ETF3500/5500 High Dimensional Data Analysis

Assignment 2

Department of Econometrics and Business Statistics

Monash University

-

An investigation report on the impacts of the Great Depression, World War II and the Global Financial Crisis on income inequality within states of United States of America.

-

Xianghe XU

14 January, 2024

Contents

1 Introduction and Motivation

In the realm of economic and social analysis, income inequality is always one of indicators that reveals the true economic and social status of an area or a country. The comprehension of the dynamics behind the income distribution over a period of time has long been a topic of significant interest for policymakers, economists, and public. Among the various factors that have influences on income inequality, historical events is one of the main factors that plays a profound and important role on income inequality.

This report embarks on a thorough investigation of the effects of three pivotal historical events, the Great Depression, World War II, and the Global Financial Crisis, on income inequality within the states of United States. Each of these historical events, characterized by unique and unexpected economic and societal shifts, has left indelible imprints on the nation's economic and financial landscape. In this report, the aim is to scrutinize how does the income distribution vary with the historical events mentioned above, and unravel the complex interplay between historical circumstances and contemporary income disparities.

2 Data Description

Source and structure of the data:

The data set used in this report was sourced from ***U.S. State-Level Income Inequality Data - Mark W. Frank.***, which offers a comprehensive panel of United States' annual state level income inequality measures and was constructed from individual tax filling data available from the Internal Revenue Service.

This report uses Gini index data to conduct an analysis on income inequality within the states of United States. The **Gini index**, also known as Gini coefficient, is a statistical measure of income or wealth inequality within a country or a social group. It quantifies the extent to which the income or wealth is distributed among areas of a country. The Gini index is a number between 0 and 1; a Gini index of 0 represents perfect income equality within a state where everyone has the same income or wealth, whereas a Gini index of 1 represents perfect income inequality within a state where one individual or household has all the income or wealth and everyone else has none.

The two data files used in this report for the analysis of income inequality are "Inequality_GD.csv" and "Inequality_GR.csv". These data files contain annual Gini index data for the states of the United States, a federal district(Washington), and the United States as a whole. The files "Inequality_GD.csv" and "Inequality_GR.csv" have dimensions of 50 observations and 19 columns, and 52 observations and 11 columns, respectively. The first data file, "Inequality_GD.csv", covers the period between 1929 and 1945 which corresponds to the occurrence of the Great Depression and World War II. The second data file, "Inequality_GR.csv", covers the period between 2007 and 2015, which corresponds to the Global Financial Crisis and the economic downturn that followed. Both data sets share a similar structure; the first column of the data set represents the states'rank according to state names in an alphabetical order, the second column contains the names of the states, and the remaining columns hold the annual Gini index for each state over the years. The first data set has 50 observations compared with 52 observations in second data set. This difference is due to Alaska and Hawaii joining in the year 1959. The inconsistency in the number of observations between the two data sets is believed to have no impact on the analysis of income inequality on the states of United state.

3 Preliminary Analysis

The Table 1 and Table 2 display the first six rows of the two data set. At a glance of the data in the table 1 and Table 2, it provides insight on how the data should be prepared for principal component analysis:

- The first columns of the data sets can be removed since they consist of index numbers, which can be considered irrelevant and redundant for the analysis.

Table 1: First six rows of Inequality-GD data set

...1	State	1929	1930	1931	1932	1933	1934	1935	1936
1	United States	0.5664781	0.4986438	0.4677332	0.4466704	0.4633539	0.4556289	0.4604073	0.4867856
2	Alabama	0.4636559	0.4341273	0.4177243	0.3861681	0.3985525	0.3897605	0.3858755	0.4278177
3	Arizona	0.4460411	0.4026354	0.3788175	0.3388463	0.3463233	0.3504196	0.3592113	0.3814880
4	Arkansas	0.4122764	0.3885373	0.3835296	0.3516845	0.3590294	0.3888165	0.3865042	0.4321355
5	California	0.5138855	0.4644068	0.4492710	0.4228227	0.4379376	0.4335451	0.4325291	0.4666549
6	Colorado	0.4913782	0.4684625	0.4539476	0.4237142	0.4372745	0.4355249	0.4514149	0.4806452

Table 2: First six rows of Inequality-GR data set

...1	State	2007	2008	2009	2010	2011	2012	2013	2014
1	United States	0.6285253	0.6259130	0.6242736	0.6268764	0.6313589	0.6449431	0.6319985	0.6383778
2	Alabama	0.6494615	0.6449882	0.6153094	0.5913605	0.5885409	0.5928221	0.5795680	0.5814627
3	Alaska	0.5624961	0.5625051	0.5669771	NA	0.5668116	0.5730165	0.5594128	0.5597337
4	Arizona	0.6024898	0.6020868	0.6082659	NA	0.6242259	0.6227468	0.6190730	0.6184750
5	Arkansas	0.6684059	0.6674106	0.6260257	NA	0.6108396	0.6115291	0.5976986	0.6058933
6	California	0.6296388	0.6264265	0.6398192	NA	0.6706768	0.6870044	0.6703479	0.6783118

- The variable 'State' can be converted to row names using `column_to_rownames()` function.
- There are missing values in the data set, operations may be needed for those missing values as PCA is not robust to missing data.

