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. A lower dimensional representation of both individuals and measured variables is used to detect and represent underlying structures in the US Health and Retirement Study, a longitudinal survey of a representative sample of Americans over age 50 that captures information on how changing health interacts with social, economic, and psychological factors and retirement decisions.

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* (Figure 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 (Le Roux and Rouanet 2010).

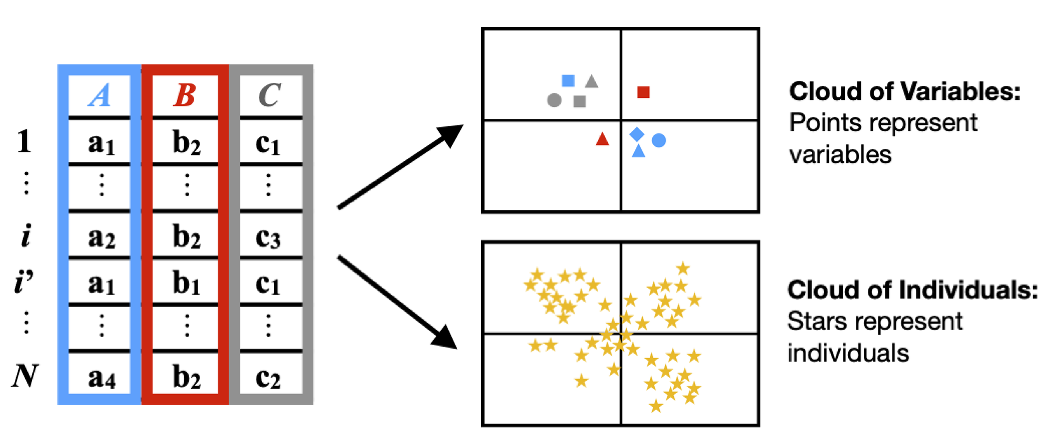

Figure 1: MCA idea: representation of participants and responses to survey

The lower dimensional representation 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 solution can be obtained compactly and neatly using the generalized singular value decomposition (SVD) of the data matrix (Greenacre and Blasius 2006).

Let $\mathcal{I}$ be the set of N individuals and $\mathcal{Q}$ the set of questions, encoded in Q variables. The data used in the MCA approach is a $N \times Q$ matrix such that entry (i, q) is the response category of the question q chosen by individual i. Let n_k be the number of respondents who chose category k, and $f_k = \frac{n_k}{N}$ the relative frequency of respondents who chose category k. 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, and f_k is the relative frequency of respondents who chose category k. 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 $L \leq K - Q$ (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. In the cloud of categories, a weighted cloud of K points (where K denotes the overall set of categories), category k is denoted by point M^k with weight n_k.

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 data from more than 22,000 Americans over 50 years old every two years. The study was created and maintained by the Institute for Social Research (ISR) Survey Research Center (SRC) at the University of Michigan. This work uses the following HRS data products: HRS Core Cognition Section (D), HRS Left-Behind Questionnaires Section LB, and the RAND HRS Longitudinal File V2. All the data used in this work was related to the survey waves of 2006, 2008, 2010 and 2012.

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.

Table 1: Coordinates of the first 4 dimensions for the top 12 categories (sorted by contribution and level of agreement)

Variable	Dimension 1	Dimension 2	Dimension 3	Dimension 4
imaginative A lot	-0.75	0.23	-0.20	0.04
creative A lot	-0.72	0.25	-0.23	0.03
talkactive A lot	-0.53	0.22	-0.03	-0.35
caring A little	1.66	1.09	-1.90	0.58
friendly A little	1.64	1.23	-1.73	0.19
responsible A little	1.71	1.31	-1.53	-0.18
careless Some	0.32	0.08	0.10	-0.68
responsible Some	0.94	-0.33	0.07	-0.06
moody Some	0.27	0.02	0.07	-0.70
nervous Not at all	-0.31	0.22	-0.19	0.81
worry Not at all	-0.34	0.39	-0.27	1.11
sophisticated Not at all	0.52	0.32	0.03	-0.14

Large coordinate measures suggest that the categories of a variable are better separated along that dimension, while similar coordinate measures for different variables in the same dimensions indicate that these variables are related to each other. Variance rates are calculated as follows: for $l = 1, 2, \ldots, l_{\max}$ such that $\lambda_l > \bar{\lambda}$

