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

Author 1
Institution 1
Department 1
City, State XXXXX–XXXX
Author1@Institution1.edu

Author 2
Institution 2
Department 2
City, State XXXXX–XXXX
Author2@Institution2.edu

***Abstract*—A geometric data analysis that builds 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. Multiple correspondence analysis is performed on a subset of the survey responses, creating a lower dimensional representation of the respondents and their response patterns, and a hierarchical clustering method is applied to test and validate specific structures in this population study.**

Introduction

In this work the use of unsupervised learning techniques is explored to discover patterns in the survey responses for the US Health and Retirement Study (HRS). The HRS is a rolling cohort of men and women 50 years old and above that began in 1992, with biennial follow-up and periodic recruitment of eligible new participants. Publicly available HRS datasets versions used here were developed at RAND with funding from the National Institute on Aging and the Social Security Administration. Among the variables included are information on the individual, household income and wealth, education, family structure, health behaviors, among others. 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. 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. Furthermore, the use of a hierarchical clustering method using the newly generated representation of each individual has the potential to reveal hidden structures within the population of interest and within the components of the questionnaire used for data collection. This can be used for better survey design, design of experiment, and population sampling based on the characteristics learned via the unsupervised learning methods explored here.

The HRS dataset is a rich source of social, psychological, and economic factors that can serve as important predictors of wellness and health. The authors in [1] investigate how machine learning may add to the understanding of social determinants of health using data derived from the HRS catalog. HRS data was also used in [2] to examine the combined effects of adverse experiences during childhood and adulthood and their relationship to telomere length. The work in [3] examines whether veteran status was associated with elevated depression and anxiety symptoms in men aged 50 and older after adjusting for sociodemographic factors, using records from about 6500 HRS 2006 wave participants and a set of multivariate-adjusted logistic regression analyses. The study in [4], in which HRS data from years 2006, 2008, 2010, and 2012 were used, indicates that subjective memory ratings are associated with factors other than memory itself, including education, personality, depressive symptoms, and subjective age. The approach presented in this work aims to explore and understand the intrinsic structure of a rich dataset such as the HRS data, that has been the object of interest from multiple fields of study and research groups to capture actionable insights to improve the health and wellness of the population.

Methodology

Multiple Correspondence Analysis (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 [5]. 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.

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.

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 an $N \times Q$ such that entry (i, q) is the category of the question

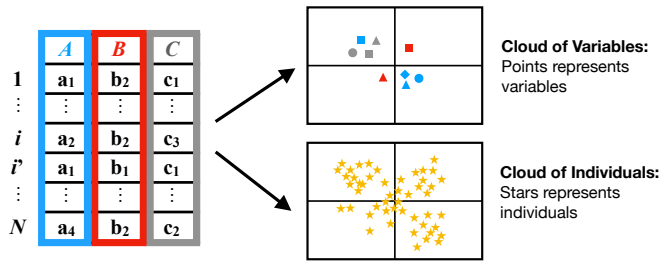

Fig. 1: Clouds of points generated by MCA

q chosen by individual i. Here, the collection of options for question q, that is, the set of levels of variable q is denoted by K_q, and K denotes the overall set of categories in the dataset. 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.

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

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 $L \leq K-Q$ [6] (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

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

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 thee points M^K and $M^{k'}$ is calculated as

$$(M^k, M^{k'})^2 = \frac{n_k + n_{k'} - 2n_{kk'}}{n_k n_{k'}/N} \tag{3}$$

with $n_{kk'}$ denoting the number of respondents who have chosen both categories k and k'. It is worth mentioning that the cloud of categories has the same dimensionality and the same variance as the cloud of individuals. The more categories k and k' have been chosen by the same individuals, the smaller the distance $M^k M^{k'}$. Furthermore, the less frequent a category k, the farther the point M^k is from the center of the cloud.

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:

$$\text{Ctr}_k = \frac{1-f_k}{K-Q}, \quad \text{Ctr}_q = \frac{K_q - 1}{K-Q} \tag{4}$$

The variance of the cloud is simply $V_{\text{cloud}} = \frac{K}{Q} - 1$ and the mean of eigenvalues is given by $\bar{\lambda} = \frac{1}{Q}$.

