**Rebuttal**

José Vinícius de Miranda Cardoso, Yixin Men, and Shunkang Zhang

First, we would like to thank the reviewers for their insightful comments. In this document we hope to address the very valid concerns draw by them.

Principal Component Analysis (PCA), which is used to summarize the information contained in a continuous (i.e, quantitative) multivariate data by reducing the dimensionality of the data without losing important information.

PCA should be used mainly for variables which are strongly correlated. If the relationship is weak between variables, PCA does not work well to reduce data. Refer to the correlation matrix to determine. In general, if most of the correlation coefficients are smaller than 0.3, PCA will not help.

We use correlation between different stocks of daily returns as the input. The reason why we use correlation matrix not covariance matrix is because you tend to use the covariance matrix when the variable scales are similar and the correlation matrix when variables are on different scales. As we observe from the below plotting, the stocks vary a lot among different industries.

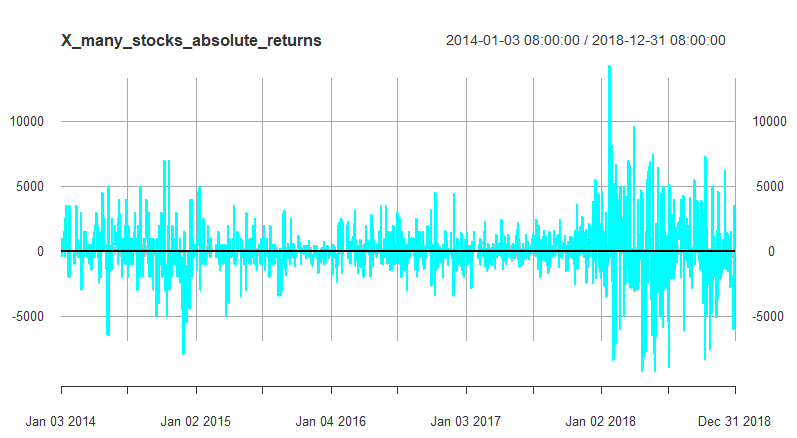


Figure 1. Plotting of Returns of the Stocks in three industries

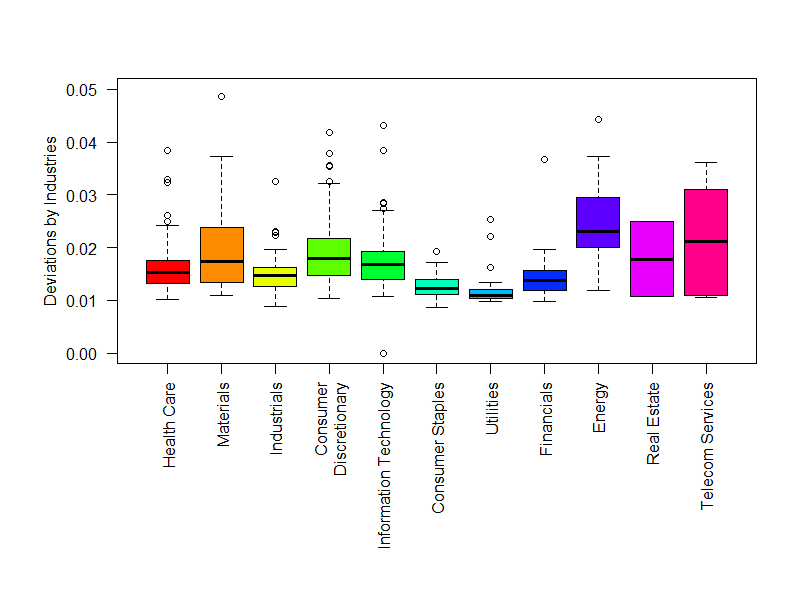


Figure 2. Deviation through industries

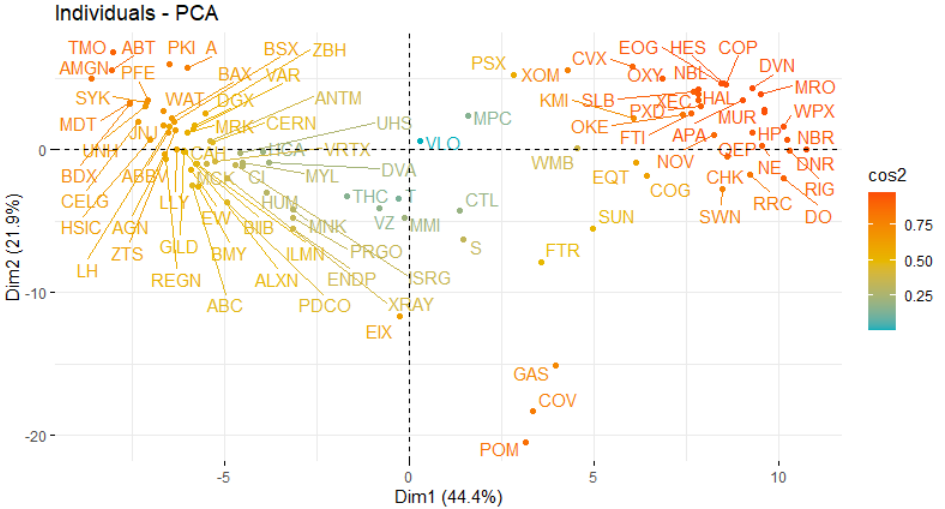


Figure 3. PCA applied correlation matrix

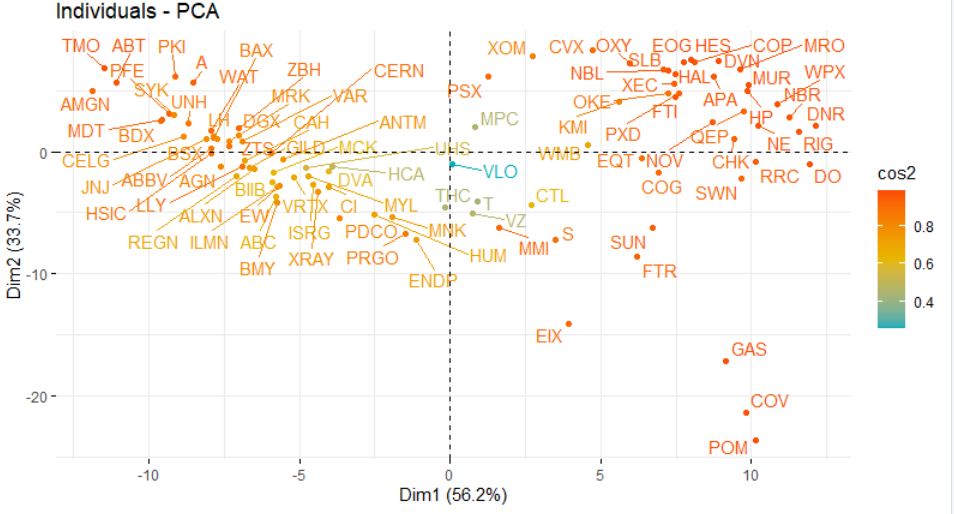


Figure 4. Robust PCA applied correlation matrix

Due to the limit of space, we did not show enough results obtained. The Robust PCA gives stronger correlation (darker colour) and denser concentration after denoising.

For better evaluation metrics, we could have compared the connectivity degrees of each stock, the smaller the connectivity the better since it’s well known that stocks are correlated with a sparse number of other stocks.

For the workload allocation, I am mainly responsible for data preprocessing and analysing. Jose contributed most of the R code for extracting correlation features of daily returns. Zhang contributed Python code and mathematical theories. I ran R code from Jose’s contribution also applied RCA and Robust RCA to verify his discovery. Since we chose poster to present our work, we did not put on all the process we have done.

