Multivariate Analysis with Pew Research Data

[redacted]

Northwestern University, Unsupervised Learning Methods

January 30th, 2022

Abstract

To gain and maintain the support of the public, politicians and their political scientists are looking to get a better understanding of the views and preferences of the voter population. Generalized views around political affiliation do not provide enough information on topics relevant to the voter, nor do they help generate political strategies for parties trying to create an advantage in an upcoming election. Both pollsters and politicians use information from surveys, such as the Pew research data, to help home in on the most important topics to voters and to predict which campaigns are most likely to generate an election victory. One important consideration is the many subgroups within the population and the unique opinions and preferences that they represent. This paper presents three unsupervised learning methods that are designed to help identify patterns, simply or reduce data and/or help unlock insights that are unattainable through basic exploratory data analysis. These techniques include Principal Component Analysis (PCA), Factor Analysis (FA) and Multidimensional scaling (MDS). Among

the methods tested, MDS showcased valuable insights where it was possible to label the niche subgroups found on the map.

Keywords: PCA, FA, MDS

I. Introduction

Especially during electoral times, a lot of research is done to understand the needs and opinions of the voters. As consultants for a politician who is running for office, the candidate would like us to supply them with information on where the population stands on various issues. More specifically, the politician heard of a Pew research survey and would like more detail on their political leanings based on the various questions that were asked. The politician would also like to discover underlying characteristics or beliefs of people of particular political leanings. The belief is that being able to break up the overall population into smaller subgroups and working to relate to each of these groups specifically will provide a better chance of winning than creating a generic campaign for the entire population.

II. Literature Review

The question becomes: how can we find the unique characteristics of these subgroups? Scholarly studies in political science, marketing, and psychology regularly use unsupervised learning methods to identify the underlying characteristics that are common among the data studied. Some methods include: PCA, MDS, and FA. There is no universal agreement on what method works best. Some assert that FA is better suited for isolating the deep-rooted factors common to a group of people versus PCA (Watkins 2018). Others contend that multiple methods used together gets the best result. In a study about marketing politics, PCA was performed on the data

prior to running FA to improve correlation among the variables, which in turn led to higher eigenvalues for the FA (Anim et al 2019). Other researchers take a different approach, using MDS to visualize one component, while using FA to identify underlying characteristics of another aspect of the study (Banerjee 2019). What can be concluded is that there is no single way to identify underlying characteristics present in the data, and it is up to the researcher to determine the appropriate method.

III. Methods: Research Design, Modeling Methods + Implementation and Programming

The analysis was conducted on data from a March 2019 Pew survey. The questions from the survey are political and demographic in nature and have been analyzed using unsupervised learning techniques to discover hidden patterns. Of the 73 question variables available, many had missing values and were omitted for the analysis conducted. Thus, the work was done on 30 multi-category variables that were expanded to 100 binary indicator variables and analyzed using the aforementioned techniques. Data imputation is always an option for missing values, but due to the complexity of the subject matter and the potential to introduce bias (politics seem to be seldom straightforward and easily categorized), it was decided to use only the complete data.

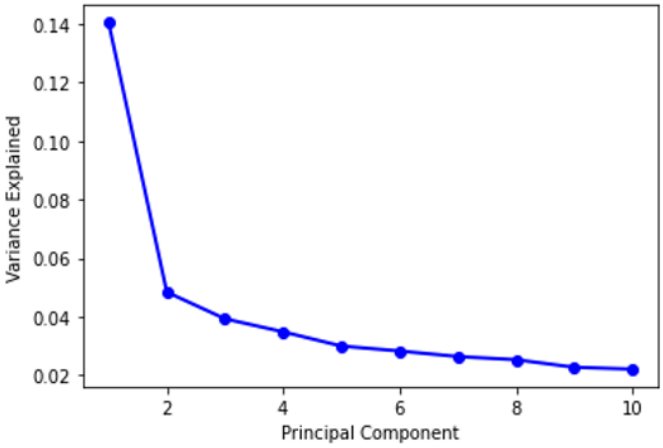
In terms of implementation, we used the programming language Python, and these accompanying packages: sklearn, numpy, pandas, pyreadstat, factor_analyzer, matplotlib, seaborn, and scipy.