For interpreting an axis, we use the method of contributions of points and deviations. Due to the high dimensionality of the clouds, the variance rates of the principal axes are generally low. [7] proposed to use modified rates in order to better understand the importance of the first principal dimensions. Variance rates are calculated following the steps below:

For $l = 1, 2, ..., l_{\max}$ such that $\lambda_l > \bar{\lambda}$ calculate:

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

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

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 low-dimensional 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 [6].

The Health and Retirement Study (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 multi-disciplinary 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. The methodology uses an interview typically conducted in person, by telephone with a duration of approximately 2.72 hours, including topics such as health status and conditions, employment history, internet usage, health care utilization, cognitive status, disability status, physical functioning, housing status, family characteristics, work environment characteristics, among others.

The HRS data is organized in public and restricted materials and contains survey records produced by ISR and third parties which publish data based on processed HRS data. The public HRS files are divided into three files called HRS Core, Exit, and Post-Exit. Also, the data produced by the

RAND ("Research ANd Development") Center for the Study of Aging, and USC Program on Global Aging, Health, and Policy, and data produced by researchers is available. This study uses the following data products: HRS Core Cognition Section (D) [8], HRS Left-Behind Questionnaires Section LB [9], and the RAND HRS Longitudinal File 2014 (V2) [10]. All the data used in this work was related to the survey waves of 2006, 2008, 2010 and 2012. The Cognition section provides the variables related to memory performance and subjective memory, not included memory diseases which are not part of variables used in the studies. The psychosocial and lifestyle questionnaires from 2006 to 2010 are self-administered questionnaires that stay with the respondents after the completion of an in-person core interview, covering six main areas: subjective well-being, lifestyle and experience of stress, quality of social ties, personality traits, work-related beliefs, and self-related beliefs.

Data used in this study comes from waves 08, 09, 10 and 11, corresponding to years 2006, 2008, 2010, and 2012 respectively. The sociodemographic data was extracted from the RAND datasets with variables such as gender (male, female), age, race (White/Caucasian, Black/African American, other), ethnic group (Hispanic, non-Hispanic), education (none, elementary, middle school, high school, college, other), and military service (veteran, non-veteran) Information associated with mental health was separated in data related to depression and anxiety. Depression variables extracted from the RAND dataset capture the result of the mental health index derivation using the CES-D scale (Center for Epidemiological Studies Depression Scale). The CES-D score is constructed based on the sum of five negative indicators minus two positive indicators. All the positive and negative indicators are obtained from a closed-ended question with *yes* or *no*, as the possible answers [11].

Results and Discussion

Multiple Correspondence Analysis

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. Therefore, the results shown in this study include responses for two different years with different participants to test the ability of the technique on a large set of observations (other years combinations were calculated and very similar results were obtained using the geometric data analysis framework discussed here). The dimensions of the final dataset used in this analysis are 9732×34, where each row of the tabular data set represents one of the survey respondents and each column is a question included in the questionnaire. A summary of the frequency of responses is shown in Table I.

Category	nk	fk	ctrk
sophisticated_Not at all	2159	0.22	0.0093
imaginative_A lot	3109	0.32	0.0081
creative_A lot	2668	0.27	0.0086
caring_A little	345	0.04	0.011
talkactive_A lot	2823	0.29	0.0085
friendly_A little	372	0.04	0.011
careless_Some	993	0.10	0.011
responsible_Some	1729	0.18	0.0098
responsible_A little	236	0.02	0.012
nervous_Not at all	3101	0.32	0.0081
worry_Not at all	1625	0.17	0.0099
moody_Some	1470	0.15	0.01

TABLE I: Response frequencies (absolute n_k and relative f_k), and contributions (Ctr_k) of top categories by Ctr_k and levels

