Multivariate Analysis of Emergency Department Patient Satisfaction

Claudio Mazzi
Department of Computer Science, University of Pisa
MeS Laboratory, Sant'Anna School for Advanced Studies, Pisa
claudio.mazzi@santannapisa.it

Applied Statistical Modelling
A.Y. 2025–2026

1 Introduction

This document presents a complete multivariate analysis workflow applied to a real-world patient satisfaction survey collected in Emergency Departments (EDs). The dataset (`DATA_PS.dta`) contains responses from 5,919 patients to a structured questionnaire covering different domains of the ED experience, such as waiting times, comfort, communication, and evaluations of medical and nursing staff.

The overall objective of this toolkit is didactic: it shows how to combine *factor analysis* and *cluster analysis* to identify latent dimensions of satisfaction and empirically derive patient experience profiles. The main steps are:

- preprocessing and selection of the questionnaire items of interest;
- separate factor analyses for conceptually coherent blocks of items;
- extraction of domain-specific latent factor scores (waiting time, comfort, etc.);
- hierarchical clustering (Ward's method) on the factor scores;
- validation via k-means clustering and visualisation through PCA.

2 Data Description

The `DATA_PS.dta` file contains ED patient satisfaction data obtained from a structured survey. Each row corresponds to one patient; columns include administrative variables and a large number of questionnaire items. In particular, we focus on the original item variables with suffix `_orig`, which store the numeric codes of the responses (e.g. Likert-type scales).

For this exercise, we concentrate on the following groups of items:

- **Triage and waiting time**: `d009_orig`, `d010_orig`, `d016_orig`;
- **Comfort and cleanliness**: `d017_orig`, `d018_orig`;
- **Assistance and communication**: `d021_orig`, `d022_orig`, `d023_orig`, `d024_orig`;
- **Doctors**: `d026_orig`, `d027_orig`, `d028_orig`, `d029_orig`;
- **Nurses**: `d031_orig`, `d032_orig`, `d033_orig`, `d034_orig`.

Each group reflects a specific conceptual domain of ED experience and is measured by two to four ordinal items (e.g. 1–4, 1–6 or 1–7 response categories).

3 Methodological Framework

3.1 Factor Analysis

Factor Analysis (FA) is a multivariate technique used to model the covariance structure of a set of observed variables in terms of a smaller number of unobserved (latent) factors. In the common factor model, each observed variable X_j is expressed as:

$$X_j = \mu_j + \lambda_{j1}F_1 + \lambda_{j2}F_2 + \cdots + \lambda_{jm}F_m + \varepsilon_j, \tag{1}$$

where:

- μ_j is the mean of variable X_j,
- $F_1, \ldots, F_m$ are latent common factors,
- λ_{jk} are the factor loadings,
- ε_j is a specific error term unique to X_j.

In this application, we specify *one* common factor per domain (e.g. one factor for triage and waiting, one for comfort), and we are interested primarily in:

1. verifying that the items within each domain share a strong common dimension;
2. computing factor scores for each patient, representing their position along that latent dimension.

We use *principal axis factoring* (R function `fa()` with `fm = "pa"`). Because the items are ordinal and may contain missing values, we first compute a pairwise correlation matrix for each item block and then run FA on that correlation matrix.

3.2 From Items to Latent Domain Scores

For each domain, we:

1. select the corresponding `_orig` items;
2. compute a pairwise-complete Pearson correlation matrix;
3. perform one-factor FA using principal axis factoring;
4. obtain factor scores for each respondent.

The resulting five scores are:

- `attesa`: latent factor for triage and waiting time;
- `comfort`: latent factor for comfort and cleanliness;
- `assistenza`: latent factor for assistance and communication;
- `medici`: latent factor for doctor-related experience;
- `infermieri`: latent factor for nurse-related experience.

These scores are continuous variables and serve as a compact, noise-reduced representation of the original questionnaire items.

3.3 Hierarchical Cluster Analysis (Ward's Method)

Cluster analysis is an unsupervised learning technique whose goal is to partition a set of observations into groups (clusters) that are internally homogeneous and externally heterogeneous. Here we apply hierarchical agglomerative clustering with Ward's method.

In hierarchical agglomerative clustering:

- we start with each observation as its own cluster;
- at each step, we merge the two clusters that are most similar according to a linkage criterion;
- this process continues until all observations are merged into a single cluster.

Ward's method uses a sum-of-squares criterion: at each step, it merges the two clusters that result in the minimum increase in the total within-cluster sum of squares. This tends to produce clusters that are compact and spherical in shape. The results are visualised via a dendrogram, which displays the nested structure of clusters and helps determine a reasonable number of groups (e.g. by inspecting the height at which branches are merged).

