MCA Clustering Analysis on Obesity Dataset Report

Wen Juntong **Liu Xiangheng** **Hu Ziwei** **Pan Hewen**

Abstract

In this work, we reproduce the Unsupervised Learning on the Health and Retirement Study using Multiple Correspondence Analysis (MCA) and Hierarchical Clustering on a new data set related to obesity. We aim to uncover the underlying patterns among individuals based on categorical survey responses related to diet, physical activity, and demographic factors. We utilize MCA for dimensionality reduction followed by hierarchical clustering to identify distinct groups. By visualizing and clustering individuals based on their MCA projections, we observe distinct clusters corresponding to different health risks.

1 Introduction

Obesity research is vital to understanding health-related behaviors and their effects on life. By exploring factors related to obesity, such as age, weight, eating habits, and physical activity, we can identify patterns that contribute to healthier lifestyles. We use MCA and Hierarchical Clustering to analyze an obesity dataset and find interesting trends among subgroups. However, several challenges arise during this process. One key challenge is effectively reducing high-dimensional categorical data using MCA while preserving the relationships between variables. Additionally, the preprocessing of continuous variables into categorical bins, followed by one-hot encoding, introduces complexity. The interpretation of the clustering results is also difficult, especially when defining clear boundaries between risk groups. Finally, selecting the appropriate number of clusters involves a degree of subjectivity, which requires careful consideration.

Using dimensionality reduction, Multiple Correspondence Analysis examines categorical data linkages. Data is reduced in dimension while variable relationships are preserved. This approach works well with categorical or ordinal survey data. MCA lets us visualize how individuals relate to categories and expose data structures by projecting persons and categories into the same coordinate space. We reduce the obesity dataset's dimensionality after MCA to find the most essential components that capture the data's key variations. After further analysis, we can find correlations between dataset attributes.

Similarity-based hierarchical clustering groups people. Hierarchical clustering treats each person as a cluster and combines the closest groups based on similarity. This research uses hierarchical clustering on MCA-reduced dimensions. We intend to find obesity-related clusters by grouping individuals by their projections into the first two axes of MCA.

2 Methodology

2.1 Data Preprocessing

This dataset contains obesity and cardiovascular disease risk data for 20758 individuals from Mexico, Peru, and Colombia, including 17 attributes, such as sex, age, eating habits, and physical condition. The dataset is designed to analyze obesity-related factors and support classification and cluster analysis to help identify different risk groups. In order to prepare for multiple correspondence analysis (MCA), all variables need to be converted to categorical form. However, the original dataset

Preprint. Under review.

contains multiple continuous numerical variables (e.g., age, height, weight, and some behavioral scores such as FCVC, CH2O, etc.), which cannot be directly used for MCA. Therefore, in this preprocessing step:

1. All numerical columns are identified and the original categorical variables are preserved;
2. Binning rules for key continuous variables are defined using domain knowledge or reasonable thresholds (e.g., binning Age into groups like 0–18, 18–30, etc.);
3. The binned variables are concatenated with the original categorical columns to form the final dataset which contains only categorical features and is ready for MCA analysis.

2.2 Multiple Correspondence Analysis

After data preprocessing, the dimension of the final dataset is 20758*58, which means that each row represents one individual point, a total of 20758 individual points, and each column represents categorical variable values, a total of 58 new categorical variable features.

In order to clearly see the structures inside the individual points and categorical variables, the variance ratios of each principal axis need to be calculated to get the projection of the whole cloud in the principal axis planes. The variance ratios are calculated following the steps below:

1. The category variables were coded into a 0/1 (binary) indication matrix with a data shape of 20758*58 and the first 6 principal components were extracted.
2. Apply SVD to the normalized indicator matrix Z,

$$Z = U\Sigma V^\top$$

Σ is diagonal matrix, including singular values $\sigma_1, \sigma_2, ...\sigma_k$

3. Eigenvalue λ_i is the square of a singular value,

$$\lambda_i = \sigma_i^2$$

4. Explained Variance Ratio for $\text{Dim}_i = \frac{\lambda_i}{\sum_j \lambda_i}$

The variance ratios on six dimensions of top 10 categorical variables is shown in Table. 1

