Unsupervised Learning on the Health and Retirement Study using Geometric Data Analysis

Reinaldo Sanchez-Arias[1], Roberto Williams Batista[1]

[1] Department of Data Science and Business Analytics, Florida Polytechnic University

Introduction

The main focus of this work is to show the ability of geometric data analysis techniques in discovering response patterns in survey data where the majority of measurements result in categorical variables.

Methods

The geometric data analysis method of Multiple Correspondence Analysis (MCA) allows the construction of a lower dimensional space that captures the variance in the original data, and in which both variables and individuals can be projected to explore patterns, validate hypotheses, and better understand the association among the observed data.

MCA is an unsupervised learning algorithm under the framework of Geometric Data Analysis (GDA), in which the elements of two sets indexing the entries of the data table become points in a geometric space and define two clouds of points: a cloud of categories and a cloud of individuals (Fig. 1). The distance between individual points is a reflection of the dissimilarity between response patterns of individuals, and both resulting clouds are on the same distance scale (B. Le Roux/2010).

The traditional data format for MCA is an Individuals $\times$ Questions rectangular table, where questions are categorical

variables with a finite number of categories (also called levels), and for each question, each individual chooses one and only one response category. Categories may be qualitative (nominal or ordinal) or may result from the splitting of continuous variables into categories.

MCA can be seen as a particular case of weighted principal component analysis, in which a set of multidimensional points exists in a high-dimensional space where distance is measured by a weighted Euclidean metric and the points themselves have differential masses. A lower dimensional solution is obtained by determining the closest plane to the points in terms of weighted least-squared distance, and then projecting the points onto the plane for visualization and interpretation. The lowdimensional subspace that fits the points as closely as possible can be obtained compactly and neatly using the generalized singular-value decomposition (SVD) of the data matrix.

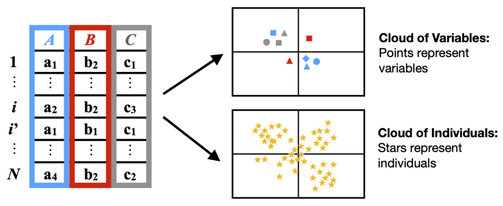

Figure 1: MCA idea

The squared distance between two respondents is calculated using the variables for which each had chosen different categories.

$$d^2(i,i') = \frac{1}{Q}\sum_{k\in K}\frac{(\delta_{ik}-\delta_{i'k})^2}{f_k}$$

where $\delta_{ik}=1$ if i has chosen k and 0 otherwise. Notice that the smaller the frequencies of disagreement categories, the greater the distance between individuals. The set of all distances between individuals determines the cloud of individuals consisting of N points in a space with dimensionality $L \leq K-Q$ (it is assumed here that $N > L$). Additionally, if respondent i chooses infrequent categories, then the point M^i representing individual i is far from the mean center of the cloud G. The squared distance from point M^i to G is given by

$$(GM^i)^2 = \left(\frac{1}{Q}\sum_{k\in K}\frac{\delta_{ik}}{f_k}\right) - 1$$

In the cloud of categories, a weighted cloud of K points, category k is denoted by point M^k with weight n_k: For each question, the sum of the weights of category points is N, and the relative weight p_k of point M^k is simply $p_k = f_k = Q$.

Given two categories k and k', the squared distance between the points M^k and $M^{k'}$ is calculated as

$$(M^k, M^{k'})^2 = \frac{n_k + n_{k'} - 2n_{kk'}}{n_k n_{k'}/N}$$

with $n_{kk'}$ denoting the number of respondents who have chosen both categories k and k'.

The contribution of a category point M^k to the overall variance is the ratio of the amount of the variance of the cloud due to category k. The contribution of a question q is the sum of the contributions of its categories. Contributions can be calculated as shown below:

$$\mathrm{Ctr}_k = \frac{1-f_k}{K-Q}, \qquad \mathrm{Ctr}_q = \frac{K_q - 1}{K-Q}$$

The HRS Dataset

Created in 1990 and launched in 1992 by the National Institute on Aging (NIA) and Social Security Administration, the Health and Retirement Study (HRS) surveys collect every two years of data from more than 22,000 Americans over 50 years old. It is the first longitudinal study of Americans approaching the economic and health aspects in the same survey and being the largest nationally representative multidisciplinary panel study of Americans aged 50 and older. The study was created and maintained by the Institute for Social Research (ISR) Survey Research Center (SRC) at the University of Michigan.

