ETF3500/5500 High Dimensional Data Analysis - Group Assignment

Department of Econometrics and Business Statistics, Monash University

Due Date: 25th September 2025

1 Context

Stock trading is an active form of financial investment. On the New York Stock Exchange (NYSE), over two thousand companies list their equity shares for the public to trade. Due to limited funds and stock trading constraints, investors often select a subset of stocks to form an investment portfolio. Selecting which companies to invest in can be dictated by many factors, including the industry in which the company participates in, the risk profile of the company, and the investor's preferences. In this assignment, we will look at using high-dimensional data analysis methods to analyze the returns on stock prices of up to one hundred companies that are listed on the NYSE. These techniques are useful in dissecting the variations of the stock returns into common variation (also known as systematic risk) and variation that is unique to the stock (also known as idiosyncratic risk).

2 Data

You are provided the stock returns data in your relevant data folder. The file `SampleX.csv` contains monthly returns of each of the one hundred stocks, recorded between January 2005 and December 2019. The final letter `X` in the filename should correspond to your group name on Moodle. Make sure you use the correct file.

The first column of the data contains the date, and the remaining columns record the returns for the large selection of stocks included in your dataset. The column label serves as the identifier for each stock. The identifier consists of one letter and five numerical digits. The first letter in this identifier indicates the industry in which the company operates:

- B = Mining
- C = Construction
- D = Manufacturing
- E = Transportation and Public Utilities
- F = Wholesale Trade
- G = Retail Trade
- H = Finance, Insurance and Real Estate
- I = Services

The five numerical digit following the letter is the stock's PERMNO code, which is used by the database provider, CRSP. See https://libguides.stanford.edu/library/wrds/identifiers-linking/files for further details. For example, the stock under the identifier `C52708` operates in the construction industry (industry code `C`), and its PERMNO identifier code is `52708`.

The market return, computed as the monthly return based on the S&P500 market index, is also provided in the file `Market.csv`.

3 Key Tasks

Your task is to use the techniques covered in the subject so far to analyze the large collection of stock returns provided in the dataset. Our objective is to provide analysis of the common factors that drive the covariation in stock returns, and to provide a way of selecting stocks based on their covariation patterns.

You must address the following questions in your analysis:

- Does the largest principal component/most loaded factor really coincide with market movements? Discuss the extent to which this belief is supported by your data.
- Analyze which of the key industry group, indicated by the industrial code given above, has the greatest degree of loading on the first latent factor. Does your analysis change when you look at the other factors?
- Are there any industry group, as defined by the letter identifier above, that move in the opposite directions of the common factors? Which industry group(s) tend to move in the same direction with the common factors? Comment on the heterogeneity in the factor loadings across the industry groups.
- Use your results provide commentary on which industry is expected to contain the most amount of systematic risk. Systematic risk is defined as risk, or variations, that penetrates through the market and cannot be eliminated by portfolio diversification.
- If you are advising an investor whose preference is to invest in stocks that tend to move with the market and has the capacity to only invest in five stocks, which stocks would you recommend?

This list is not exhaustive. **You are strongly encouraged** to address and investigate additional issues that are not listed in this assignment brief.

You **must** use Principal Component Analysis (PCA) and Factor Models in your analysis. You may also use the other techniques covered in the unit such as cluster analysis, multidimensional scaling and Gaussian mixture modelling, but each of these is optional. You must summarise your results in a report of **no more than 1500 words**. Your R code and any additional work not directly described in your report must be included in an Appendix (this will not count towards the word limit). The maximum page limit for your Appendix is 10 pages.

4 Guidance on the report

To assist you with structuring your report, a list of questions/hints are provided below. These are designed to prompt you to think about your analysis and the presentation of results, and will influence the grading of the assignment.

- Is the data tidy? Are there missing values, outliers or other data credibility issues?
- Can you derive any insights from the data using simple exploratory analysis including summary statistics and visualization tools?
- How did you select the number of factors for your factor model? Are your choices supported by any preliminary analysis?
- Hint: to fit the factor model for this large data set, you need to adjust the default convergence criteria for the uniqueness of the model. Consider inputting `lower=0.05`, or something a bit higher. The default is `lower=0.005`.
- Does the report contain enough information to be reproduced by somebody with knowledge of the techniques used?
- Are all plots clearly presented, labelled and correctly explained?
- What assumptions needed to be made to conduct your analysis? Make sure you discuss them.
- Are the limitations of your analysis clearly discussed?
- Your report should focus on providing insights to the key questions posed in the **Key Tasks** above. Avoid simply listing R codes/outputs/tables/graphs without providing discussions in the context of the tasks.

5 Submission

The assignment is a **group assignment**. The maximum group size is five people. You are strongly encouraged to form groups within your tutorial, so that you can make the most use of your class time to complete the assignment tasks. Group members may come from different unit codes. A single soft copy should be submitted per group via Moodle, with the group number and members' information clearly visible on the front page.

You will also need to fill in the **Feedback Fruit** for peer evaluation by the due date. This is an individual task, so make sure that you do this independently. The link to this form will be provided on Moodle under "Group Assignment" section. Each member of the group must provide feedback on all members.

Peer review of your contribution to your team will contribute to your final group assignment score, with a weighting of 80% given to the report and 20% given to the peer evaluation.

The Chief Examiner reserves the rights to adjust the report score if a certain group member is deemed to have minimal contribution to the group assignment.