In terms of the modeling method, we tested and evaluated three: PCA, FA, and MDS. Ultimately, we chose MDS to represent our data because, while it was subjective to interpret, it gave us the most information, flexibility, and nuance for grouping the variables and forming

underlying characteristics. The below paragraphs describe how we chose the dimensions for each method.

Principal Components Analysis

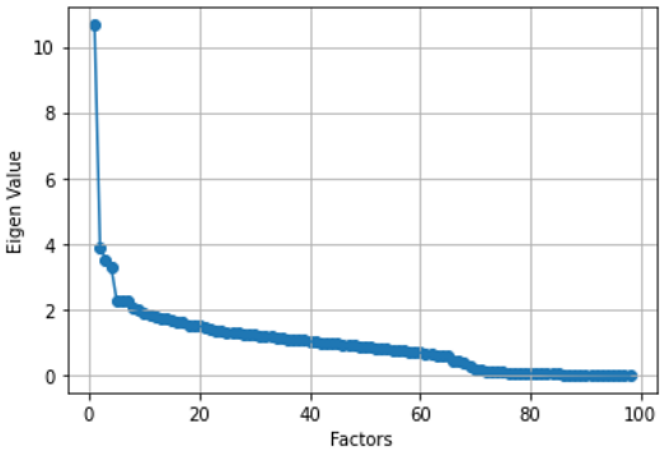
Figure 1: PCA Scree Plot



Our goal for PCA was to reduce the number of data dimensions through the discovery of hidden relationships among the 30 variables chosen from the Pew data. *Figure 1: PCA Scree Plot* describes the incremental variance explained by adding additional principal components. The data can be reduced to approximately 2 principal components since this is where the variation explained decreases to a far more gradual slope on the Scree Plot.

Factor Analysis

Figure 2: FA Scree Plot



Similar to PCA, our goal for FA was to reduce the number of data dimensions through the discovery of hidden relationships. Like the scree plot above, we look for where the slope of the line is reduced to a more gradual level to determine the optimal number of factors for data reduction. We determined that five factors were appropriate for our analysis based on our interpretation of the elbow in *Figure 2: FA Scree Plot*.

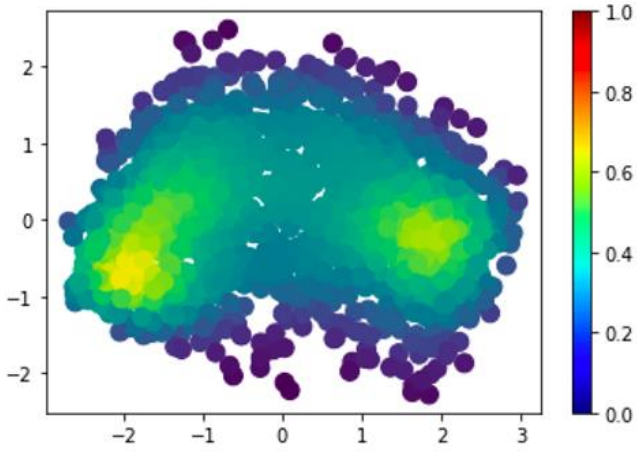
Multidimensional Scaling

Multidimensional Scaling was also tested as a dimension reduction technique and used to better understand the patterns in responses to multiple political questions. A multidimensional scaling map was created to display the relative positions of responses to each other to help visually ascertain response groupings.

