High Dimensional Data Analysis Assignment 2

Department of Econometrics and Business Statistics, Monash University

Due Date: 21st September 2022 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.

The data contains the following 11 variables.

- **Gini1918:** Gini index in the year 1918
- **Gini1928:** Gini index in the year 1928
- **Gini1938:** Gini index in the year 1938
- **Gini1948:** Gini index in the year 1948
- **Gini1958:** Gini index in the year 1958
- **Gini1968:** Gini index in the year 1968
- **Gini1978:** Gini index in the year 1978
- **Gini1988:** Gini index in the year 1988
- **Gini1998:** Gini index in the year 1998
- **Gini2008:** Gini index in the year 2008
- **Gini2018:** Gini index in the year 2018

The full dataset can be found on Moodle under the name *Inequality_data.csv*.

2 Task

The idea is that we want to investigate how inequality has evolved over time for different states. Your are required to conduct some preliminary analysis on the data. The only mandatory requirement is that you **MUST** use principal component analysis (PCA). In addition, 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 additional work not crucial to the analysis can be included in an Appendix (this will not count towards the word limit).

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. If you can think of issues not listed here then you are encouraged to address them.

- Is the data clean? Are there missing values, outliers or other data credibility issues?

- Can you derive any insights from the data from simple exploratory analysis including summary statistics and basic plots?
- Can the data be easily visualised?
- 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 and correctly explained?
- Is the analysis robust to minor changes in the methodology?
- Are any assumptions made for the analysis or in drawing conclusions. If so, are these clearly explained?
- Does the report focus on a small number of interesting features of the analysis or does the report simply list everything that was attempted (the former is preferable to the latter)?
- Are the limitations of the analysis clearly discussed?

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. All assignments should be assignment via Moodle. Peer review of your contribution to your team will be taken into consideration when marking the assignment.