Homework 2: Multiple Correspondence Analysis and Clustering

Exploratory Data Analysis

Import the data set correctly to R and assign the type of each variable correctly.

Variables *Fresh*, *Milk*, *Grocery*, *Frozen*, *Detergents_Paper*, and *Delicatessen* are numeric variables, whereas *Channel* and *Region* variables are factors.

Convert the numerical variables to factors consisting of two categories, "low" and "high" by cutting them from their median and saving them as new variables in the data set.

To categorize the numerical variables into "low" and "high" based on their median values, we defined a reusable function, *numerical_to_factor*. This function takes the dataset and a column name as arguments and creates a new factor variable with categories "low" and "high," saving it as a new column in the dataset.

Application of Multiple Correspondence Analysis (MCA)

Apply MCA to the categorical variables, taking Region and Chanel as supplementary variables by using the MCA() function in the FactoMineR package. Interpret your findings.

We applied Multiple Correspondence Analysis (MCA) using the indicator matrix and evaluated the number of dimensions to retain using three criteria: percentage of inertia explained (eigenvalues), cumulative inertia, and the elbow rule.

To retain dimensions with eigenvalues greater than the average inertia (1/6 = 0.167), we would keep the first two dimensions (λ_1 = 0.406, λ_2 = 0.224). For cumulative inertia, retaining the first three dimensions (λ_1 = 0.406, λ_2 = 0.224, λ_3 = 0.130) explains 76% of the variance, meeting the standard 70%-80% threshold, while the first two dimensions explain 63%. The scree plot (Figure 1) shows a noticeable elbow after Dim. 3.

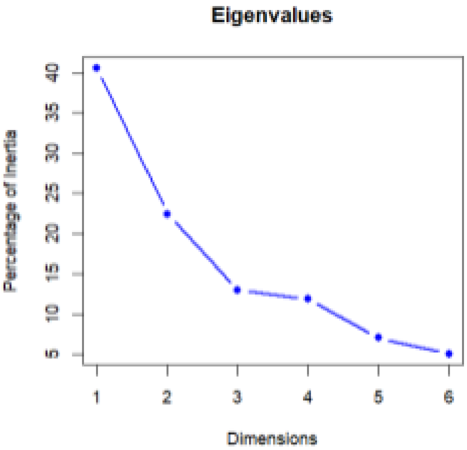

Figure 1. Percentage of inertia explained by dimensions.

To conclude, the number of dimensions to retain is somewhat subjective. Based on the first rule, only the first two dimensions should be kept. However, if the explained variance is prioritized over simplicity, adding the third dimension increases the explained variance by 13%, reaching 76%, which is further supported by the elbow rule. For simplicity, retaining just two dimensions (63%) may suffice.

The variables *f.grocery*, *f.milk*, and *f.detergents_paper* dominate the first dimension, as indicated by their high contributions to the explained variance (res.mca$var$contrib). On the factor map (Figure 2), higher levels of these variables align with the positive side of Dimension 1, while lower levels align with the negative side.

This dimension represents daily essentials. Customers with high scores on Dimension 1 tend to purchase larger quantities of these, while those with low scores purchase less of these product categories. The supplementary variable Channel_2 (Hotel/Restaurant/Café customers) aligns positively with Dimension 1, suggesting a preference for buying household goods, likely due to operational needs. In contrast, Channel_1 (retail customers) aligns negatively, reflecting smaller-scale purchases that are more typical of individual consumers. The cos^2 values (res.mca$var$cos2) for *f.milk*, *f.grocery*, and *f.detergents_pape*r are exceptionally high for Dimension 1, confirming that these variables are well-represented by this dimension.

The strongest contributors to the second dimension are the variables *f.frozen*, *f.fresh*, and *f.delicassen*. High levels of these variables align with the positive Y-axis, while low levels align with the negative Y-axis. This dimension reflects preferences for fresh, frozen, and specialty items. Customers scoring high on Dimension 2 focus on perishable and high-value products, often associated with health-conscious consumption patterns. Conversely, low scores suggest a lower emphasis on these items. The cos^2 values also indicate that variables such as *f.frozen* and *f.fresh* are better represented in Dimension 2, reinforcing their significant contributions to this axis.

The variables *f.fresh* and *f.frozen* also contribute significantly to the third dimension (res.mca$var$contrib). Thus, the third dimension offers limited additional insight and may overcomplicate interpretation without adding substantial value.