The tables 1 and 2 offer a glance of the Gini index data sets. The next step will involve utilizing summary statistics and visualization techniques for preliminary data analysis. The objective of this step is to prepare the data sets for principal component analysis.

The Table 3 shows the a statistical summary of the Gini index data sets; the left side is the summary for the 'Inequality_GD' data set, and the right side is for the 'Inequality_GR' data set."

Table 3: Summary statistics for the Gini index data sets.

	Gini_index
	Min. : 0.2987
	1st Qu.: 0.3937
	Median : 0.4229
	Mean : 1.6039
	3rd Qu.: 0.4539
	Max. :1000.0000

	Gini_index
	Min. :0.5411
	1st Qu.:0.5843
	Median :0.6022
	Mean :0.6087
	3rd Qu.:0.6304
	Max. :0.7114
	NA's :8

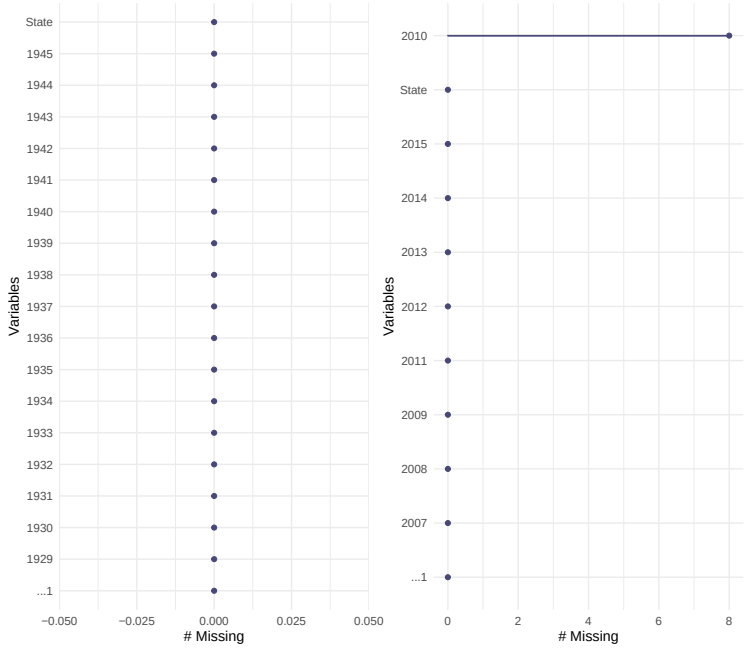

Figure 1: Visulazations of missing values in two data sets

The Figure 1 displays plots that provide at-a-glance representations of missing data within the data sets. The second plot (on the right) reveals the presence of eight missing values in the year 2010 in the second data set (`Inequality_GR`).

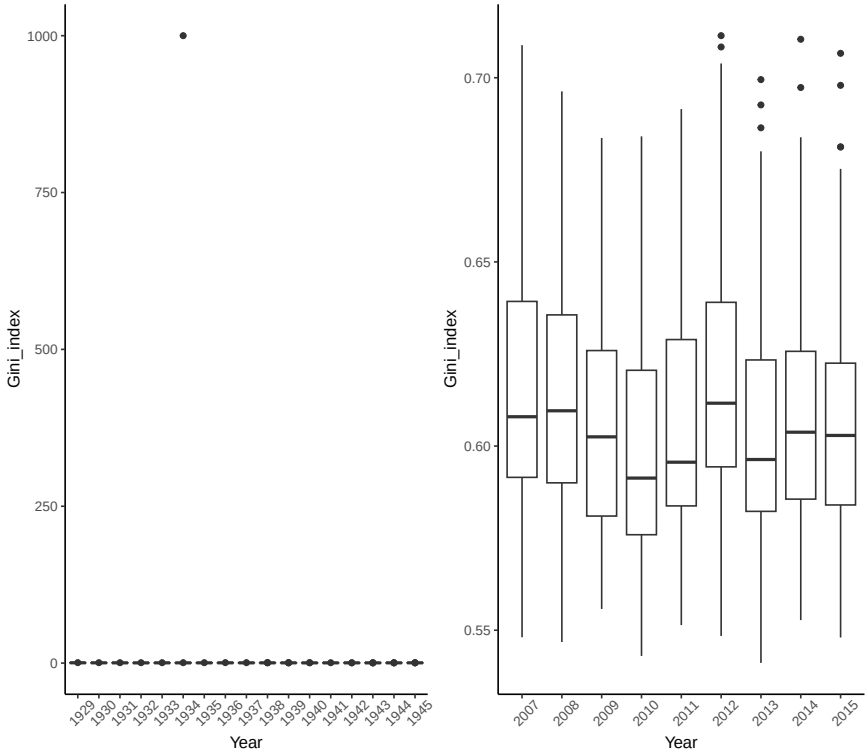

Figure 2: Visualisations of distributions of Gini index for each year

In Figure 2, the figure uses box plots to depict the Gini index data distributions for each year within each of the data sets. The plot on the right appears normal; all the distributions fall within the range of 0 and 1 for all the years. However, the plot on the left looks distinctly different compared to the right. In the left plot, all the distributions are compressed into a single line. The main reason for this is that the Gini index data in the left plot contains an extremely large value of 1000.