1. calculate the pseudo-eigenvalue $\lambda' = \left(\frac{Q}{Q-1}\right)^2 (\lambda_l - \bar{\lambda})^2$,
2. compute the sum $S = \sum_{l=1}^{l_{\max}} \lambda'_l$

Then for $l < l_{\max}$ the modified rates are equal to $\tau'_l = \lambda'_l / S$

Table 2: Variances of Axes and Modified Variance Rates

Axes	1	2	3	4	5	6
Eigenvalue (λ_l)	0.244	0.142	0.086	0.085	0.070	0.063
Modified variance rate (τ_l)	0.688	0.180	0.039	0.038	0.019	0.012

Notice: The first two dimensions explain about 86.8% of the variance in the data.

The MCA algorithm used in this study corresponds to the implementation of the method available in the R package `FactoMineR` (Lê, Josse, and Husson 2008).

Table 3: Coordinates of the supplementary variables (not used to define distance between individuals)

Category	Dimension 1	Dimension 2	Dimension 3	Dimension 4
Female	-0.11	0.04	0.06	-0.18
Male	0.17	-0.05	-0.09	0.28
non-veteran	-0.05	0.03	0.03	-0.09
veteran	0.17	-0.10	-0.09	0.32
hispanic	0.18	0.19	0.20	-0.21
non-hispanic	-0.01	-0.02	-0.02	0.02
black	-0.06	0.34	-0.05	0.08
other	0.00	0.18	0.03	-0.10
white	0.01	-0.05	0.00	-0.00
Depressed No	-0.05	-0.05	-0.01	0.09
Depressed Yes	0.43	0.45	0.11	-0.73

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 respondents 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.

The four regions, shown in (Figure 3), express consistency category levels of the variables related to the personality scale (Lachman and Weaver 1997) supplied by the HRS Core LB dataset.

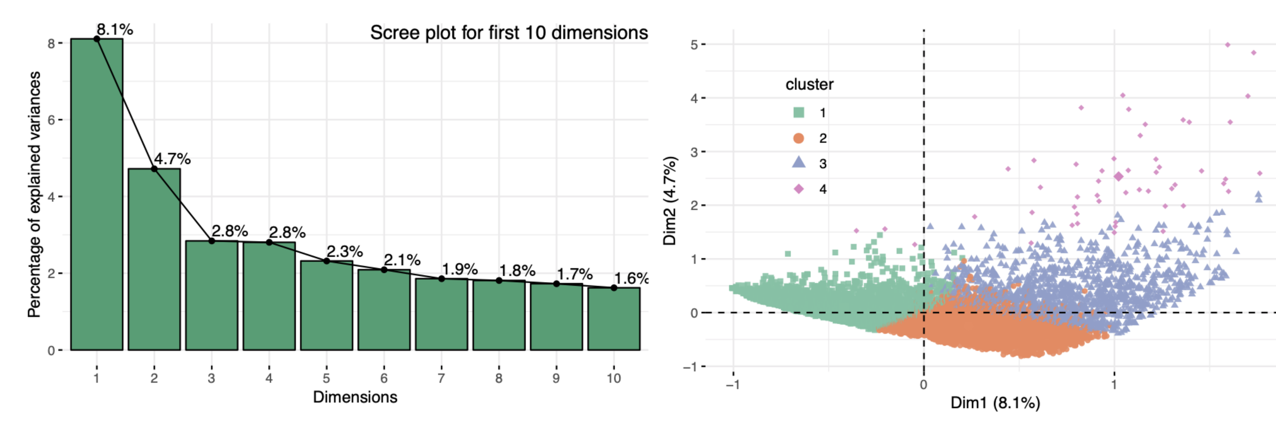

Figure 2: (Left): Variance scree plot. (Right): Hierarchical clustering

The individuals present in Region 1 have an open personality and actively seek for new experiences, while individuals in Region 3 and 4 do not exhibit this characteristic, holding all the low levels of this perception which is defined by a "Not at all" response in most cases. Similarly, the aspect of conscientiousness is a substantial characteristic for individuals in Region 1, and its weakest trace is found in individuals located in Regions 3 and 4.