Variables	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6
Age_60-100	0.68	1.84	0.42	1.76	2.98	6.75
MTRANS_Bike	0.76	0.72	-0.25	1.04	1.60	1.01
MTRANS_Motorbike	0.53	0.33	-0.17	1.03	0.58	0.79
Age_50-60	0.29	2.24	-0.51	0.73	-0.40	4.33
Height_1.9-2.0	-1.01	0.15	2.58	1.60	0.18	-1.60
SMOKE_yes	-0.31	0.52	1.46	1.30	0.23	-0.31
Age_40-50	0.03	1.81	-0.32	1.19	-2.39	-0.11
FCVC_0-1	1.12	-0.25	-0.22	-0.19	0.11	-0.26
CAEC_no	0.84	0.43	-1.68	2.25	2.89	-2.11
MTRANS_Walking	1.14	0.17	0.06	-0.49	0.87	1.50

Table 1: Variance Ratios on Six Dims of Top 10 Categorical Variables

However, In high-dimensional data, variance rates are often spread out and small. It becomes hard to interpret which dimensions (axes) are truly important. A new method called modified rates [1] was proposed to highlight the most informative dimensions:

1. For i=1,2,3...n, when $\lambda_i > \overline{\lambda}$

$$\text{the pseudo-eigenvalue } \lambda' = (\frac{Q}{Q-1})^2(\lambda_i - \overline{\lambda})^2$$

2. the sum $S = \sum \lambda_i^{'}$

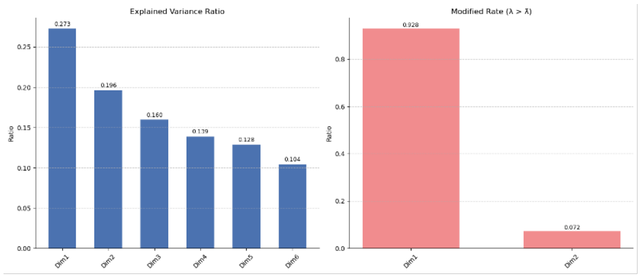

Figure 1: Results of variance ratio and modified rate

3. Therefore, the modified rate $\tau_i^{'} = \lambda_i^{'}/S$ [2]

After applying modified rates, the dim 1 and dim 2 are conclude to be the most important dimensions, which means they are more explanatory and have a clearer distinction between variables. The results of using variance ratio and modified rate are shown in Fig. 1.

2.3 Hierarchical Clustering

After completing the multiple correspondence analysis, the low-dimensional space constructed by individual points in dim1 and dim2 was utilized for hierarchical clustering. Euclidean distance and complete linkage were used for clustering and the number of clusters was set to 3 based on interpretability and visual separation.

3 Results

3.1 MCA analysis

Multiple Correspondence Analysis (MCA) is a dimensionality reduction technique designed for categorical variables that maps high-dimensional data into a low-dimensional space (typically 2-3 dimensions) to reveal latent associations between variables and individuals. In the MCA results visualized in the figure 2, health behavior variables (e.g. BMI categories, commuting modes, frequency of alcohol consumption) are projected onto a two-dimensional plane defined by Dim1 (horizontal axis) and Dim2 (vertical axis). The core logic revolves around the interaction between two types of coordinates: category coordinates (spatial positions of categorical variables) and individual coordinates (positions of respondents).

Each category of a variable (e.g., "Weight 120-166" for high body weight, "TRANS Bike" for cycling commuters) is assigned a distinct coordinate point on the plane. These category coordinates define "gravitational centers" within the space.

The dense scatterplot distribution and regional partitioning (Regions 1–3) visually encapsulate these multidimensional gravitational forces, providing a spatially driven data-centric explanation for health risk stratification.