After conducting data exploration analysis on the data using summary statistics and visualization techniques, the tables and figures presented above demonstrate that there are two problems within the data sets:

- The data set `Inequality_GD` contains an extremely large value of 1000.
- The data set `Inequality_GR` has eight missing values in the year 2010.

From the statistical summary table 3 and Figure 2, it is evident that `Inequality_GD` data set has a maximum value of 1000. This is impossible in Gini index data because the Gini index cannot exceed 1. Also, the Table 3 and the Figure 1 has demonstrated that there are eight missing values in `Inequality_GR` data set. The extremely large value will introduce bias into the accuracy of principal component analysis, and traditional principal component analysis does not accept any missing data points. Therefore, we need to perform some operations to eliminate the presence of both the extremely large value and the missing values before conducting principal component analysis. Since there is only one extremely large value and a few missing values, the mean imputation method can be applied to both the missing values and the the outler value in order to perform the principal analysis and reduce the bias. (Mean imputation is a method in which the mean of the observed values for each variable is computed, and the missing values or outlier for that variable are replaced with this mean.)

Table 4: An extremely large Gini index

State	Year	Gini_index
Oregon	1934	1000

Table 4 displays the state and year information of the extremely large Gini index data point.

Table 5: First six rows of Inequality-GR data set

	Gini_index		Gini_index
	Min. :0.2987		Min. :0.5411
	1st Qu.:0.3937		1st Qu.:0.5845
	Median :0.4229		Median :0.6018
	Mean :0.4279		Mean :0.6086
	3rd Qu.:0.4537		3rd Qu.:0.6299
	Max. :0.7268		Max. :0.7114

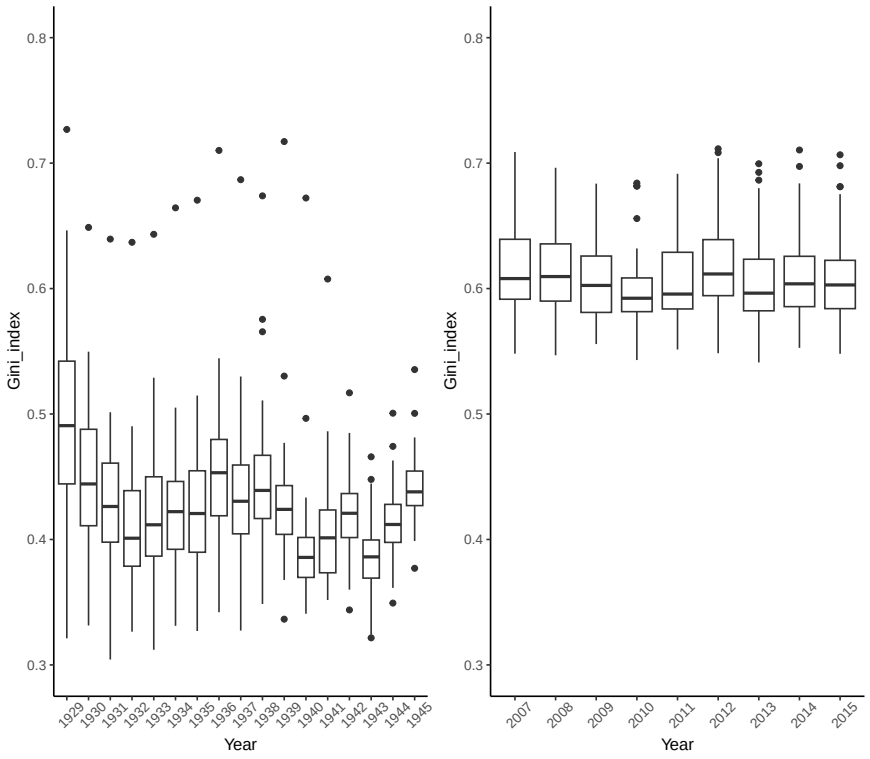

Figure 3: Visualisations of distributions of Gini index for each year

The above Table 5 displays the summary statistics of the two data sets after cleaning the data sets. From the table 5, it is clear that all summary values are within the range of 0 and 1, with no extremely large values or missing values present.

Furthermore, the use of boxplot in the Figure 3 displays the Gini index data distributions for each year within each of the data sets. After cleaning the data sets, both boxplots appear normal, with all data points falling within the range of 0 and 1.

The left boxplot illustrates the Gini index data distributions for each year over the period of the Great Depression and World War II. The spread of the distributions converges over the years, indicating that the difference in income inequality between the states became smaller over time. Moreover, the overall trend reveals that income inequality decreased in the first few years of the Great Depression, increased again, and then decreased once more in 1940, corresponding to the start of World War II before increasing again.