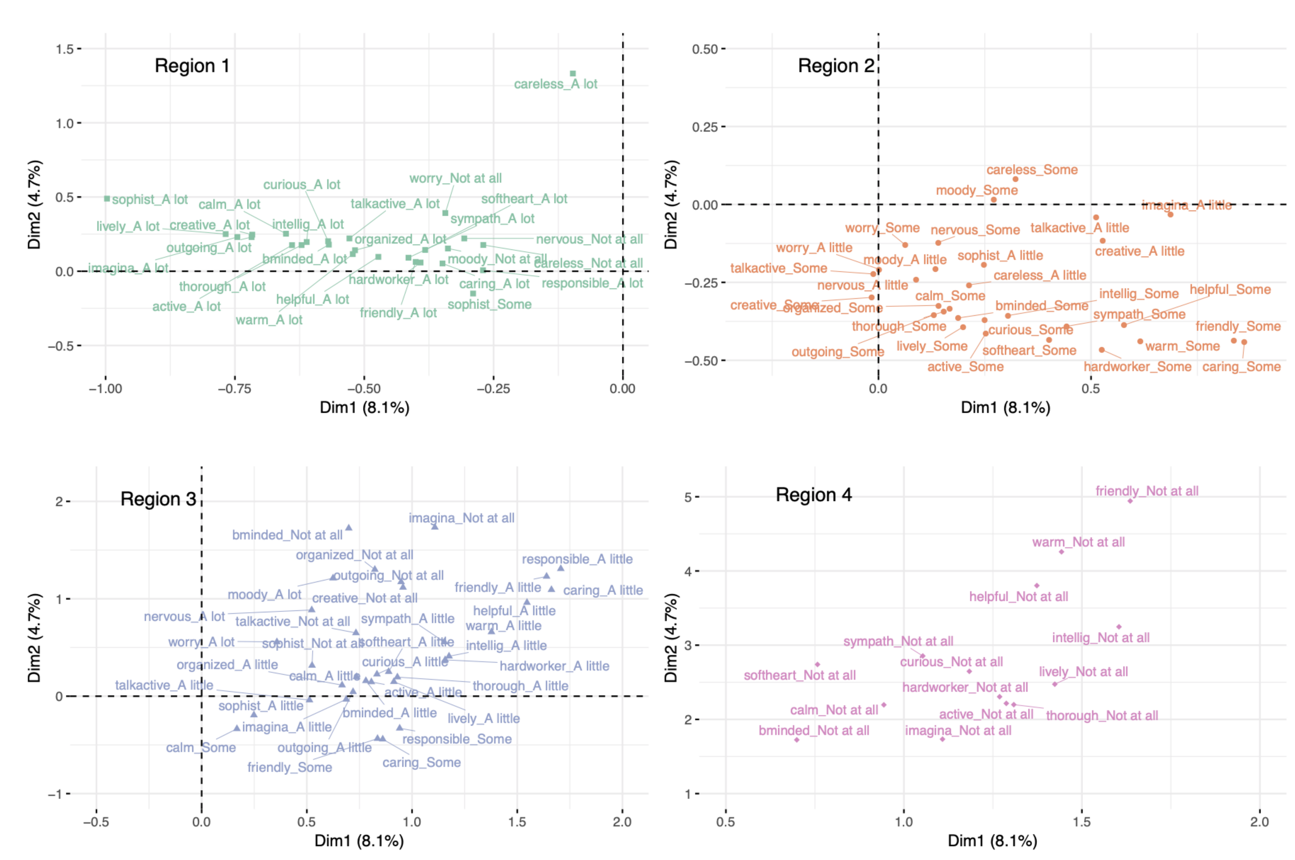

Figure 3: Regions as identified by clustering technique

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.

Acknowledgements

The HRS 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 informed 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

Greenacre, Michael, and Jorg Blasius. 2006. *Multiple Correspondence Analysis and Related Methods.* Chapman; Hall/CRC.

Lachman, Margie E, and Suzanne L Weaver. 1997. "The Midlife Development Inventory (MIDI) Personality Scales: Scale Construction and Scoring." *Brandeis University*, 1–9.

Le Roux, Brigitte, and Henry Rouanet. 2010. *Multiple Correspondence Analysis.* Vol. 163. Sage.

Lê, Sébastien, Julie Josse, and François Husson. 2008. “FactoMineR: An R Package for Multivariate Analysis.” *Journal of Statistical Software* 25 (1).

Multivariate Analysis.” *Journal of Statistical Software* 25 (1).