3.4 k-means Clustering and PCA for Visualisation

To validate the structure suggested by hierarchical clustering, we also apply *k-means* clustering on the same factor score matrix. k-means is a partitioning algorithm that:

1. initializes k cluster centroids;
2. assigns each observation to the nearest centroid;
3. recomputes centroids as the mean of assigned points;
4. iterates until convergence.

In our case, both the dendrogram and heuristic criteria support a three-cluster solution, which we further explore using k-means with $k = 3$.

For visualisation, the function `fviz_cluster()` internally applies Principal Component Analysis (PCA) to reduce the five-dimensional factor score space into two principal components (PC1 and PC2). The resulting scatterplot shows cluster membership in this reduced space, where PC1 typically captures the main gradient of overall satisfaction across domains.

4 R Implementation

4.1 Loading the Dataset and Preparing Item Variables

Listing 1: Loading the dataset and preparing *_orig items

```r
library(haven)
library(dplyr)
library(psych)
library(factoextra)
library(cluster)

# Load the original dataset in Stata format
ps <- read_dta("DATA\_PS.dta")
```

```r

# Convert labelled variables into factors
ps <- ps %>% mutate(across(where(is.labelled), ~as_factor(.)))

# Convert *_orig item variables to numeric codes
ps_num <- ps %>%
  mutate(across(matches("^d[0-9]+_orig$"),
                ~as.numeric(as.character(.))))
```

4.2 Factor Analysis Helper Function and Domain Scores

Listing 2: Factor analysis helper function and factor scores

```r
# Helper function: 1-factor principal axis FA + factor scores
compute_factor <- function(data, vars) {
  # Pairwise-complete correlation matrix
  R <- cor(data[, vars], use = "pairwise.complete.obs")
  # One-factor principal axis FA
  fa_res <- fa(r = R, nfactors = 1, fm = "pa")
  # Factor scores for all observations
  scores <- factor.scores(data[, vars], fa_res)$scores[, 1]
  return(scores)
}

# TRIAGE & WAITING TIME: d009_orig, d010_orig, d016_orig
ps_num$attesa <- compute_factor(ps_num,
                                c("d009_orig", "d010_orig", "d016_orig"))

# COMFORT & CLEANLINESS: d017_orig, d018_orig
ps_num$comfort <- compute_factor(ps_num,
                                 c("d017_orig", "d018_orig"))

# ASSISTANCE & COMMUNICATION: d021_orig, d022_orig, d023_orig, d024_orig
ps_num$assistenza <- compute_factor(ps_num,
                                    c("d021_orig", "d022_orig",
                                      "d023_orig", "d024_orig"))

# DOCTORS: d026_orig, d027_orig, d028_orig, d029_orig
ps_num$medici <- compute_factor(ps_num,
                                c("d026_orig", "d027_orig",
                                  "d028_orig", "d029_orig"))

# NURSES: d031_orig, d032_orig, d033_orig, d034_orig
ps_num$infermieri <- compute_factor(ps_num,
                                    c("d031_orig", "d032_orig",
                                      "d033_orig", "d034_orig"))
```

4.3 Hierarchical Clustering and Cluster Summary

Listing 3: Hierarchical clustering and cluster means

```r
# Data for clustering: complete cases on the five factor scores
cluster_data <- ps_num %>%
  select(attesa, comfort, assistenza, medici, infermieri) %>%
  na.omit()

# Standardize factor scores
cluster_scaled <- scale(cluster_data)

# Euclidean distance matrix
dist_mat <- dist(cluster_scaled, method = "euclidean")

# Ward's hierarchical clustering
hc <- hclust(dist_mat, method = "ward.D2")

# Dendrogram
plot(hc,
     main = "Ward Hierarchical Clustering Dendrogram",
     xlab = "Observations", sub = "")
rect.hclust(hc, k = 3, border = "red")

# Cut tree into 3 clusters
clusters3 <- cutree(hc, k = 3)
cluster_data$cluster3 <- clusters3

# Mean factor scores by cluster
cluster_summary <- cluster_data %>%
  group_by(cluster3) %>%
  summarise(across(c(attesa, comfort, assistenza, medici, infermieri),
                   \(x) mean(x, na.rm = TRUE)))

print(cluster_summary)
```