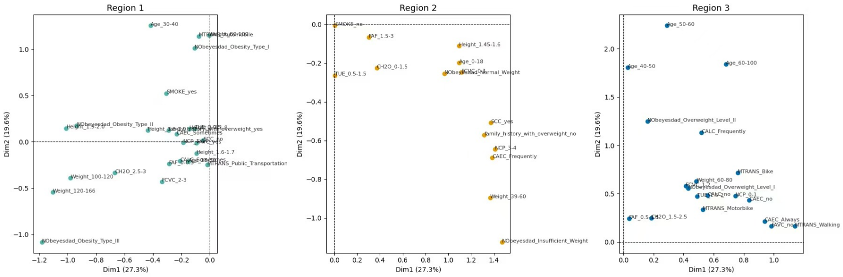

Figure 2: Variables coordinate region

3.2 Hierarchical clustering analysis

The MCA and hierarchical clustering analysis figure 3 revealed three distinct health risk clusters: low-risk (Cluster 0, blue) concentrated on the right side of Dim1 (27.3% variance), characterized by healthy behaviors (active transportation, normal BMI, no smoking/alcohol); high-risk (Cluster 2, green) clustered on Dim1's left (aligned with obesity, motorized commuting, and smoking); and medium-risk (Cluster 1, orange) in intermediate Dim1 and upper Dim2 (19.6% variance), showing partial metabolic risks. Gender markers (gray/black crosses) show no spatial clustering, suggesting risk profiles are behavior-driven rather than gender-specific. This spatial stratification supports prioritizing interventions for the high-risk group (left) while reinforcing healthy habits in the low-risk cluster (right).

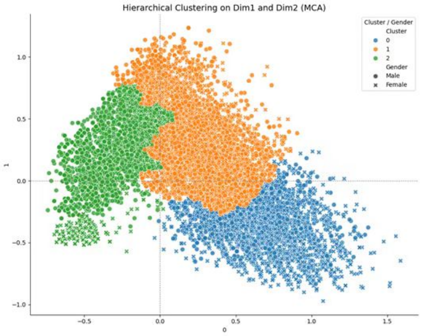

Figure 3: Clustering result

4 Conclusion

4.1 Research Findings and Practical Implications

This study integrates Multiple Correspondence Analysis (MCA) and hierarchical clustering to establish a data-driven framework for stratifying populations into distinct obesity risk categories (low, medium, high risk). Spatial visualization on the MCA plane (Dim1: 27.3% variance contribution, Dim2: 19.6%) reveals core risk drivers, including sedentary lifestyles, poor dietary habits, and metabolic-genetic predispositions. These findings directly inform targeted interventions: low-risk groups (Cluster 0) benefit from community health programs reinforcing active commuting and balanced diets; medium-risk groups (Cluster 1) require preventive measures like smoking cessation counseling and dietary monitoring to curb risk escalation; high-risk groups (Cluster 2) demand prioritized access to clinical weight management and metabolic health surveillance. Additionally, the results provide scientific evidence for optimizing public health resource allocation (e.g., targeted funding for high-risk subgroups) and policy design (e.g., urban planning for walkability-friendly infrastructure).

4.2 Future Research Directions

To enhance model robustness and applicability, future work will incorporate genetic data (e.g., obesity-related SNPs), wearable device-derived exercise metrics, and socioeconomic factors to refine risk stratification. Integrating machine learning models (e.g., random forests, neural networks) will enable prediction of individual risk trajectories and dynamic optimization of intervention timing. Longitudinal cohort studies will assess intervention efficacy and identify critical transition points (e.g., shifts from medium to high risk). Finally, developing digital tools like AI-driven health assistants will deliver personalized prevention strategies based on real-time behavioral data, transforming obesity management from a one-size-fits-all approach to precision-based strategies.

The link of code in Github is: Project Link

References

[1] Benzécri, J.-P. (1992) *Correspondence Analysis Handbook.* Boca Raton: CRC Press.

[2] Sanchez-Arias, R. & Batista, R.W. (2019) Unsupervised learning on the Health and Retirement Study using geometric data analysis. In *Proceedings of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Boca Raton, FL, pp. 335–340. IEEE. doi: 10.1109/ICMLA.2019.00063.