Finally, while the supplementary variable Channel differentiates consumer groups effectively in the first dimension, the Region variable is located near the center of the plot. This indicates that it does not provide meaningful insights regarding whether customers from different regions purchase more or fewer household products (Dimension 1) or delicatessen products (Dimension 2).

Figure 3 shows how individuals seem to be similarly distributed along the quadrants of the two-dimensional factor map.

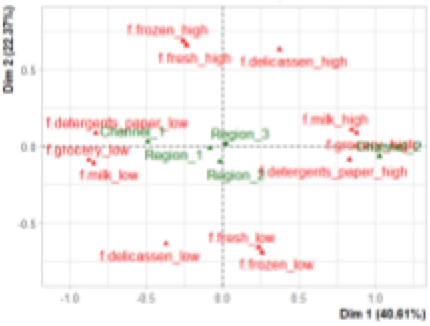

Figure 2. MCA factor map

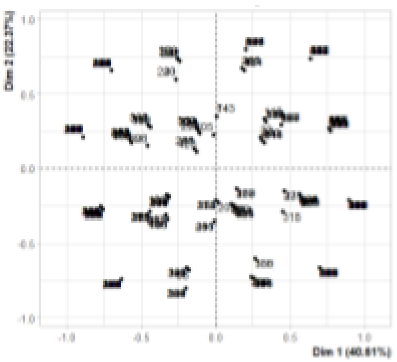

Figure 3. Individuals over the MCA factor map

Hierarchical Clustering based on MCA and PCA scores

Apply hierarchical clustering on MCA scores obtained from the previous step by using the HCPC function. How many clusters are constructed?

The dendrogram obtained when applying the HCPC function (Figure 3) indicates that four clusters are constructed. The horizontal black line on the dendrogram represents the cut-off point, and there are four distinct branches above this line.

In addition, Figure 4 provides further support for this conclusion. The inertia gained per cluster rapidly decreases after the first few clusters, and the inertia gain flattens notably from the fifth cluster onward.

Combining these two observations (and the *res.hcpc$call$t$nb.clust* = 4 value), it is reasonable to conclude that 4 clusters are the optimal number of clusters to capture the structure of the data effectively.

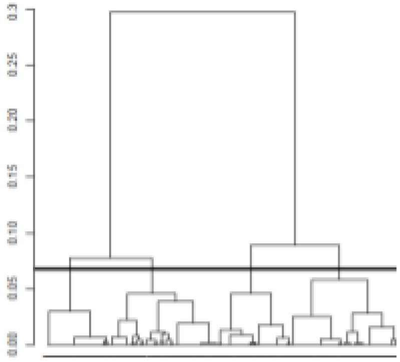

Figure 3. Hierarchical clustering dendrogram

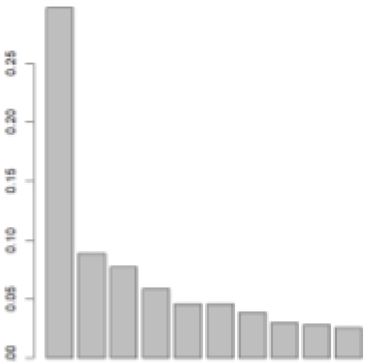

Figure 4. Inertia gained per cluster

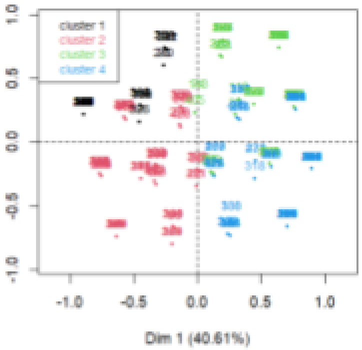

Figure 5. Cluster MCA factor map

Apply hierarchical clustering to PCA scores, considering only the numerical variables and taking Region and Channel variables as supplementary variables. How many clusters are constructed?

We began obtaining the PCA scores of our dataset with the PCA method. This included all dimensions that captured significant variance in the data. Based on this result, we applied HCPC with the default parameters, which formed 3 clusters. Within this approach, Cluster 1 has 393 individuals, Cluster 2 has 44 individuals, and Cluster 3 has three individuals.

Note that the third cluster is not representative since it only contains three individuals. It is essentially created because the HCPC function has a parameter that sets the minimum number of clusters to be made, which by default has a value of three. Setting this parameter to 2 (min = 2) generates two clusters (see Figure 6).