The console output for the cluster means is:

```
# A tibble: 3 x 6
  cluster3  attesa comfort assistenza medici infermieri
     <int>   <dbl>   <dbl>      <dbl>  <dbl>      <dbl>
1        1  0.0327 -0.0527     -0.215 -0.101     -0.122
2        2  0.557   0.432       0.507  0.621      0.658
3        3 -1.21   -0.819      -0.654 -1.06      -1.11
```

4.4 k-means Clustering and Visualisation

Listing 4: k-means clustering and PCA-based cluster plot

```r
set.seed(123)
kmeans_res <- kmeans(cluster_scaled, centers = 3, nstart = 25)

# Visualisation: fviz_cluster applies PCA for plotting
fviz_cluster(kmeans_res,
             data = cluster_scaled,
             main = "K-means Clustering on Factor Scores")
```

In the resulting plot, 2, the horizontal axis (Dim1) corresponds to the first principal component, which captures the main gradient of overall satisfaction across the five domains. The vertical axis (Dim2) represents a secondary source of variation, typically less interpretable but useful to visually separate the clusters.

5 Results and Interpretation

5.1 Cluster Profiles

The mean factor scores by cluster (Section 3) allow a straightforward interpretation of the three groups:

- **Cluster 2: Highly Positive Experience**
  This group exhibits consistently positive scores across all five dimensions (waiting time, comfort, assistance, doctors, nurses). Patients in this cluster perceive shorter waiting times, better physical conditions, clear communication, and very positive interactions with both medical and nursing staff. This is the "best-experience" group.

- **Cluster 1: Moderately Negative / Average Experience**
  This group shows scores around zero or slightly negative, particularly for assistance and staff-related dimensions. Waiting time is close to the overall mean. These patients report a somewhat below-average experience, but not extremely negative. Their dissatisfaction is mainly related to relational and communication aspects rather than delays.

- **Cluster 3: Strongly Negative Experience**
  This cluster displays markedly negative scores on all dimensions, with the lowest values for waiting time and staff evaluations. Patients in this group perceive long waiting times, poor comfort, weak communication, and unsatisfactory interactions with doctors and nurses. It represents the most critical profile in terms of ED patient experience.

Both Ward's hierarchical clustering and k-means identify essentially the same structure: three well-separated groups ordered along a latent "overall satisfaction" dimension.

5.2 Dendrogram and k-means Plot

Figure 1 shows the Ward dendrogram. The height jumps between the last merges support a three-cluster solution, which corresponds to cutting the tree at an appropriate level (red rectangles).

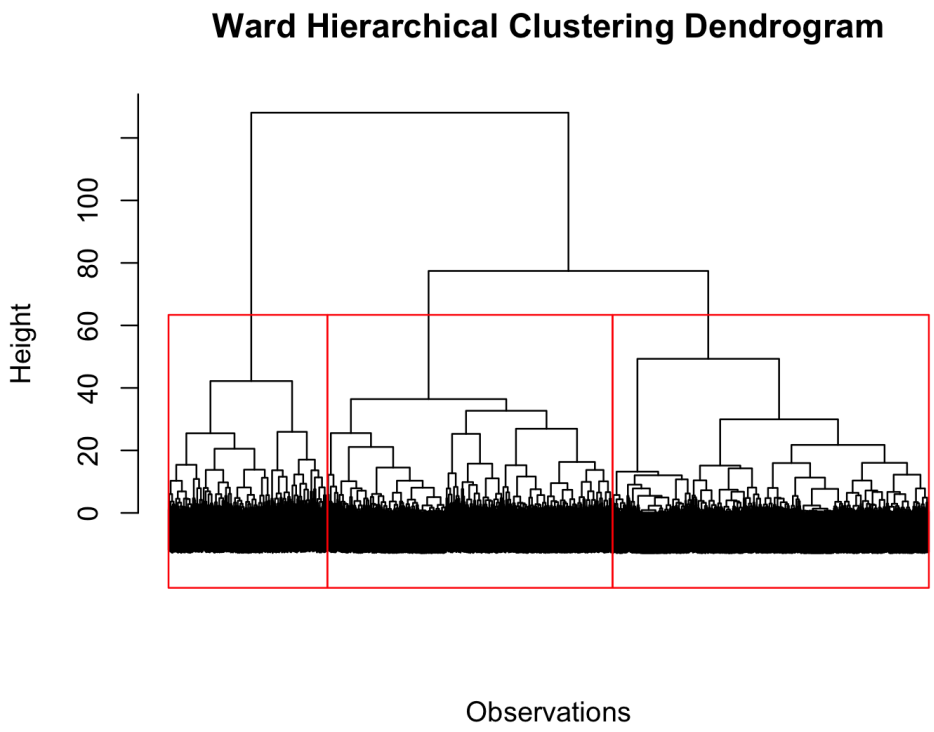

Figure 1: Ward hierarchical clustering dendrogram on the five factor scores.