The boxplot on the right illustrates the Gini index data distributions for each year over the period of the Global Financial Crisis and the economic downturn that followed. Unlike the boxplot on the left, the right boxplot has a relatively steady spread of the Gini index distribution over the years. A small drop in the Gini index is observed in the year 2010, followed by an increase in the following years.

Overall, the distributions of the Gini index on the right plot are generally higher than those on the left plot. This indicates that higher income inequality is present across the states of the United States in recent years compared to the period from 1929 to 1945.

Now the cleaned Gini index data can be taken to principal component analysis to investigate the impact of historical events on income inequality.

4 Assumptions

Principal component analysis (PCA) has following assumptions :

Linearity : PCA is based on linear transformations of the data, which means that it assumes that the relationships between variables are linear. PCA will not be accurate if the relationships are not linear.

Orthogonality : PCA assumes that the principal components are uncorrelated to each other.

Normal distribution of data : PCA assumes the data is normally distributed.

5 Principal Component Analysis

In this report, Principal Component Analysis (PCA) will be conducted on two data sets. PCA is a statistical method used to simplify and reduce the dimensionality of a data set while retaining as much essential information as possible. It achieves this by transforming the original variables into a new set of uncorrelated variables known as 'principal components.' These principal components are uncorrelated linear combinations of the original variables and are ordered so that the first component explains the most variance in the data, the second component explains the second most, and so on. The goal is to use as few principal components as possible to capture the majority of the variance. To determine the optimal number of principal components, we will employ methods such as the scree plot and Kaiser's Rule. Moreover, principal component analysis provides new variables (principal components) that are linear combinations of the original variables. These components are hard to interpret in terms of the original features, making it challenging to explain the results.

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.5886 1.6110 0.78718 0.57040 0.35566 0.35120 0.30521
## Proportion of Variance 0.7575 0.1527 0.03645 0.01914 0.00744 0.00726 0.00548
## Cumulative Proportion  0.7575 0.9102 0.94664 0.96578 0.97322 0.98047 0.98595
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.23891 0.21639 0.20273 0.17386 0.15526 0.11394 0.10631
## Proportion of Variance 0.00336 0.00275 0.00242 0.00178 0.00142 0.00076 0.00066
## Cumulative Proportion  0.98931 0.99206 0.99448 0.99626 0.99768 0.99844 0.99911
##                           PC15    PC16      PC17
## Standard deviation     0.09402 0.07980 2.626e-07
## Proportion of Variance 0.00052 0.00037 0.000e+00
## Cumulative Proportion  0.99963 1.00000 1.000e+00