Results and Discussion

MCA was performed on a combined dataset from respondents of the 2008 and 2010 waves. Notice that the participants of the 2008 survey are different than those from the 2010 survey. The clouds patterns for every wave were examined to confirm that the overall geometric representations were similar regardless of the number of participants in each wave, or the year in which the survey responses were collected.

You can reference tables like so: Table ??.

`NEED TO ADD TABLES TO POSTER`

Clustering

Geometric data analysis methods have the potential to be used as a pre-processing step for clustering, given the representation in a lower dimensional space provided by the principal component technique of choice. In this work, a hierarchical clustering algorithm is performed using the coordinates of each respondent in the lower dimensional space generated by the MCA procedure.

The findings of this hierarchical clustering confirm a natural grouping for the participants of the survey: the tendency of survey respondent to use the levels of agreement with the different questions that are part of the questionnaire, namely, "a lot", "not at all", "some" and "a little". These levels of agreements are well separated in distinct regions within the plane of the first 2 principal dimensions.

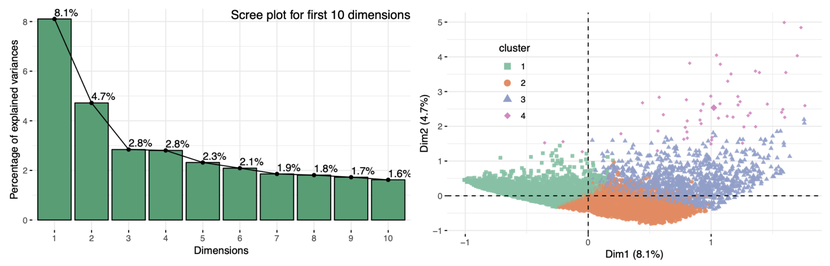

Figure 2: Hierarchical Clustering

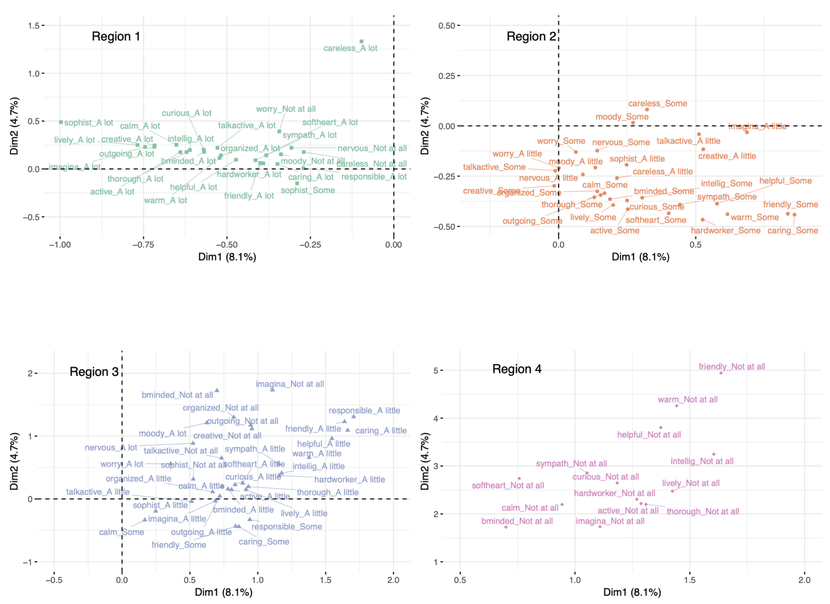

Figure 3: Regions

Conclusions

The use of unsupervised techniques presented in this work represents an opportunity to extract valuable insights from longitudinal datasets like the one made available by the US Health and Retirement Study. MCA allows for new interpretations and discovery of patterns that take advantage of the qualitative nature of the data collected from survey respondents. The hierarchical clustering technique applied to the low dimensional representation of participants, provided by the MCA method, suggested a reasonable separation of the respondent profile as characterized by a personality scale. Results provided by this approach may be used to explore other areas that have yet to be captured using the items in the questionnaires, helping in the design of the survey and sampling procedure, and allowing for correlation studies with other physical and mental health indicators.

Praesent dictum mauris at diam maximus maximus (Thorne 2019). `FactoMineR`, `tidyverse`

Acknowledgements

The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. The HRS has been approved by the Institutional Review Board at the University of Michigan. The HRS obtains in- formed verbal consent from voluntary participants and follows strict procedures to protect study participants from disclosure (including maintaining a Federal Certificate of Confidentiality). The public data, made available to registered researchers and used in this study, is de-identified.

References

Thorne, Brent. 2019. *Posterdown: Generate Pdf Conference Posters Using R Markdown*. https://CRAN.R-project.org/package=posterdown.