IV. Results

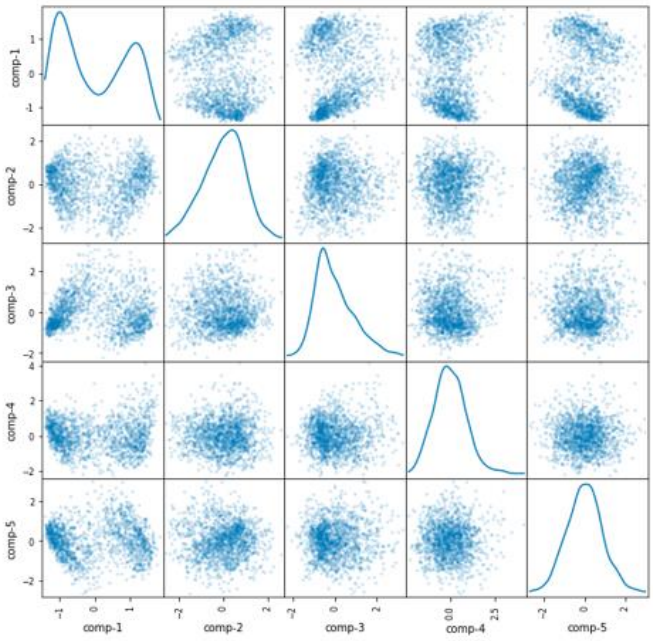
Here we consider the results of all methods studied and justify why we chose MDS as the best option for identifying underlying characteristics of the data. PCA is a promising method that has been used in many studies. However, in using PCA, two principal components explained only 14% of variation of the dataset. This very low result suggests a weak model and opportunities for gaining additional insights. In order to gain some intuition on how the two PCs behave, a scatterplot with density information was generated. *Figure 3: PCA Density Plot* allows us to see the center of each of the components, but it is unclear where one starts and the other ends. The results were not conclusive and thus other methods were investigated.

Figure 3: PCA Density Plot



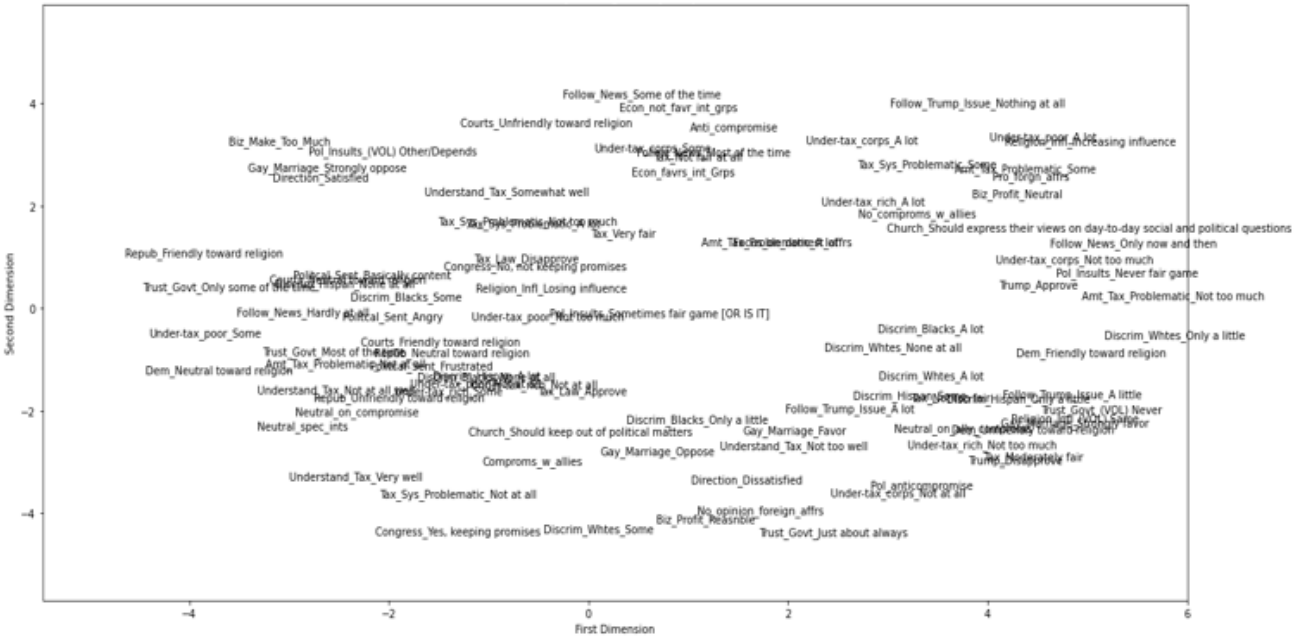
Factor Analysis seemed to be an ideal candidate for this task because the results are far more interpretable than PCA. FA resulted in a reduction of data dimensions to five factors which explained 20.7% of the data spread. While better than PCA, this seemed quite low at first glance. To further observe the relationship between these factors, *Figure 4: Scatterplot between Factors* was created.