## Importance of components:
##                            PC1    PC2     PC3     PC4    PC5     PC6     PC7
## Standard deviation     2.6438 1.2976 0.47878 0.19353 0.1671 0.12942 0.08702
## Proportion of Variance 0.7766 0.1871 0.02547 0.00416 0.0031 0.00186 0.00084
## Cumulative Proportion  0.7766 0.9637 0.98919 0.99335 0.9965 0.99832 0.99916
##                            PC8     PC9
## Standard deviation     0.07337 0.04686
## Proportion of Variance 0.00060 0.00024
## Cumulative Proportion  0.99976 1.00000
```

The above output displays standard deviation, proportion of variance and cumulative proportion of variance for each principal component for the `Inequality-GD` and `Inequality-GR` data sets, respectively. The above output conveys that over 90% of total variance is explained by the first two principal components together for each of the data sets.

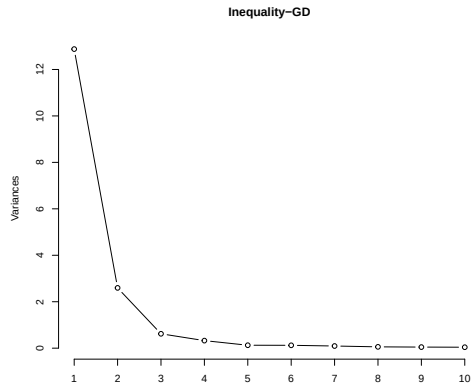

Figure 4: Scree plots for 'Inequality-GD' data set

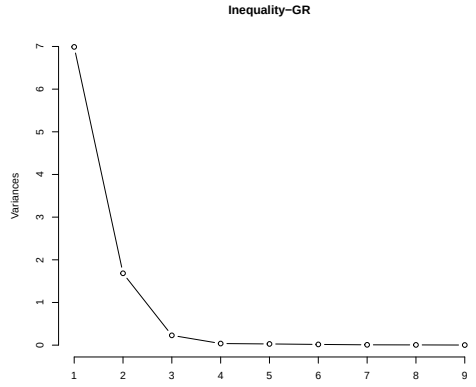

Figure 5: Scree plots for 'Inequality-GR' data set

The Figure 4 and Figure 5 represent the scree plots for the data sets `Inequality_GD` and `Inequality_GR`, respectively. According to the elbow method, the point at which the scree plot flattens out indicates the principal components we should choose. In this case, both scree plots indicate that the first three principal components will be selected. However, Kaiser's Rule suggests selecting only the first two principal components, as they both have variances greater than one for both data sets. Therefore, the first two principal components will be chosen for both data sets.

A biplot can be constructed using the first two principal components, which allows for the comparison of observations to variables. Since 91% of the total variance is explained by the first two principal components together for `Inequality_GD` data set, and 96% of the total variance is explained by the first two principal components together for `Inequality_GR` data set, both of the first two principal components explain the majority(over 90%) of the variation in the two data sets. Therefore, the biplots can be considered accurate visualizations for comparing observations and variables.

A biplot essentially involves plotting the weight vectors on the same scatterplot as the data. There are several things that can be done with a biplot, including observing how the observations relate to one another, how the variables relate to one another, and how the observations relate to the variables. A correlation biplot can be used to compare the variables, a distance biplot can be used to compare the observations, and either a correlation biplot or a distance biplot can be used to explore the relationship between observations and variables.

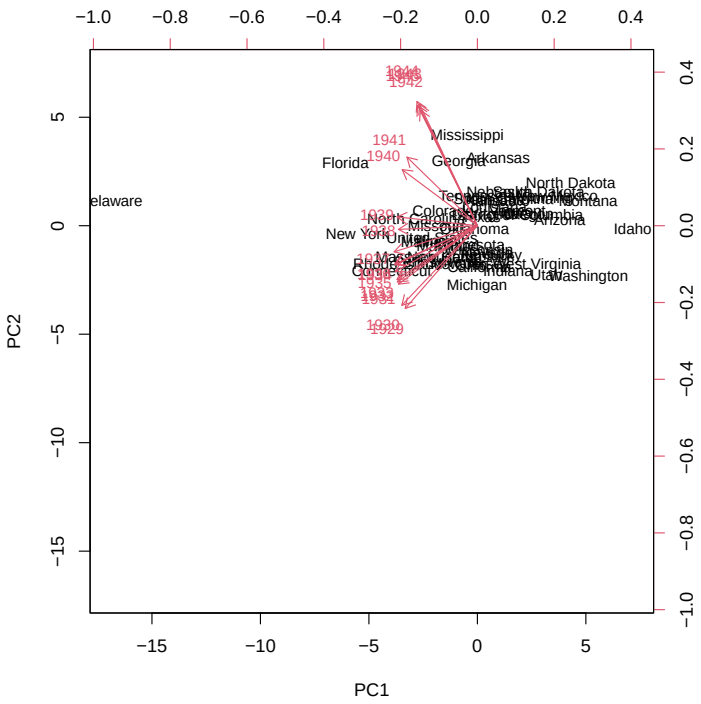

Figure 6: Correlation biplot for the Great Depression and World War II.

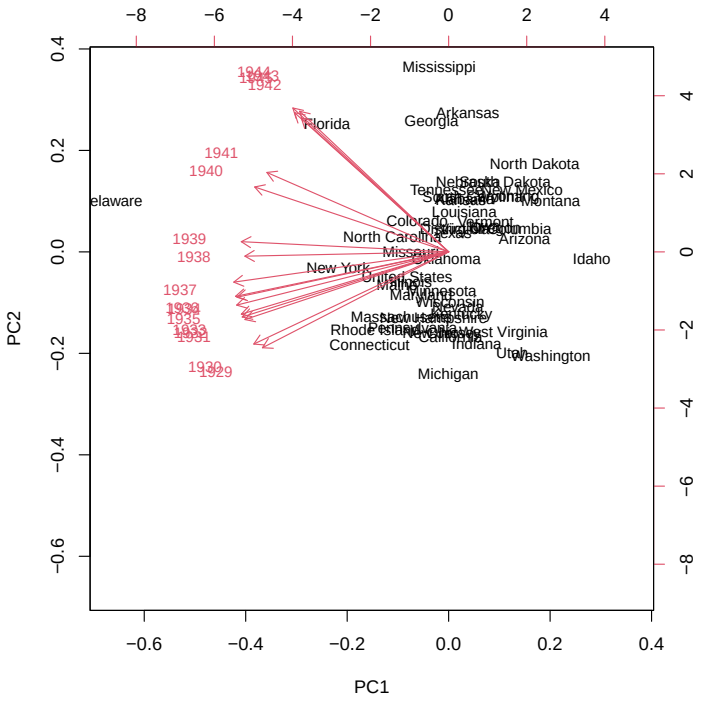

Figure 7: Distance biplot for the Great Depression and World War II.

Figure 6 represents the correlation biplot, while Figure 7 represents the distance biplot for the `Inequality_GD` Gini index data set, which covers the period between 1929 and 1945. This period corresponds to the Great Depression and World War II.

From the correlation biplot in Figure 6, all the yearly variables are more or less negatively associated with PC1, and the yearly variable shifts from a negative association to no association and then to a positive association with PC2. The yearly variable red line shifts clockwise at a small and almost steady pace from 1929 to 1945. We are interested in how the income inequality of the states changes over time, so we first look at consecutive years. The angles between the red lines provide insights into the correlations between two yearly variables. Almost any two consecutive years within the period from year 1929 to 1945 have a small angle, indicating a strong association between the two consecutive years and the Gini index, meaning that the income inequality in the states is quite similar for two consecutive years. The only exception is the angle between 1939 and 1940, where a quite large angle can be observed. This indicates that the income inequality in 1939 is quite different from the income inequality in 1940, which also corresponds to the start of World War II.