Figure 2 reports the k-means clustering plot. The axes correspond to the first two principal components of the standardized factor scores. Dim1 typically reflects an overall satisfaction gradient (from negative to positive experiences), while Dim2 captures secondary variation. The three clusters are clearly separated in this low-dimensional representation.

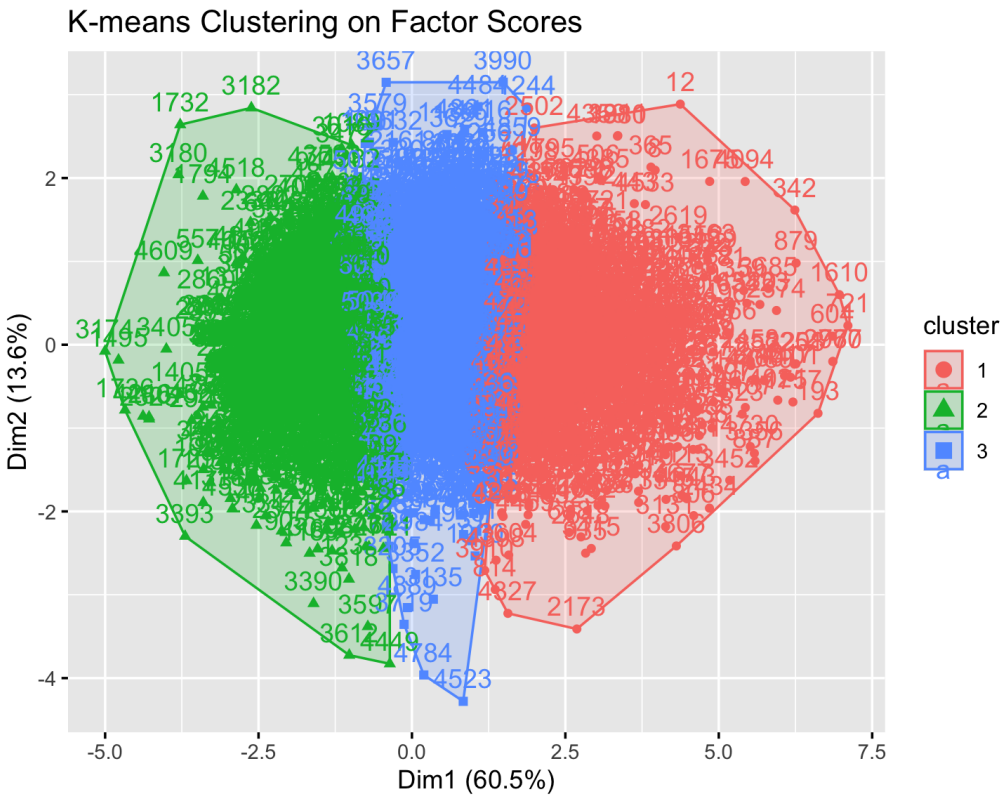

Figure 2: k-means clustering on the five factor scores (PCA projection).

6 Theoretical Summary

From a theoretical standpoint, the analysis illustrates how:

- **Factor Analysis** can be used to construct latent variables summarising blocks of correlated questionnaire items, improving robustness and interpretability;
- **Hierarchical clustering (Ward's method)** provides a data-driven way to identify homogeneous groups of patients based on their latent satisfaction profiles;
- **k-means clustering** confirms and refines the cluster solution suggested by hierarchical methods;
- **PCA** serves as a useful tool for visualising high-dimensional clustering results in a two-dimensional space, without altering the cluster assignments themselves.

Pedagogically, this example shows how unsupervised methods (FA, clustering, PCA) can be combined to derive meaningful segments of ED patient experience from complex, high-dimensional survey data.

7 Conclusions

This multivariate toolkit provides a complete, reproducible workflow for analysing patient satisfaction in the Emergency Department using real administrative and survey data. By moving from raw items to latent factors, and then from factors to clusters, we obtain a rich yet manageable representation of ED experience, which can be used for:

- monitoring and benchmarking patient satisfaction;

- identifying critical patient segments for targeted interventions;
- illustrating advanced multivariate methods in an applied teaching context.

Students are encouraged to extend the analysis by: adding covariates, exploring additional clustering solutions (e.g. $k = 4$), or linking cluster membership to clinical and organisational outcomes.