Figure 4: Scatterplot between Factors



At first glance, it seems that the only relationship in the data is between factor 1 and the rest of the factors, with the prominent factors being 1 and 5. More troubling was that when we discovered exactly what variables were grouped in factor 1, we realized that the variables contradicted each other. For example, factor 1 contained the variables representing both "satisfied" and "dissatisfied" respondents on the topic related to satisfaction in the direction of the country. This is problematic because it left us wondering how exactly to tease out underlying groups when the data contradicted itself. Lastly, it seemed that the data was disorganized. While factor 5 contains questions on federal tax systems, factor 5 is varied on presidency, economic system, and taxes. Due to the contradiction and confusion, we thought it was best to try another method that can provide better visual representation between the variables.

Figure 5: Dissimilarity Among Survey Respondents



To showcase the actual variables against each other, we generated a dissimilarity plot (see *Figure 5: Dissimilarity Among Survey Respondents*) that displays the positions of each variable in a 2-dimensional space. This provided the most amount of information for us to interpret as compared to the other two methods. There is an apparent cluster that is formed on the right end of the map. A lot of the responses were in relation to race, the trust of government, and the following of misconduct allegations during Trump's presidency. It seems that race, lack of trust for the government and disapproval of Trump's presidency are strongly related. This area can be labeled as something along the lines of "Unfair Discrimination Leads to Distrust in Government". Another grouping can be seen towards the middle left, where religion has taken up a lot of the variables. Many of the variables are positive towards the religion questions, and they interloop religion with the government. This is interesting because many of the variables indicate that the government is easy going towards religion, and if so, those respondents trust the government. By observing closely, it is also in this grouping that these respondents also feel that

there are no problems with tax amounts. It appears that by having a popular religion in government that is followed by many, denotes a positive relationship with government actions and generates trust. Overall, this group can be labeled as “Religious Influence on Government Trust”. With this interpretation of MDS, we have reduced the data to two underlying characteristics. We now turn to our conclusion and apply our findings to help the politician understand the electorate.

V. Conclusion

The purpose of conducting the various methods was to understand which variables showcased prominent relationships. The method outcomes consisted of these relevant variables that can later be used in feature engineering. These variables can be transformed into significant variables to predict an outcome variable, which can aid in understanding the political leanings of a population. Based on the analysis in this study, it can be recommended to a politician to campaign towards religion. They should also be mindful of the discrimination towards various races in order to gain the trust of minorities. In doing so, a politician could gain support from various voter bases. Additional analysis using methodologies such as K-Means clustering could help to unlock additional insights.

VI. References

Anim, Patrick Amfo, Frederick Okyere Asiedu, Matilda Adams, George Acheampong, and Ernestina Boakye. 2019. "'Mind the Gap': To Succeed in Marketing Politics, Think of Social Media Innovation." Journal of Consumer Marketing 36 (6): 806–17. https://doi.org/10.1108/jcm-10-2017-2409.

Banerjee, Sougata, and Mohsin Aziz Baba. 2019. Review of Positioning of Vishal Mega Mart, a Hypermarket and Its Consumer Preferences through the Implementation of Multi Dimensional Scaling, Factor and Conjoint Analysis W.r.t. Delhi Market. *International Journal of Marketing and Business Communication* 8 (1): 25–37. https://www.proquest.com/docview/2297130249?accountid=12861&parentSessionId=aYoNc9Gq1LmJM5szR%2Fj3QbfV2igspoB2DoY%2B%2BN0UlBc%3D&pq-origsite=primo.

Cochrane, Christopher. 2010. "Left/Right Ideology and Canadian Politics." Canadian Journal of Political Science 43 (3): 583–605. https://doi.org/10.1017/s0008423910000624.

Watkins, Marley W. 2018. "Exploratory Factor Analysis: A Guide to Best Practice." Journal of Black Psychology 44 (3): 219–46. https://doi.org/10.1177/0095798418771807.

León-Borges, José A., Noh-Balam, Roger-Ismael, Rangel Gómez, Lino, Philip Strand, Michael. 2015. “Machine Learning in the Prediction of Elections.” ReCIBE. Revista electrónica de Computación, Informática, Biomédica y Electrónica. http://www.redalyc.org/articulo.oa?id=512251502001