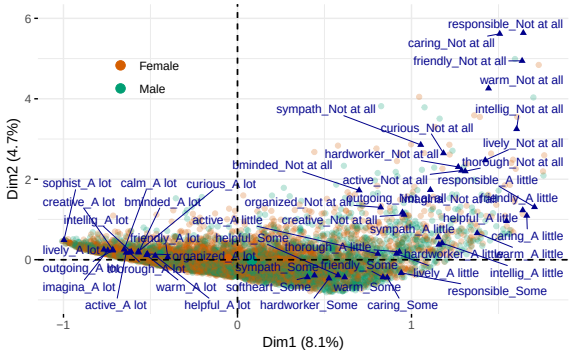

Fig. 2: Projection onto the first two principal dimensions

The cloud in Figure 2 is the projection of the full cloud onto the plane of the principal axes 1 and 2, that is, onto the principal plane 1-2. The location of the top contributing categories is also shown, in a graph known as the *biplot* of individuals and variables. In Figure 3 the projection on plane 3-4 is presented, and the participants are colored by gender. Notice that in this plane, a clear separation is found by this categorical variable, and the collection of associated categories is informative of the pattern shown by the survey respondents.

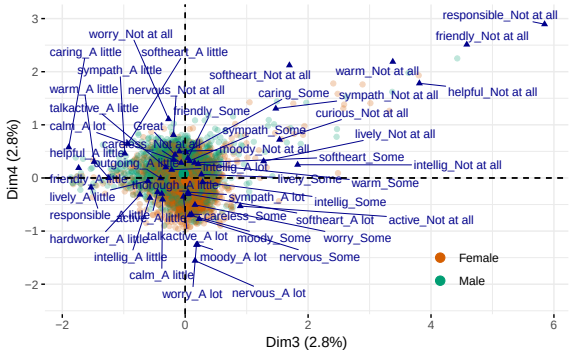

Fig. 3: Biplot in dimensions 3 and 4 with top contributing variables. Respondents are color-coded based on the supplementary variable gender

The coordinates of the first 4 dimensions for the top 12 categories (sorted by contribution Ctr_k and level of agreement) are shown in Table II. 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.

Variable	Dim 1	Dim 2	Dim 3	Dim 4
sophisticated_Not at all	0.52	0.32	0.03	-0.14
imaginative_A lot	-0.75	0.23	-0.20	0.04
creative_A lot	-0.72	0.25	-0.23	0.03
caring_A little	1.66	1.09	-1.90	0.58
talkactive_A lot	-0.53	0.22	-0.03	-0.35
friendly_A little	1.64	1.23	-1.73	0.19
careless_Some	0.32	0.08	0.10	-0.68
responsible_A little	1.71	1.31	-1.53	-0.18
responsible_Some	0.94	-0.33	0.07	-0.06
nervous_Not at all	-0.31	0.22	-0.19	0.81
worry_Not at all	-0.34	0.39	-0.27	1.11
moody_Some	0.27	0.02	0.07	-0.70

TABLE II: Coordinates of the first 4 dimensions for the top 12 categories (sorted by contribution Ctr_k and level of agreement)

Table III shows the modified rates for the first principal axes produced by MCA on the HRS survey data.

Axes	1	2	3	4	5	6
Eigenvalue (λ_l)	0.244	0.142	0.086	0.085	0.070	0.063
Variance rate	0.081	0.047	0.029	0.028	0.023	0.021
Modified rate (τ_l)	0.688	0.180	0.039	0.038	0.019	0.012

TABLE III: Variances of axes, variance and modified rates

Figure 4 shows the percentages of variance of the first 10 dimensions. The first principal axis explained 8.11% of the principal inertia, the second principal axis explained 4.72%, and none of the remaining principal axes explained more than 3%. Using the modified variance rates τ_l one can see that the first two dimensions explain about 86.8% of the variance in the data.