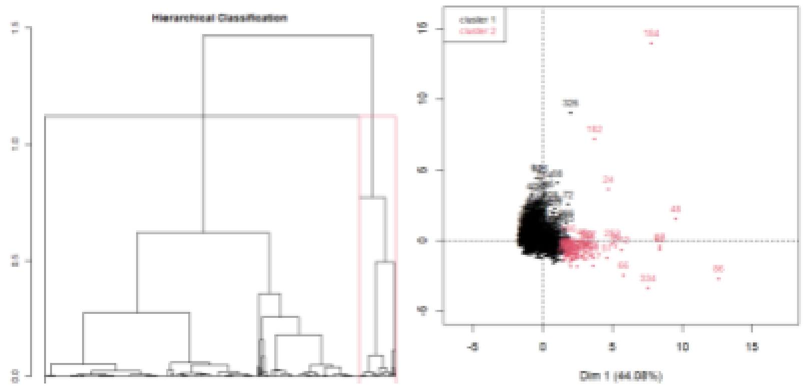

Figure 6. Dendrogram and scatter plot representing clusters obtained with the HCPC method and no minimum number of clusters.

Upon examining the eigenvalues from the PCA output, we observed that only the first two principal components had eigenvalues greater than 1. Based on this observation, we decided to limit the number of dimensions to two (ncp = 2) on the PCA for further analysis.

With this adjustment, we re-applied hierarchical clustering using the PCA scores restricted to these two principal components. Interestingly, this approach resulted in four clusters, a different outcome from the two clusters obtained without restricting the dimensionality (see Figure 7).

Note that the most populated clusters are Cluster 1 (319 observations) and Cluster 2 (96 observations). The 3rd and 4th clusters have 15 and 10 individuals assigned, respectively.

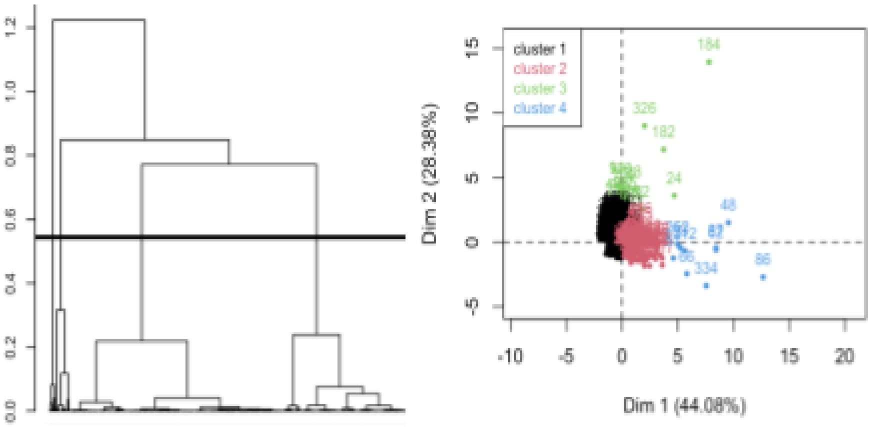

Figure 7. Dendrogram and scatter plot representing clusters obtained with the PCA limited to 2 dimensions

To conclude, although the four-cluster approach identifies two additional groups of individuals with higher values on dimensions 1 and 2, this segmentation arises from reducing the dimensionality provided to the HCPC method (limiting it to the first two PCA dimensions). This restriction leads to the loss of variability captured in higher dimensions. Furthermore, the 3rd and 4th clusters are marginal, representing only 3% and 2% of the population, respectively.

Given these considerations, we determine that the division of the data into two clusters is the most representative and meaningful outcome of hierarchical clustering on the PCA results.

Profiling

Interpret the clusters obtained from both methods.

1. Hierarchical clustering based on MCA components

Before analyzing the clusters directly, we examine the relationships between the categorical variables and the cluster variable. Using chi-squared tests, we confirm that all categorical variables included in the MCA are significantly associated with the cluster variable. For the supplementary variables, *Channel* shows a significant association with the clusters, whereas *Region* does not. Consequently, *Region* is excluded from the interpretation.

The first cluster is characterized by individuals with higher values on the second dimension and lower values on the first dimension, placing them in the first quadrant of the MCA factor map (see Figure 5). These individuals exhibit high expenditures on fresh, frozen, and delicatessen products while spending less than the median on detergents, paper, milk, and groceries. Notably, the cluster predominantly represents retail customers (Channel_1), indicating that it comprises retail buyers who prioritize fresh, frozen, and delicatessen items over household essentials.

