ETF3500/5500 High Dimensional Data Analysis - Group Assignment

Department of Econometrics and Business Statistics, Monash University

Due Date: 22nd September 2023 at 4:30PM

1 Data

The Gini index is a measure of income inequality. This assignment uses Gini index data on the set of 52 states of the U.S. The data was sourced from U.S. State-Level Income Inequality Data - Mark W. Frank.

For this assignment, we have provided you with the Gini index of the 52 U.S. states over a selection of years. The first block, available in *Inequality_GD.csv*, covers the period between 1929 and 1945. This period covers the Great Depression and World War II. The second block, available in *Inequality_GR.csv*, covers the period between 2007 and 2015, which corresponds to the Global Financial Crisis and the economic downturn that followed.

2 Task

The task is to investigate how the historical events of the Great Depression, the World War II and the Global Financial Crisis impacts income inequality for the 52 U.S. states. Your are required to conduct some preliminary analysis on the data, including data cleaning and providing basic summary and visualization of the data, and conduct analyses using the techniques covered so far in this unit. The only mandatory requirement is that you **MUST** use principal component analysis (PCA).

You may also use the other techniques covered in the unit such as cluster analysis and multidimensional scaling, 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.

3 Guidance

To assist you, a list of questions are provided below. These are designed to prompt you to think about the analysis and will influence the grading of the assignment. **You are strongly encouraged** to address and investigate issues that are not listed here. This list is not exhaustive.

- Is the data clean? 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 can you profile the principal components? Do they have some interpretation in terms of the data itself?
- 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 interesting features discovered by the analysis and should not simply list everything that was attempted.

4 Submission

The assignment is a **group assignment**. The maximum group size is four people. You may form groups with students from different tutorial groups and from different unit codes. A single soft copy should be submitted with a group assignment cover page added to the front via Moodle by the due date. Peer review of your contribution to your team will be taken into consideration when marking the assignment.

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 "Assignment 2" section.

Note that we reserve the rights to adjust your individual assignment score, subject to the peer evaluation you received from your fellow group members.