ELEN90094 – Large Data Methods & Applications (Semester 2, 2024)
The University of Melbourne

Project: Investigation of methods for resolving statistical noise and understanding correlation structure in high dimensional data

Project Guidelines:

The course project involves investigation of methods related to high dimensional data analysis, as well as independent exploratory analysis on real data. You are required to identify and critique approaches for resolving true correlation information and to characterize the effects of statistical noise in high dimensional data. The analysis will be conducted in two stages.

In the **first stage**, you are required to study and explain different data analysis methods and results that are presented in Ref. [1] for investigating correlation patterns in complex, high-dimensional stock return data.

Ref. [1]: V. Plerou, et al., Random matrix approach to cross correlations in financial data, Physical Review E 65.6, p 066126, 2002.
[Visit https://journals.aps.org/pre/abstract/10.1103/PhysRevE.65.066126 -> pdf -> Log in via your institution -> Find your institution: University of Melbourne]

You are expected to:

- Describe the basic problem being addressed, motivation for the study.
- Describe the rationale for the use of high-dimensional data analysis methods, including why more standard statistical analysis tools may not be sufficient.
- Describe in your own words the different mathematical and data analysis approaches that were applied in the analysis in Ref. [1]. In your description, you should refer clearly to specific sections of Ref. [1]. (*You can exclude the methods used in Sections V and VIII.)*
- Interpret the results presented for the different analysis approaches, describing what the results are and why they are useful or important.
- In your descriptions and interpretations, you should demonstrate your understanding of concepts covered in ELEN90094 by linking your discussions to content that was covered in class, tutorials, and/or assignments.

In the above tasks, you are expected to present explanations and summaries in your own words, with your own interpretations.

For the **second stage,** you are required to apply data analysis methods from the first stage, complemented by additional methods covered in the course, to investigate correlation patterns in complex, high-dimensional data sets.

Two data sets of financial stock returns, along with associated metadata and data descriptions are provided on Canvas. These data sets will be discussed in class and tutorials. They comprise stock returns for periods (i) prior to, and (ii) after, the emergence of COVID-19. You are required to analyse both data sets for your analysis.

You are expected to:

- Conduct numerical experiments that apply the data analysis methods from Ref. [1] and methods/tools covered in the class, tutorials, and/or assignments, to the data sets provided.
- Clearly describe each of the experiments that you have performed, why you have chosen the specific methods and interpret the results of each experiment.
- For any methods not covered in the first stage (i.e., methods covered in class, tutorials, and/or assignments, but not used explicitly in Ref. [1], e.g., bi-plots, QQ-plots, inverse PCA interpretation, etc), you should explain what these methods are along with the key or unique information they are providing.

Your analysis and interpretation in this second stage of the project should seek to:

- Distinguish "true" correlation information in the data sets from spurious correlation information reflecting high-dimensional statistical noise.
- Quantify the amount of statistical noise in the estimated correlation values.
- Reveal structured correlation patterns/networks (if any) that are not strongly affected by noise.
- Explore the dynamical (time varying) stability of estimated correlation matrices and their properties (e.g., eigenvalues, eigenvectors).
- Provide meaningful and well-founded interpretations of your analysis, including practical interpretation. For example, you may reflect broadly on what insights your analysis provides regarding the dependencies among stock returns in your analysed data sets, the complexity of these dependencies, comparisons and interpretations across data sets, etc.

Your results should be documented into a clearly written report in the required format, as outlined below.

Structure of Written Report:

The written report should **have the following sections** (marks for each section are indicated; total score of 35 points):

1) Introduction (including problem description and motivation)	6pt
2) First Stage analysis/investigation	11pt
3) Second Stage analysis/investigation	12pt
4) Conclusions (including overall insights & reflections)	4pt
5) List of References of published material	1pt
6) Clear statement of contributions of team members (with signatures)	1pt

Submission Requirements:

You are expected to provide a **written report together with the code and associated files** which addresses the criteria outlined above. The complete package should be submitted via Canvas.

The report should satisfy the following requirements:

- The length should not exceed 30 pages, single column, with a minimum of 11-point font. Be brief and to the point. The project report should be submitted in PDF format. It can be written using Microsoft Word or other text editors.
- **The project should be done in groups of 2 or 3**. Requests to work individually will be considered on a case-by-case basis; you should discuss with the instructor and demonstrator if you wish to work individually.
- A single report is to be submitted for each group, however, a **clear and detailed description of the individual contributions of each group member must be included** (stating which members performed which task). All group members need to **add their signature** to acknowledge the contribution statements.
- All supporting code should be submitted. Make sure that all codes are readily executable/interpretable, and your submission includes all dependencies (if any) to re-run the code on a different machine. Your programs should be written in either MATLAB, Python or R. Ensure that you clearly mention the programming language and version used to develop your code.

Schedule of submission and grading

Component	**Grade**	**Deadline**
Preliminary project report	10%	9am, October 9, 2024
Final report	35%	9am, October 28, 2024
Presentation	15%	9am, October 28, 2024 (Note: presentation grading includes oral evaluation. This will be scheduled subsequently)

Late penalty: 20% of the individual component's grade per day late.

Expectations for the Preliminary Report submission

As a minimum, the preliminary report submission is expected to contain a clear Introduction and to document significant progress in the First Stage analysis. As for the full report, the preliminary report should clearly indicate the **contribution of each member** of the group, and all group members need to **add their signature** to acknowledge the contribution statements.

A clear plan for the remaining tasks (including a plan of the simulations to conduct) should also be provided. The plan should include a clear breakdown of the tasks and how each member of the group will contribute to these tasks.

Feedback on the preliminary report will be provided to assist with preparation of the final report.

Important note on plagiarism

You must **cite all individual sources of information in your report**. For example, you should make proper reference to the source of any code/software that you have used but not written yourself; if you have extracted any figures or tables or other material from published documents, these must be referenced. Conversely, when you generate your own data, run your own experiments, generate your own figures, etc., this should also be explicitly identified in your report. It is your duty to make it clear what you have done and what you have extracted from literature. The review and evaluation of methods in the report must convey your own thoughts and ideas, and you must not copy extracts of text (full sentences, paragraphs, etc.) from published material. This will show up clearly on the anti-plagiarism software. Note that an anti-plagiarism check will be performed once you submit your final report onto Canvas, and this will be carefully scrutinised. The University of Melbourne promotes the highest ethical standards in research, learning and teaching. Note that plagiarism is an offence against the University's regulations. Carefully read about the University's Policy and Regulations on Academic Integrity including plagiarism and collusion: https://academicintegrity.unimelb.edu.au/.