The second cluster mainly represents retail customers with low values on both dimensions, suggesting that individuals in this group generally exhibit low spending across all product categories (groceries, milk, detergents, paper, fresh, delicatessen, and frozen). However, some members of this cluster are close to the mean consumption patterns. Overall, this cluster encapsulates low-spending retail customers, indicative of individuals with constrained purchasing power or basic shopping needs.

The third cluster contrasts sharply with the second, comprising individuals who consistently spend high amounts across all product categories (groceries, milk, detergents, paper, fresh, delicatessen, and frozen goods). It includes both retail customers (43 individuals) and hotel/restaurant/café (HORECA) clients (49 individuals). The spending patterns suggest a group with high purchasing power, likely including upper-class individuals as well as premium commercial clients who require substantial quantities and quality products for their operations.

The fourth cluster represents individuals who spend above the median on groceries, detergents, paper, and milk but below the median on frozen and fresh products. This group primarily consists of HORECA customers, with a minority of retail customers (~30%). The focus on household products over fresh or frozen items suggests more utilitarian purchasing patterns, reflecting less refined or specialized consumption behavior compared to other clusters.

Overall, the first two clusters represent retail customers with contrasting spending patterns on specific product categories. The third cluster captures high-spending individuals and businesses, while the fourth characterizes utilitarian expenditure patterns in the HORECA sector.

2. Hierarchical clustering based on PCA components

First, the cluster variable and the Channel factor are strongly associated, as confirmed by the rejection of the chi-squared test's null hypothesis at all significance levels. This indicates a meaningful relationship between clusters and customer channels. Additionally, ANOVA tests demonstrate significant associations between all numerical variables (Detergents_Paper, Grocery, Milk, and Delicatessen) and the cluster variable, with Grocery and Detergents_Paper contributing the most to the variance across clusters.

Cluster 1 is primarily characterized by retail customers (over-represented), with 75% of its individuals coming from Channel 1 and 25% from Channel 2. Almost all of the individuals from Channel_1 and 69% from Channel_2 belong to this cluster, meaning Channel_2 is not underrepresented here. This cluster accounts for the majority of individuals from both channels, especially from Channel_1, and statistical tests show extreme significance (p-values around 10^{-21}).

In terms of quantitative variables, the first cluster shows much lower mean and standard deviation values across all significant products, such as milk, detergents, paper products, delicatessens, and general groceries. Among these, Grocery exhibits the largest deviation from the overall mean, indicating its role in defining this cluster. These clients are likely everyday household consumers who prioritize groceries and primary household products over luxury or specialty items. Therefore, they could represent families or general consumers with a focus on essential goods.

Cluster 2, on the other hand, is characterized by an over-representation of Channel_2, which makes up almost all of the individuals in this cluster. Customers in this cluster display significantly higher mean values for all quantitative variables, especially groceries, detergents_paper, milk, and delicatessen. These clients likely represent businesses such as restaurants, hotels, or large establishments, as indicated by their higher expenditures in these categories. The fact that "Delicatessen" also has a significantly higher mean suggests that these establishments may appeal to higher-end consumer preferences or offer more specialized products.

In summary, the PCA-based clustering effectively distinguishes between everyday household consumers in Cluster 1 and high-expenditure businesses in Cluster 2, providing clear and actionable insights into consumer segmentation.

Which of the above hierarchical clustering methods would you choose? Why?

Choosing between PCA-based and MCA-based hierarchical clustering involves considering both precision in representing data and the interpretability of main characteristics. PCA-based clustering has been effective for capturing broad patterns in numerical data, particularly when focusing on overall spending behavior. It helped identify high or low spenders, which can be useful for segmenting top revenue-generating customers or understanding broad spending trends, such as in Zipfian/Pareto distributions. However, PCA-based clustering produced broad clusters with less emphasis on specific individual characteristics.

On the other hand, MCA-based clustering provided us with a more detailed and interpretable segmentation of consumer behavior, such as distinguishing between high-spending individuals, utilitarian purchasers (i.e., HORECA), and retail customers.

Given these differences, **MCA-based clustering** is the preferred approach in this case. It generates four stable and representative clusters that align well with the underlying structure of the data and allows us to characterize the individuals better and group them given their behavior.