Figure 6 is the distance biplot, where the distance between observations implies similarity between observations. In this distance plot, most states are close to each other. The closer they are, the more similar those states are in terms of income inequality within their borders. However, a few states are slightly farther away from the majority, including New York, Florida, Mississippi, and Idaho. This suggests that these states differ significantly in terms of income inequality compared to the majority of states. Delaware stands out as the only state far apart from the others, indicating that Delaware has distinct income inequality compared to the rest of the states in the United States.

Either the correlation biplot in Figure Figure 6 or a distance biplot in Figure 7 can be used to explore the relationship between observations and variables. To achieve this, we need to project the observations perpendicularly onto the year variable line (the red line with an arrow), and the closer an observation is to the red arrow, the higher its value, indicating higher income inequality. Conversely, observations further away from the red arrow have smaller values, meaning lower income inequality.

After projecting the observations onto the year variable line, about half of all states are positioned in the middle of the line in 1929, while the other half are projected to the opposite side of the arrow. This suggests that approximately half of the states have moderate income equality, while the remaining half have low income inequality. During the Great Depression period from 1929 to 1939, more and more states move away from the middle of the red arrow to the opposite side, indicating a decrease in income inequality during the Great Depression. Only New York and Florida appear to move closer to the red arrow, indicating that these two states had higher income equality during the Great Depression.

During World War II, from 1940 to 1945, the majority of states are located on the opposite end of the red arrow, suggesting low income inequality during the war. Only a few states exhibit higher income inequality during the war, including Florida, Mississippi, Georgia, and Arkansas. Interestingly, Delaware is the only state that consistently exhibits a very high level of income inequality throughout the periods of the Great Depression and World War II.

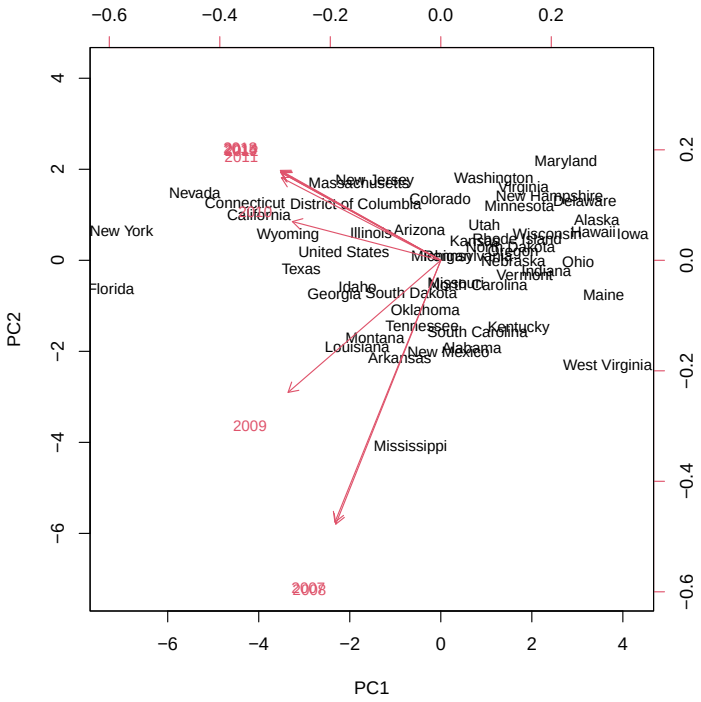

Figure 8: Correlation biplot for the Global Financial Crisis.

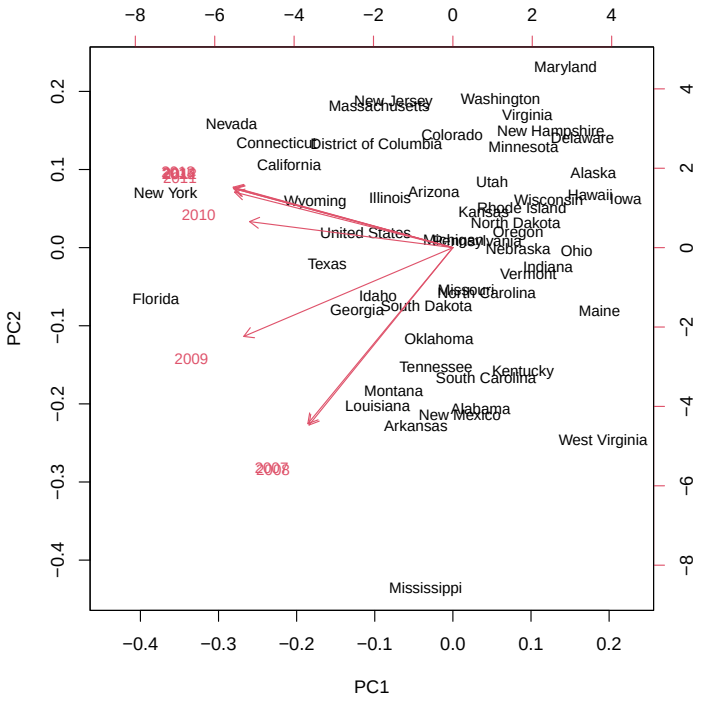

Figure 9: Distance biplot for the Global Financial Crisis.

Figure 8 and Figure 9 represent the correlation biplot and distance biplot, respectively, for the `Inequality_GR` Gini index data set, covering the period between 2007 and 2015. This period corresponds to the Global Financial Crisis and the economic downturn that followed.

In Figure 8, we observe that all the yearly variables are more or less negatively associated with PC1, with the yearly variable shifting from a moderately negative association to a weak positive association with PC2. Our primary interest lies in understanding how income inequality in the states changes over time, so we begin by examining consecutive years.

The yearly variable red lines shift clockwise from 2007 to 2015. The angles between these red lines offer insights into the correlations between two consecutive years. Specifically, the angles between 2007 and 2008, as well as between 2010 and 2015, are all very small. This indicates strong associations between these respective pairs of yearly variables, implying a high correlation in income inequality for those years. Conversely, there are two large angles observed between 2008 and 2010, indicating a weak association between these three years in terms of income inequality.

Figure 8 represents the distance biplot, where the distance between observations implies similarity between observations. The closer observations are, the more similar they are in terms of income inequality over the period, and vice versa. In this distance plot, all the states are spread out, indicating significant differences among them in terms of income inequality within their borders. Furthermore, several states are notably farther away from the majority. These states include New York, Florida, Mississippi, and West Virginia, impling substantial differences in income inequality compared to the majority of states.

