Section 10.2 | Exploratory Multivariate Analysis -CA

Mohammad Saqib Ansari

2023-12-22

Correspondence Analysis (CA) is a multivariate statistical technique used to explore and analyze relationships within categorical data. It's particularly useful for examining associations between categorical variables in a contingency table. Here's an explanation covering its basics, assumptions, uses, and disadvantages:

Basics of Correspondence Analysis:

1. Data Input:

- **Categorical Variables**: CA works with categorical data represented in a contingency table or a two-way frequency table.

2. Analysis:

- **Visualization of Relationships**: CA simplifies and visualizes relationships between categorical variables by projecting them onto a lower-dimensional space.
- **Dimension Reduction**: It reduces the dimensionality of the data while preserving the associations between variables.

3. Outputs:

- **Biplot**: The main output of CA is a biplot, which displays points representing categories of variables in a reduced space.

Assumptions of Correspondence Analysis:

- **Independence**: The technique assumes independence between categories within variables.
- **No Zero Marginals**: The absence of zero marginal totals is preferred to avoid numerical issues.

Uses of Correspondence Analysis:

1. Exploratory Analysis:

- **Identifying Patterns**: CA helps identify patterns and associations between categorical variables.
- **Visualization**: It provides a visual summary of relationships in categorical data.

2. Market Research:

- **Brand Association**: Analyzing associations between brands and customer demographics.
- **Product Preferences**: Understanding relationships between products and consumer characteristics.

3. Social Sciences:

- **Survey Data**: Analyzing survey responses, opinions, or behaviors across different demographic groups.
- **Text Analysis**: Analyzing word frequencies in textual data.

Disadvantages of Correspondence Analysis:

1. Data Limitations:

- **Limited to Categorical Data**: CA is applicable only to categorical variables and might not work well with continuous variables.
- **Sparse Data**: It might not perform well with extremely sparse data or when categories have zero counts.

2. Interpretation Challenges:

- **Complex Interpretation**: The interpretation of results from CA can be complex, especially when dealing with multiple variables or categories.
- **Subjectivity**: Interpreting the distances between points in the biplot might require subjective judgment.

Using R for Correspondence Analysis:

Certainly! Here's how you can structure an R Markdown document to conduct Correspondence Analysis (CA) using the `HairEyeColor` dataset from the `datasets` package, including explanations for each step:

Load necessary libraries

```r
# Check if packages are installed, if not, install them
if (!requireNamespace("ca", quietly = TRUE)) {
  install.packages("ca")
}

# Load required libraries
library(ca)
library(FactoMineR)
```

Step 1: Reading the dataset

```r
# Load the HairEyeColor dataset
data(HairEyeColor, package = "datasets")

# Assign dataset to 'datc'
datc <- HairEyeColor

# View the structure, summary, and first few rows of the dataset
str(datc)
```

```
##  'table' num [1:4, 1:4, 1:2] 32 53 10 3 11 50 10 30 10 25 ...
##  - attr(*, "dimnames")=List of 3
##   ..$ Hair: chr [1:4] "Black" "Brown" "Red" "Blond"
##   ..$ Eye : chr [1:4] "Brown" "Blue" "Hazel" "Green"
##   ..$ Sex : chr [1:2] "Male" "Female"
```

```
summary(datc)
```

```
## Number of cases in table: 592
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 164.92, df = 24, p-value = 5.321e-23
##  Chi-squared approximation may be incorrect
```

```
head(datc)
```

```
## , , Sex = Male
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
##
## , , Sex = Female
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    36    9     5     2
##   Brown    66   34    29    14
##   Red      16    7     7     7
##   Blond     4   64     5     8
```

Step 2: Choosing the active rows and columns (Not applicable for this dataset)