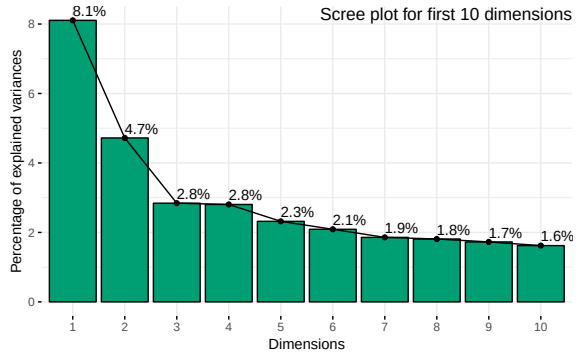

Fig. 4: Variation explained by each principal component

	Dim 1	Dim 2	Dim 3	Dim 4
level_of_education	0.03	0.02	0.01	0.08
subjective_memory	0.10	0.04	0.01	0.07
sophisticated	0.21	0.07	0.03	0.01
broadminded	0.25	0.18	0.08	0.01
curious	0.28	0.19	0.09	0.01
intelligent	0.37	0.21	0.13	0.03
imaginative	0.33	0.14	0.06	0.01
creative	0.28	0.15	0.05	0.01
sympathetic	0.27	0.19	0.10	0.11
softhearted	0.22	0.15	0.12	0.17
caring	0.36	0.21	0.26	0.14
warm	0.44	0.22	0.25	0.09
helpful	0.39	0.18	0.16	0.04
talkactive	0.17	0.06	0.04	0.07
active	0.35	0.22	0.09	0.02
lively	0.42	0.23	0.14	0.01
friendly	0.40	0.21	0.20	0.08
outgoing	0.33	0.14	0.06	0.02
careless	0.06	0.07	0.00	0.10
thorough	0.33	0.18	0.07	0.02
hardworker	0.31	0.20	0.08	0.01
responsible	0.30	0.18	0.19	0.03
organized	0.21	0.13	0.05	0.02
calm	0.27	0.17	0.08	0.08
nervous	0.05	0.09	0.02	0.47
worry	0.04	0.09	0.02	0.46
moody	0.07	0.07	0.01	0.22

TABLE IV: Squared correlation between each variable and the principal dimension

In Table IV the squared correlation between each of the questions in the survey considered in this study and the principal dimensions is shown.

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 [12]. 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. Hierarchical clustering requires to define a distance and an agglomeration criterion. Here the traditional Euclidean distance for calculating dissimilarities between observations and the complete linkage agglomeration method were used [13].

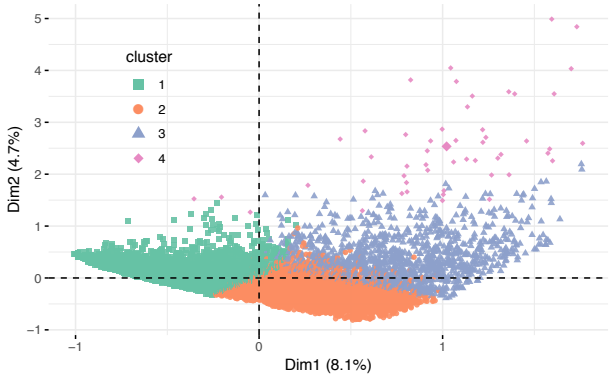

Fig. 5: Hierarchical clustering using principal dimensions generated by multiple correspondence analysis.