Either the correlation biplot in Figure 8 or the distance biplot in Figure 9 can be utilized to explore the relationship between observations and variables. To achieve this, we need to project the observations perpendicularly onto the year variable line (the red line with an arrow). The closer an observation is to the red arrow, the higher its value, indicating greater income inequality. Conversely, observations farther away from the red arrow have smaller values, representing lower income inequality.

After projecting the observations onto the year variable line, the states are distributed along the red line for each year from 2007 to 2015. During the Global Financial Crisis and economic downturn period, some states consistently remain close to the red arrow, while others consistently stay distant from it. This implies that certain states consistently experience higher income inequality, while others maintain relatively lower income levels throughout the period of the Global Financial Crisis and economic downturn.

Although the states with high income inequality fluctuate during this period, they often mirror the states with low income inequality. For instance, Mississippi exhibited high income inequality within the state in 2007 and 2008, but by 2015, it had transitioned to a relatively lower level of income inequality. Conversely, Nevada experienced relatively low income inequality in 2007 and 2008 but shifted to a higher level of income inequality in 2015. Interestingly, Florida appears to consistently maintain a very high level of income inequality throughout this entire period.

6 Limitation

As mentioned in the assumption above, Princial component analysis (PCA) has several limitations. For example, PCA relies on assumptions including linearity and orthogonality of the components. If any of these assumptions do not hold, the accuracy of PCA can be affected. Furthermore, in this report, we only use the first and second principal components for the analysis, which means that other components will not be used, the information contained by those components will be lost, even just by a samll amount. Additionally, the interpretability of Principal Component Analysis is also limited, as the principal components are linear combinations of the variables. The loadings, the coefficients in the components, have no meaning in the context of the original data. Therefore, there are several limitations to Principal Component Analysis.

7 Conclusion

In conclusion, historical events such as the Great Depression, World War II, and the Global Financial Crisis have all had significant impacts on income inequality in the states of United States. The impacts of the Great Depression and World War II on income inequality appear different from the impact of the Global Financial Crisis. During the Great Depression and World War II, the majority of states in the United States experienced a reduction in income inequality, resulting in relatively lower levels of income inequality. Only a very few states exhibited high levels of income inequality at the end of this period. In contrast, the Global Financial Crisis also influenced the level of income inequality in the states. This historical event often mirrored states with low levels of income inequality, shifting them to high levels of income inequality, and vice versa. Consequently, some states with low levels of income inequality experienced an increase in income inequality, while some states with high levels of income inequality have a relatively lower level of income inequality following this historical event.

8 Reference

Citation of R package

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi:10.21105/joss.00355 https://doi.org/10.21105/joss.00355, http://dx.doi.org/10.21105/joss.00355.

Tierney N, Cook D (2023). "Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations." *Journal of Statistical Software*, *105*(7), 1-31. doi:10.18637/jss.v105.i07 https://doi.org/10.18637/jss.v105.i07.

Zhu H (2023). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.

Auguie B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3, https://CRAN.R-project.org/package=gridExtra.

9 Appendix: All code for this report