Step 3: Conduct the CA

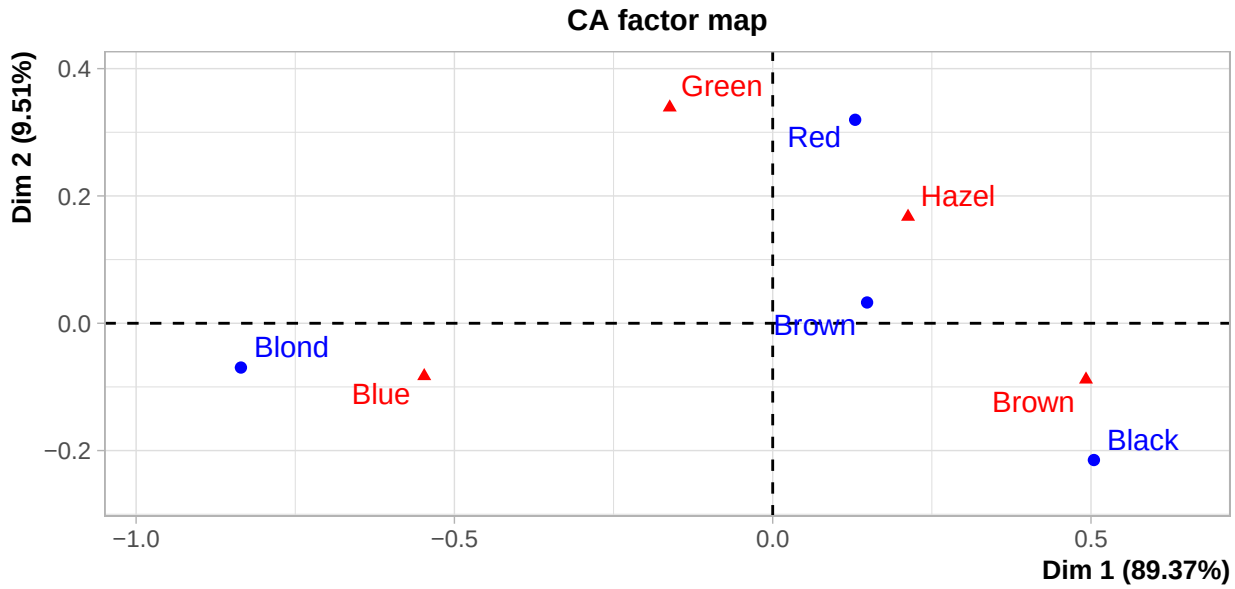
```
# Create a contingency table for Hair and Eye variables
datc <- xtabs(Freq ~ Hair + Eye, data = HairEyeColor)

# View the created contingency table
datc
```

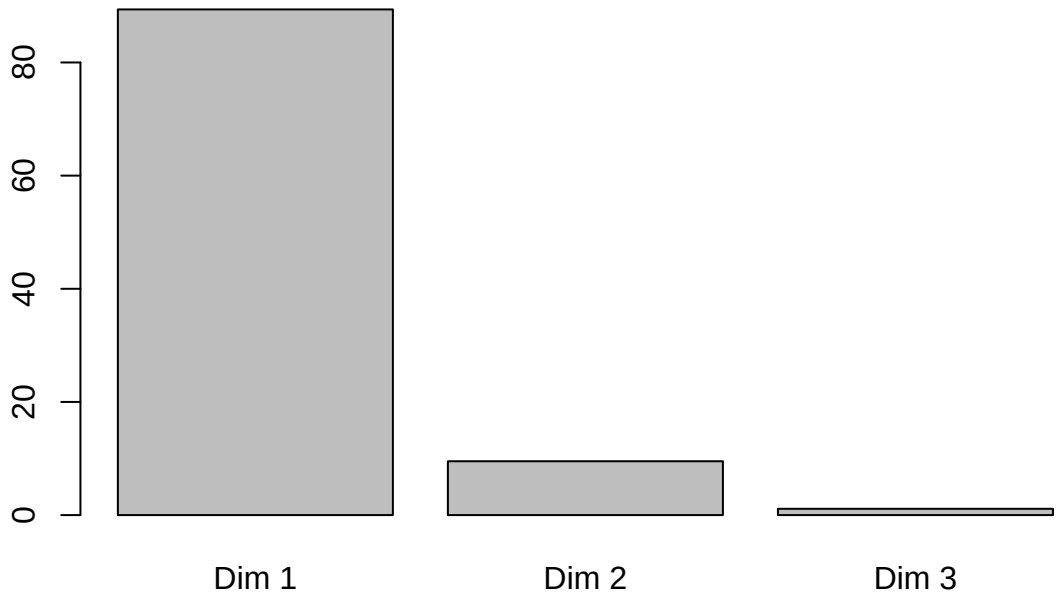
```
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    68   20    15     5
##   Brown   119   84    54    29
##   Red      26   17    14    14
##   Blond     7   94    10    16
```

Step 4: Compute Correspondence Analysis

```
# Perform Correspondence Analysis
res.ca <- CA(datc)
```




```
# Plotting scree plot to visualize eigenvalues
barplot(res.ca$eig[,2], names = paste("Dim", 1:nrow(res.ca$eig)))
```

