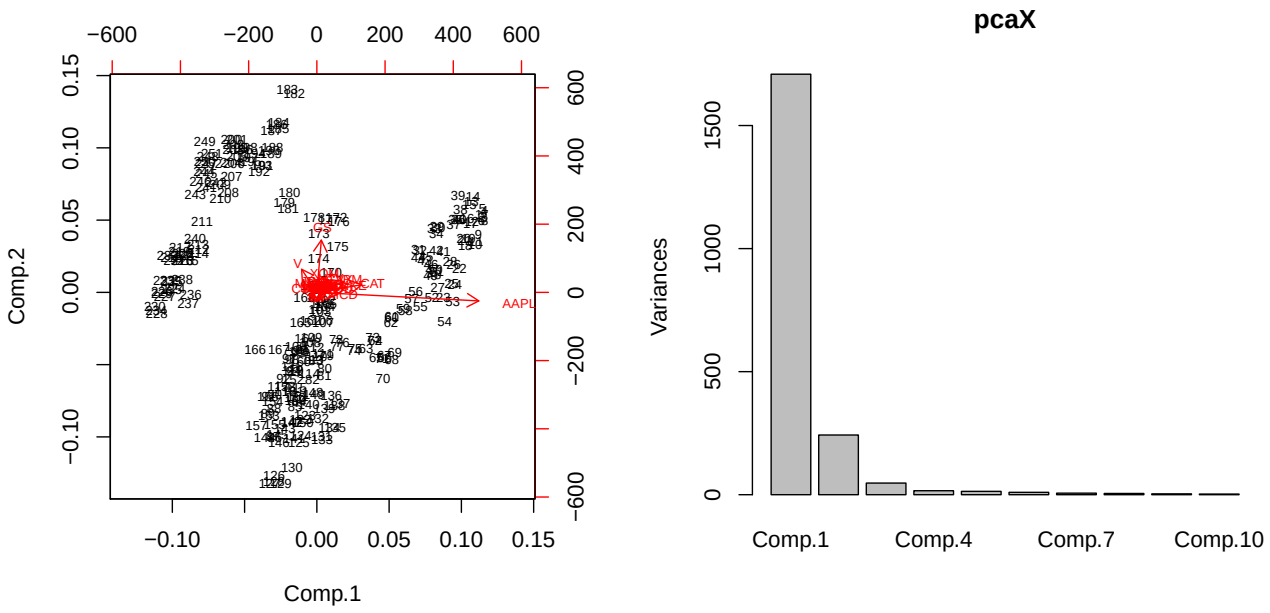

Figure 3: `biplot` and `screeplot` when `cor=F`.

- 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 of them are more correlated with each others (since the vectors are closer between them).

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

4. From Figure 5 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). They are equally important in the first principal component. (Does this tell us that the entire market played a more important role. Time and single company do not really matter that much?)

