MATH 271.1: Statistical Methods

Group Project 3 on Principal Components Analysis
Instructor: Jakov Ivan S. Dumbrique
First Semester 2021-2022

For this group project, you will work in pairs or trios and apply principal components analysis (PCA) to analyze and visualize the world development indicators of countries. The deadline for the submission of your project is on **December 18, 2021**.

1 Overview of the WDI Dataset

The World Bank (http://www.worldbank.org/) annually collects global development indicators of countries and economies from officially-recognized international sources, in an effort to inform the world leaders, businessmen, economists, and the general public of the most current and accurate global development data available including national, regional, and global estimates. You can find the World Development Indicators (WDI) dataset and more information at this website:

https://datacatalog.worldbank.org/dataset/world-development-indicators

Download the Excel file in zip format (around 67.1 MB in size). The unzipped Excel file contains annual information on 1,443 world development indicators of 265 countries and economies from 1960 to 2020. You will be getting your dataset from this file. The Excel file contains the following six sheets:

1. Data: contains the annual data on world development indicators of 265 countries and economies. This will be the Excel sheet you will be primarily working on.
2. Country: contains the description of each country and economy. This tab also contains the classification of each country according to its *geographical region, income group, lending group, system of trade, government accounting concept*, and other categories.
3. Series: contains the description of each of the 1,443 world development indicators (WDIs). Note that a country may not have data for a particular WDI.
4. Country-Series: contains the description and sources for selected WDIs for each country.
5. Series-Time: contains the description and details on data interpolation of WDIs for specific years.
6. FootNote: contains additional information for specific values of WDIs of countries on selected years.

2 Instructions for the Project

1. **Choosing your Data of Interest**
   At the end of the day, this is your *own* project, so it would be good to have the freedom to choose the countries, WDIs and the timeframe you want to study. It is best to start with a research question you want to investigate.

   (a) *Countries of Interest*
   The WDI dataset encompasses development data across 265 countries and economies. You have to choose which **countries** (*not economies*) you want to analyze. This may

be based on geographical region, income group, First/Second/Third World classification, or it may just be a random collection of countries (if so, please provide justification for your choice). You must have **at least twenty (20) countries** in your final list. You must include in the R Notebook the reasons behind your selection of countries.

(b) *WDIs of Interest*
You also have to choose the WDIs you want to analyze for the countries. Your final list of WDIs should be a good balance of the following areas/topics concerned (check the Series sheet of the Excel file):

i. Economic Policy & Debt
ii. Education
iii. Environment
iv. Financial Sector
v. Gender
vi. Health
vii. Infrastructure
viii. Poverty
ix. Private Sector & Trade
x. Public Sector
xi. Social Protection & Labor

(c) *Timeframe of Interest*
Choose a specific timeframe you want to study. This may be a year (e.g. 2020) or a range of years (e.g. 2008-2019). If your group decides to choose a year, then you'll be working on the column of actual values of the WDIs for that year. Otherwise, if your group chooses a range of years, then you need to come up with a new column of values for each WDI of every country that best represent the actual data for your chosen timeframe. This may require you to use measures of central tendency (i.e. mean, median, mode, weighted mean) to generate those values. You must include in your R Notebook the reasons behind your selection of timeframe, and the choice of your measure of central tendency, if applicable.

(d) *Expected output for this part*
It is expected by the end of this part that you should have a column of WDI values for your countries of interest on your selected timeframe.

2. **Transforming your Data**
Transform your single column of values into a $n \times p$ data matrix, where n is the number of countries and p is the number of world development indicators. You may use any software (e.g. Excel, R) to reformat your data.[1] How one should treat missing values in a dataset is generally a difficult, and often domain-specific and subjective, problem. For missing values in your data matrix, you may use techniques discussed in class on how to handle them, as long as you justify why you chose to implement a particular technique. You might want to read the answers on this online forum. You may also use other methods–you just need to specify them in your R Notebook.

3. **Principal Components Analysis**
Perform Principal Components Analysis on your data, after centering and scaling it. Visualize the data by generating biplots, graphs of variables and observations, correlation plots, and bar graphs. Some guide questions on the analysis of results:

[1] For those who want to transform their data using R, you may want to check one of our references: Wickham H. and G. Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media (2017).

- Quantify the proportion of variance explained by each principal component. Display these PVEs using a scree plot.
- If we use principal components as a summary for the data, how many components are sufficient? Justify your answer.
- Analyze your visualizations (e.g. biplots, graphs of variables and observations). You may use classifications found in the Country tab or other reliable sources to label the countries in the plots and generate insights based on these graphs.
- What do the PC loading vectors represent?
- Based on the graph of variables, are there WDIs that are positively/negatively correlated? Justify your claims.
- Assess the quality of representation of WDIs on the graph of variables and their contribution to the construction of the PC loading vectors.
- What insights on the countries can you generate from the graph of observations? You are encouraged to find related literature to support your claims.
- Analyze the importance of the principal components for a given country, and the contribution of each country in constructing the PCs.

4. **Extending your Project** (this is optional)
   You may opt to extend the scope of your project by performing PCA on

   (a) the same set of countries but on a different timeframe, or
   (b) the same timeframe but on a different set of countries.

   You can then compare your PCA results across time (e.g. pre-pandemic vs. during pandemic) and/or across different groups of countries (e.g. low-income vs high-income groups).

3 Deliverables

There are two required deliverables for this Project:

1. R Notebook

   - You are not required to come up with a separate typeset report, so please make sure that your R Notebook is **organized, well-documented and clearly-annotated**.
   - your R Notebook should at least contain the following:
     (a) Project Title and Student Names
     (b) Abstract of the Project
     (c) Research Question/s
     (d) Motivation for Choice of Dataset (countries, WDIs, and timeframe of interest)
     (e) details on transformation of data and how you dealt with missing values
     (f) PCA code
     (g) Analysis of PCA Results
     (h) Conclusions and Recommendations
     (i) References
   - You may use any citation format, but make sure to follow in-text citations.
   - Please submit both the ***.Rmd and *.nb.html** files.

2. final dataset used as input to PCA

  - preferably in csv format

Aside from these required deliverables, kindly submit as well other pertinent files you used in implementing your project, if any (e.g. Excel files for data preprocessing).

4 E-mail Submission Mechanics

- Compress all pertinent attachments in one zip file.
- If the Excel file/s are too large to be contained in the email, a download link for the file/s will suffice.
- The filename of the R Notebook *and* the zip file should be of the following format:

  MATH271_PCA_[SurnamesofMembersInAlphabeticalOrder]

  For the e-mail submission, use the filename as the subject line of the e-mail.
- All e-mail submissions must be sent to jdumbrique@ateneo.edu on or before **December 18, 2021 11:59 PM**.