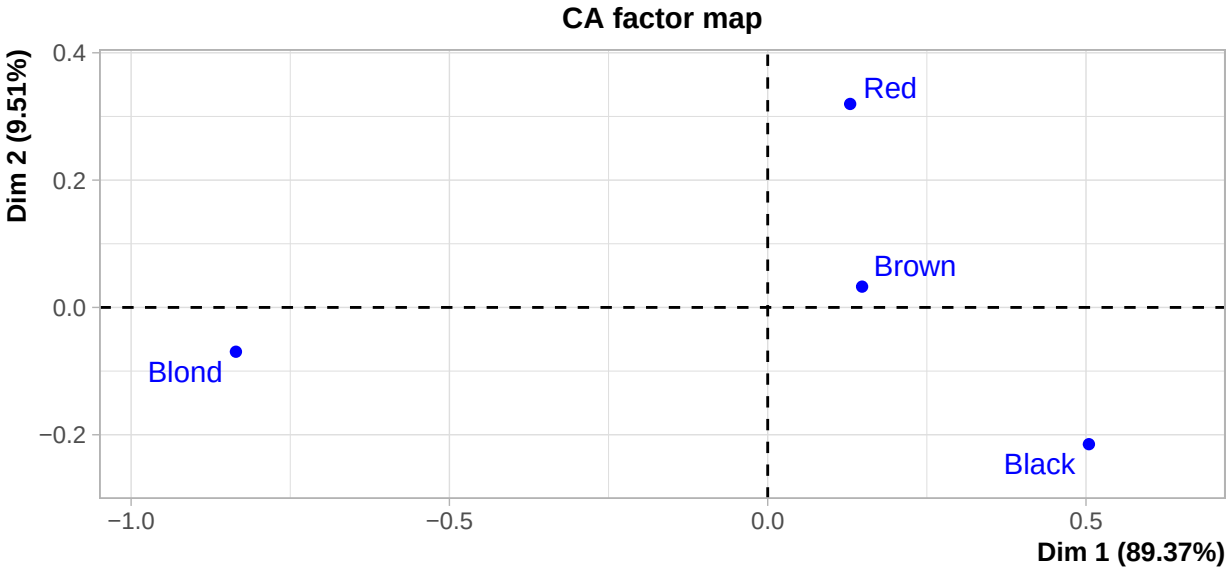



```
# Display eigenvalues and percentages of variance explained
round(res.ca$eig, 3)
```

```
##       eigenvalue percentage of variance
## dim 1      0.209                 89.373
## dim 2      0.022                  9.515
## dim 3      0.003                  1.112
##       cumulative percentage of variance
## dim 1                            89.373
## dim 2                            98.888
## dim 3                           100.000
```

Step 5: Analyzing the result

```
# Plotting the results: row and column coordinates
plot(res.ca, invisible = c("col", "col.sup"))
```



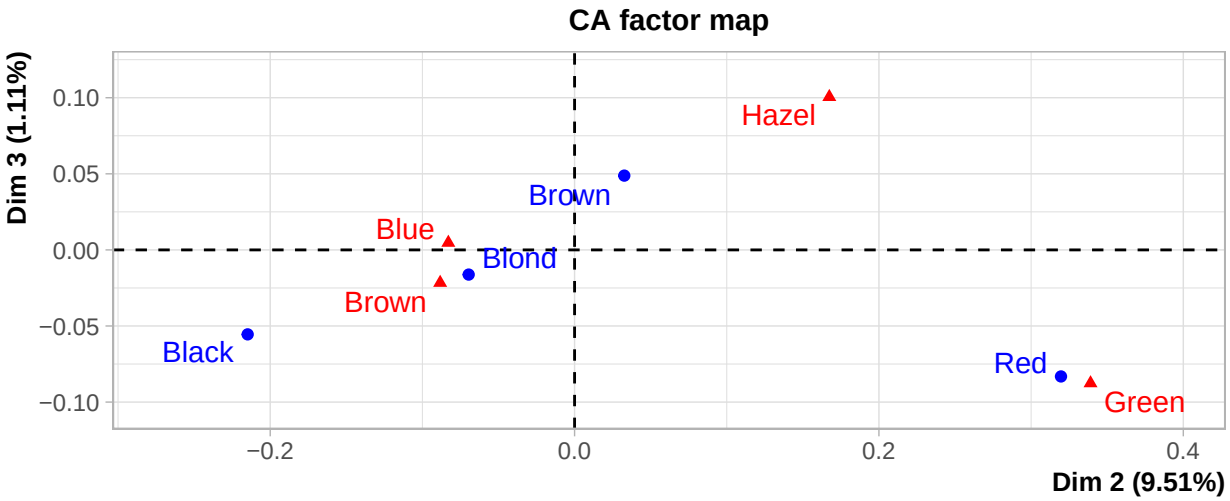

```
# Displaying row coordinates and quality of representation
round(cbind(res.ca$row$coord[, 1:3], res.ca$row$cos2[, 1:3]), 2)
```

```
##       Dim 1 Dim 2 Dim 3 Dim 1 Dim 2 Dim 3
## Black  0.50 -0.21 -0.06  0.84  0.15  0.01
## Brown  0.15  0.03  0.05  0.86  0.04  0.09
## Red    0.13  0.32 -0.08  0.13  0.81  0.05
## Blond -0.84 -0.07 -0.02  0.99  0.01  0.00
```

```
# Displaying column coordinates and quality of representation
round(cbind(res.ca$col$coord[, 1:3], res.ca$col$cos2[, 1:3]), 2)
```

```
##       Dim 1 Dim 2 Dim 3 Dim 1 Dim 2 Dim 3
## Brown  0.49 -0.09 -0.02  0.97  0.03  0.00
## Blue  -0.55 -0.08  0.00  0.98  0.02  0.00
## Hazel  0.21  0.17  0.10  0.54  0.34  0.12
## Green -0.16  0.34 -0.09  0.18  0.77  0.05
```

```
# Plotting dimensions 2 and 3
plot(res.ca, axes = 2:3)
```



In conclusion, Correspondence Analysis is a powerful tool for exploring relationships within categorical data, providing insights into associations between variables. However, its applicability depends on the nature and structure of the categorical data being analyzed.