```
knitr::opts_chunk$set(
  echo = FALSE,
  eval = TRUE,
  message = FALSE,
  warning = FALSE,
  error = FALSE,
  out.width = "70%",
  fig.align = "center",
  fig.width = 8,
  fig.height = 7,
  fig.retina = 3,
  fig.pos = "H",
  out.extra = "")
# Use out.extra to apply CSS styles
# out.extra='style="background-color: #9ecff7; padding:10px; display: inline-block;"'
# Load library and data sets
library(tidyverse)
library(visdat)
library(naniar)
library(kableExtra) # if knit error with kabelExtra, reinstall package: devtools::install_github("kupie
library(gridExtra)
data_gd <- read_csv("Inequality_GD.csv")
data_gr <- read_csv("Inequality_GR.csv")
# Check data sets
data_gd %>%
  head() %>%
  kable(caption = "First six rows of Inequality-GD data set")
# Check data sets
data_gr %>%
  head() %>%
  kable(caption = "First six rows of Inequality-GR data set")
# Remove index column
new_gd <- data_gd %>%
  select(-"...1")

new_gr <- data_gr %>%
  select(-"...1")

# Pivot data sets
pivot_gd <- new_gd %>%
  pivot_longer(cols = -c(State), names_to = "Year", values_to = "Gini_index") %>%
  mutate(Year = as.factor(Year))

pivot_gr <- new_gr %>%
  pivot_longer(cols = -c(State), names_to = "Year", values_to = "Gini_index") %>%
  mutate(Year = as.factor(Year))

# Use summary statistics to check Gini index data in Inequality_GD and Inequality_GR
t1 <- pivot_gd %>%
  select(Gini_index) %>%
  summary()
```

```r
t2 <- pivot_gr %>%
  select(Gini_index) %>%
  summary()

kable(list(t1, t2), caption = "Summary statistics for the Gini index data sets.")
# Plot of missing value
miss1 <- gg_miss_var(data_gd)
miss2 <- gg_miss_var(data_gr)

grid.arrange(miss1, miss2, nrow = 1)
# Use boxplots to show distributions of data
pivot_p1 <- pivot_gd %>%
  ggplot(aes(x = Year,
             y = Gini_index,
             group = Year)) +
  geom_boxplot(na.rm = TRUE) +
  theme_classic() +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, vjust = 0.5, hjust= 0.3))

pivot_p2 <- pivot_gr %>%
  ggplot(aes(x = Year,
             y = Gini_index,
             group = Year)) +
  geom_boxplot(na.rm = TRUE) +
  theme_classic() +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, vjust = 0.5, hjust= 0.3))

grid.arrange(pivot_p1, pivot_p2, nrow = 1)
# Get information of state & year of the extremely large value
pivot_gd %>%
  filter(Gini_index >1) %>%
  kable(caption = "An extremely large Gini index") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
# Remove the extremely large value and missing values
new_gd[new_gd == 1000] <- NA

gd <- new_gd %>%
  mutate(`1934` = impute_mean(`1934`))
gr <- new_gr %>%
  mutate(`2010` = impute_mean(`2010`))
table1 <- gd %>%
  pivot_longer(cols = -c(State), names_to = "Year", values_to = "Gini_index") %>%
  mutate(Year = as.factor(Year)) %>%
  select(Gini_index) %>%
  summary()

table2 <- gr %>%
  pivot_longer(cols = -c(State), names_to = "Year", values_to = "Gini_index") %>%
  mutate(Year = as.factor(Year)) %>%
  select(Gini_index) %>%
  summary()

kable(list(table1, table2), caption = "First six rows of Inequality-GR data set") %>%
```

```
  kableExtra::kable_styling(latex_options = "hold_position")
clean_gd <- gd %>%
  pivot_longer(cols = -c(State), names_to = "Year", values_to = "Gini_index") %>%
  mutate(Year = as.factor(Year)) %>%
  ggplot(aes(x = Year,
             y = Gini_index,
             group = Year)) +
  geom_boxplot(na.rm = TRUE) +
  theme_classic() +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, vjust = 0.5, hjust= 0.3)) +
  ylim(0.3, 0.8)

clean_gr <- gr %>%
  pivot_longer(cols = -c(State), names_to = "Year", values_to = "Gini_index") %>%
  mutate(Year = as.factor(Year)) %>%
  ggplot(aes(x = Year,
             y = Gini_index,
             group = Year)) +
  geom_boxplot(na.rm = TRUE) +
  theme_classic() +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, vjust = 0.5, hjust= 0.3)) +
  ylim(0.3, 0.8)

grid.arrange(clean_gd, clean_gr, nrow = 1)
# Convert State column to row names
gd <- gd %>%
  column_to_rownames("State")

gr <- gr %>%
  column_to_rownames("State")
# Pipe data into prcomp()
pca_gd <- gd %>%
  prcomp(scale. = TRUE)

pca_gr <- gr %>%
  prcomp(scale. = TRUE)
# To inspect std, (cumulative) proportion variaance
summary(pca_gd)

summary(pca_gr)
scree1 <- screeplot(pca_gd, type = "l", main = ggtitle("Inequality-GD"))
scree2 <- screeplot(pca_gr, type = "l", main = ggtitle("Inequality-GR"))
biplot(pca_gd, scale = 0, cex = 0.8)
biplot(pca_gd, cex = 0.8)
biplot(pca_gr, scale = 0, cex = 0.8)
biplot(pca_gr, cex = 0.8)
```