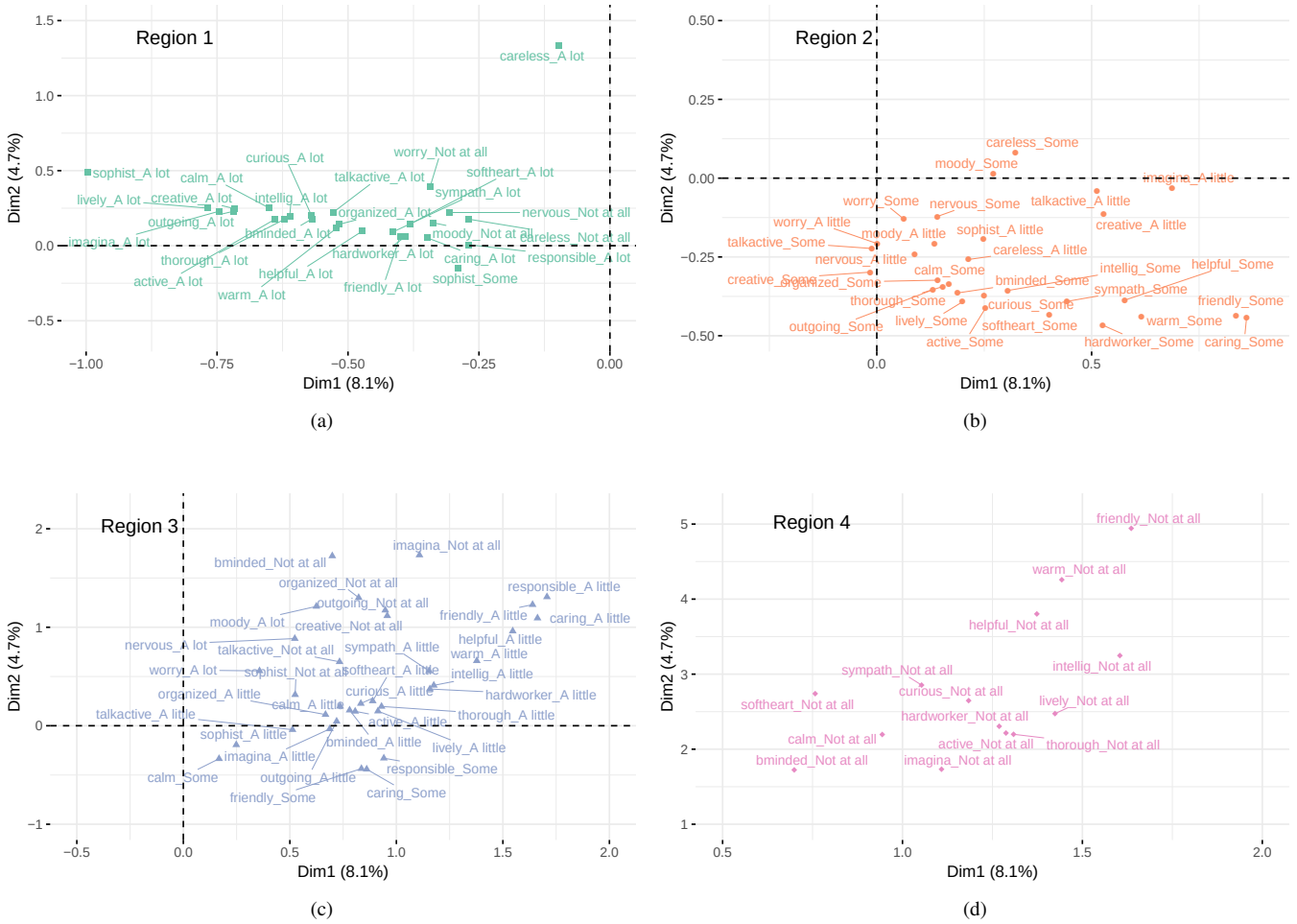

Fig. 6: Clouds of categories: (a) Region 1, (b) Region 2, (c) Region 3, (d) Region 4

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 6 shows the categories that fall in the regions suggested by the clustering procedure results shown in Figure 5.

Variables indicating gender, veteran status, ethnicity, race, and depression status were used as supplementary categories. A supplementary category (sometimes called illustrative category), is a category that is not used to define the distance between individuals. Recall that every variable has multiple levels related to the rate of agreement with the associated question. Table V shows the coordinates in the first four dimensions for the supplementary variables considered here. The deviation between female and male respondents on axis 4 is notable ($|-0.183-0.277| = 0.46$) confirming the suggested grouping observed in Figure 3 for plane 3-4.

Category	Dim 1	Dim 2	Dim 3	Dim 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

TABLE V: Coordinates of the supplementary variables on the first four axes

The four regions identified here, and shown in Figure 6, express consistency category levels of the variables related to the personality scale [14] supplied by the HRS Core LB dataset. The personality scale used in this work covers five aspects: conscientiousness, neuroticism, extroversion, agreeableness, and openness. The individuals present in Region 1 have an open personality and actively seek for new experi-

ences, 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 (which is related to organization, persistence, control, and motion in goal) is a substantial characteristic for individuals in Region 1, and its weakest trace is found in individuals located in Regions 3 and 4. In general, Region 2 holds individuals which scale all the four main characteristics as *"Some"* and *"A little"* levels, with a high frequency of extraversion and neuroticism, that combined depict the *"style of well-being"* described in the sample report in [15]. Individuals with low levels of neuroticism and extraversion (N-E-Low-keyed) maintain considerable indifference towards good or bad news, sometimes impacting their interpersonal relationships. In contrast to Region 3, which holds the extreme lower levels like *"A little"* and *"Not at all"*. Region 2 is the middle point for all the "Big 5" personality characteristics collected in the HRS Core LB survey.