We think the graph can show the relationship between different stocks. Actually, we also have the heatmap result which can also show the correlation result between different stocks. What’s more, we will add the cluster result in the following version. As for the writing part, we have a lot of analysis result and we do not have enough space to show it. Let us put it on appendix.

As for our project, we just want to compare the PCA and robust PCA and see the consistency result between the real case and estimated case. To be honest, I am not sure what the new is. Maybe we can add some clustering and classification results into our report and try to run some nonlinear dimension reduction methods. However, in the real application, I think the linear model is enough. By doing this, we can filter some noisy signal. If you tried to do some more complicated dimension reduction methods, I am not sure whether it is valuable. For example, as for the SNP data, people used to do dimension reduction to extract PC by using PCA. Why not researchers use more complicated method? In a word, we will add more analysis result in appendix.

We can use all the 10 classes to draw the data. In our poster, we just want to show the graph more clear. We do not try some manifold learning method because we believe that the basic linear project can help us learn a global representation. Let us add the manifold learning method latter. Thank you.

As noticed by Mr. Avik DAS, the figure numbers were missing, which is now corrected. As regarding the discussion for stock industry selection, we decided to focus on three industries in order to show the modular aspects of stock data. We could have indeed used more industries, or different industries, but the results are likely to be similar.

Mr. HUANG points out that there was a lack of comparison among the clustering results presented in the figures. We agree with that and will include a more thorough discussion. Meanwhile, we suggest Mr. HUANG to also take a look at our supplementary material provided as a Jupyter Notebook. It contains a more detailed description of the meaning of the stocks graphs.

Mr. KUNDHU asks “are there any numerical metrics which show clearly how much the RPCA performs better than PCA?”. The answer is yes. We could have used some numeric objective measures like the edge density of the graphs to clearly show that RPCA gives improved results as compared to PCA. The numerical value of the hyperparameter is discussed in the supplementary material available as a Jupyter Notebook. We strongly agree with Mr. KUNDHU that the effect of the hyperparameter could have been explored.