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

□

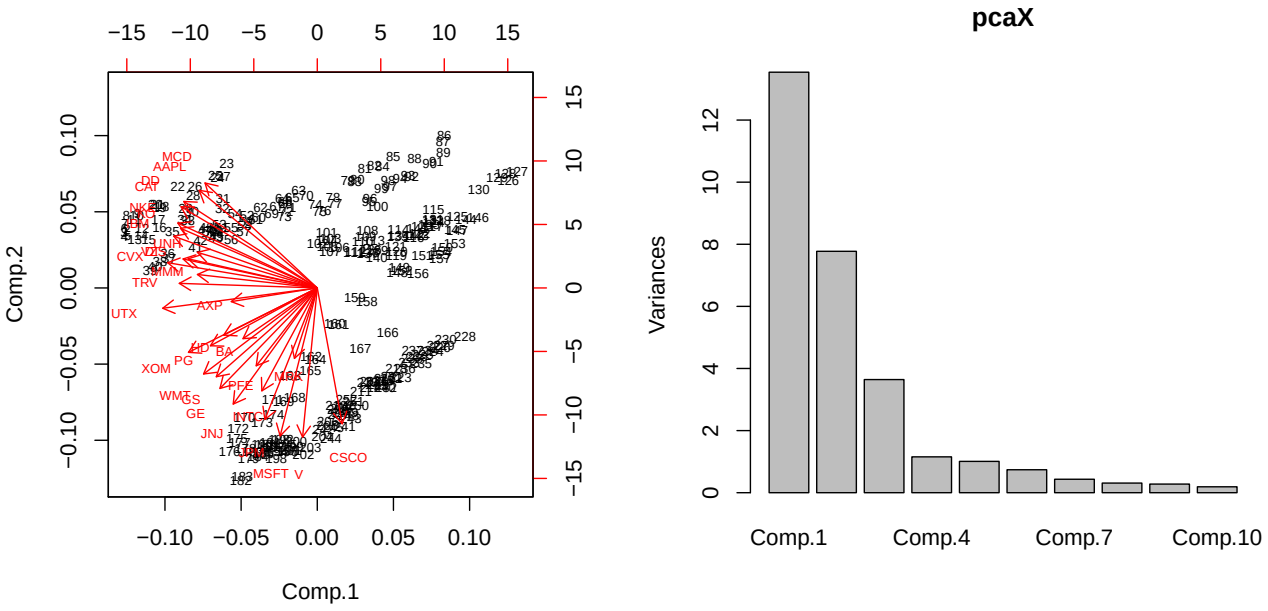

Figure 4: `biplot` and `screeplot` when `cor=T`.

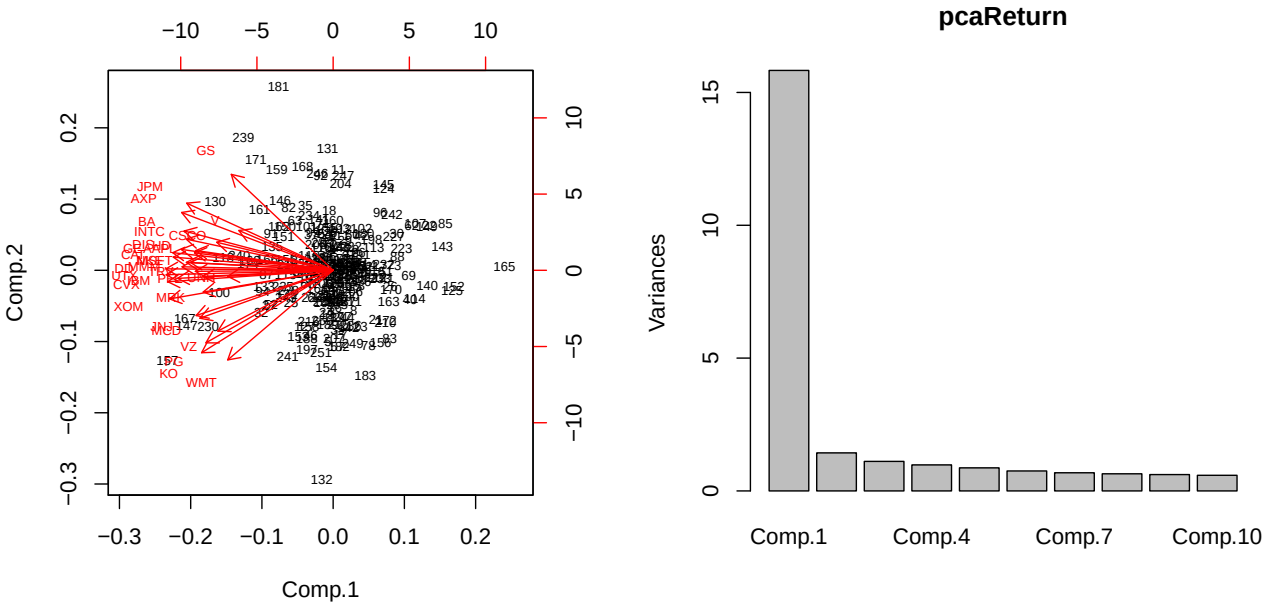

Figure 5: `biplot` and `screeplot` of return when `cor=T`.