In this work, the R packages `dplyr` [16] and `ggplot2` [17] were used for data wrangling and visualization, and the `haven` package [18] for importing data. The MCA algorithm used in this study corresponds to the implementation of the algorithm available in the `FactoMineR` package [19], that includes a collection of methods for multivariate data analysis. Additionally, the `factoextra` R package was used for visualization and interpretation [20].

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.

Acknowledgment

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 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

[1] B. Seligman, S. Tuljapurkar, and D. Rehkopf, "Machine learning approaches to the social determinants of health in the health and retirement study," *SSM-population health*, vol. 4, pp. 95–99, 2018.

[2] E. Puterman, A. Gemmill, D. Karasek, D. Weir, N. E. Adler, A. A. Prather, and E. S. Epel, "Lifespan adversity and later adulthood telomere length in the nationally representative us health and retirement study," *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, pp. E6335–E6342, 2016.

[3] C. E. Gould, T. Rideaux, A. P. Spira, and S. A. Beaudreau, "Depression and anxiety symptoms in male veterans and non-veterans: the health and retirement study," *Int J Geriatr Psychiatry*, vol. 30, pp. 623–630, 01 2015.

[4] G. Hlr, C. Hertzog, A. M. Pearman, and D. Gerstorf, "Correlates and moderators of change in subjective memory and memory performance: Findings from the health and retirement study," *Gerontology*, vol. 61, pp. 232–240. [Online]. Available: http://dx.doi.org/10.1159/000369010

[5] B. Le Roux and H. Rouanet, *Multiple correspondence analysis*. Sage, 2010, vol. 163.

[6] M. Greenacre and J. Blasius, *Multiple correspondence analysis and related methods*. Chapman and Hall/CRC, 2006.

[7] J.-P. Benzećri, *Correspondence analysis handbook*. CRC Press, 1992.

[8] A. Sonnega and D. Weir, "The health and retirement study: A public data resource for research on aging," *Open Health Data*, vol. 2, no. 1, 2014.

[9] J. Smith, G. Fisher, L. Ryan, P. Clarke, J. House, and D. Weir, "Psychosocial and lifestyle questionnaire," *Survey Research Center, Institute for Social Research*, 2013.

[10] Health and Retirement Study, *(HRS Core) public use data set*. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740): Ann Arbor, MI, 2014.

[11] L. S. Radloff, "The CES-D scale: A self-report depression scale for research in the general population," *Applied psychological measurement*, vol. 1, no. 3, pp. 385–401, 1977.

[12] I. Jolliffe, "Principal component analysis for special types of data," *Principal component analysis*, pp. 338–372, 2002.

[13] P.-N. Tan, M. Steinbach, and V. Kumar, "Data mining cluster analysis: basic concepts and algorithms," *Introduction to data mining*, 2013.

[14] M. E. Lachman and S. L. Weaver, "The midlife development inventory (MIDI) personality scales: Scale construction and scoring," *Waltham, MA: Brandeis University*, pp. 1–9, 1997.

[15] C. Paul and M. Robert, "Revised NEO personality inventory," 01 2016. [Online]. Available: 10.1037/t03907-000

[16] H. Wickham, R. Francois, L. Henry, and K. Müller, *dplyr: A Grammar of Data Manipulation*, 2018, R package version 0.7.8. [Online]. Available: https://CRAN.R-project.org/package=dplyr

[17] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[18] H. Wickham and E. Miller, "haven: Import and export SPSS, STATA and SAS files," *R package version*, vol. 1, no. 0, 2018.

[19] S. Lê, J. Josse, F. Husson *et al.*, "FactoMineR: an R package for multivariate analysis," *Journal of statistical software*, vol. 25, no. 1, pp. 1–18, 2008.

[20] A. Kassambara and F. Mundt, *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2017, R package version 1.0.5. [Online]. Available: https://CRAN.R-project.org/package=factoextra