Chapter 10
Decomposition of Data Matrices by Factors

In Chap. 1 basic descriptive techniques were developed which provided tools for "looking" at multivariate data. They were based on adaptations of bivariate or univariate devices used to reduce the dimensions of the observations. In the following three chapters, issues of reducing the dimension of a multivariate data set will be discussed. The perspectives will be different but the tools will be related.

In this chapter, we take a descriptive perspective and show how using a geometrical approach provides a "best" way of reducing the dimension of a data matrix. It is derived with respect to a least-squares criterion. The result will be low dimensional graphical pictures of the data matrix. This involves the decomposition of the data matrix into "factors". These "factors" will be sorted in decreasing order of importance. The approach is very general and is the core idea of many multivariate techniques. We deliberately use the word "factor" here as a tool or transformation for structural interpretation in an exploratory analysis. In practice, the matrix to be decomposed will be some transformation of the original data matrix and as shown in the following chapters, these transformations provide easier interpretations of the obtained graphs in lower dimensional spaces.

Chapter 11 addresses the issue of reducing the dimensionality of a multivariate random variable by using linear combinations (the principal components). The identified principal components are ordered in decreasing order of importance. When applied in practice to a data matrix, the principal components will turn out to be the factors of a transformed data matrix (the data will be centred and eventually standardised).

Factor analysis is discussed in Chap. 12. The same problem of reducing the dimension of a multivariate random variable is addressed but in this case the number of factors is fixed from the start. Each factor is interpreted as a latent characteristic of the individuals revealed by the original variables. The non-uniqueness of the solutions is dealt with by searching for the representation with the easiest interpretation for the analysis.

Summarising, this chapter can be seen as a foundation since it develops a basic tool for reducing the dimension of a multivariate data matrix.

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-662-45171-7_10

10.1 The Geometric Point of View

As a matter of introducing certain ideas, assume that the data matrix $\mathcal{X}(n \times p)$ is composed of n observations (or individuals) of p variables.

There are in fact two ways of looking at $\mathcal{X}$, row by row or column by column:

1. Each row (observation) is a vector $x_i^\top = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$.
   From this point of view our data matrix $\mathcal{X}$ is representable as a cloud of n points in $\mathbb{R}^p$ as shown in Fig. 10.1.
2. Each column (variable) is a vector $x_{[j]} = (x_{1j}, \ldots, x_{nj})^\top \in \mathbb{R}^n$.
   From this point of view the data matrix $\mathcal{X}$ is a cloud of p points in $\mathbb{R}^n$ as shown in Fig. 10.2.

When n and/or p are large (larger than 2 or 3), we cannot produce interpretable graphs of these clouds of points. Therefore, the aim of the factorial methods to be developed here is twofold. We shall try to simultaneously approximate the column space $C(\mathcal{X})$ and the row space $C(\mathcal{X}^\top)$ with smaller subspaces. The hope is of course that this can be done without losing too much information about the variation and structure of the point clouds in both spaces. Ideally, this will provide insights into the structure of $\mathcal{X}$ through graphs in $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{R}^3$. The main focus then is to find the dimension reducing factors.

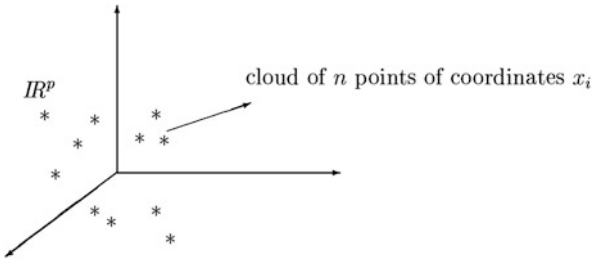


Fig. 10.1 Cloud of n points in $\mathbb{R}^p$

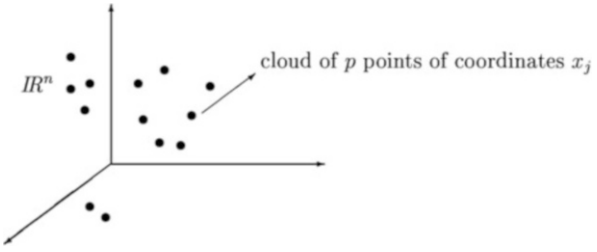


Fig. 10.2 Cloud of p points in $\mathbb{R}^n$

Summary

- ↪ Each row (individual) of $\mathcal{X}$ is a p-dimensional vector. From this point of view $\mathcal{X}$ can be considered as a cloud of n points in $\mathbb{R}^p$.
- ↪ Each column (variable) of $\mathcal{X}$ is a n-dimensional vector. From this point of view $\mathcal{X}$ can be considered as a cloud of p points in $\mathbb{R}^n$.

10.2 Fitting the p-Dimensional Point Cloud

Subspaces of Dimension 1

In this section $\mathcal{X}$ is represented by a cloud of n points in $\mathbb{R}^p$ (considering each row). The question is how to project this point cloud onto a space of lower dimension. To begin consider the simplest problem, namely finding a subspace of dimension 1. The problem boils down to finding a straight line F_1 through the origin. The direction of this line can be defined by a unit vector $u_1 \in \mathbb{R}^p$. Hence, we are searching for the vector u_1 which gives the "best" fit of the initial cloud of n points. The situation is depicted in Fig. 10.3.

The representation of the ith individual $x_i \in \mathbb{R}^p$ on this line is obtained by the projection of the corresponding point onto u_1, i.e. the projection point p_{x_i}. We know from (2.42) that the coordinate of x_i on F_1 is given by

$$p_{x_i} = x_i^\top \frac{u_1}{\|u_1\|} = x_i^\top u_1. \tag{10.1}$$

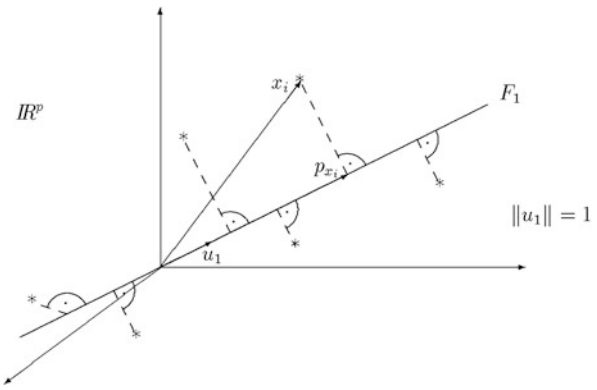


Fig. 10.3 Projection of point cloud onto u space of lower dimension

We define the *best line* F_1 in the following "least-squares" sense: Find $u_1 \in \mathbb{R}^p$ which minimises

$$\sum_{i=1}^{n} \|x_i - p_{x_i}\|^2. \tag{10.2}$$

Since $\|x_i - p_{x_i}\|^2 = \|x_i\|^2 - \|p_{x_i}\|^2$ by Pythagoras's theorem, the problem of minimising (10.2) is equivalent to maximising $\sum_{i=1}^{n} \|p_{x_i}\|^2$. Thus the problem is to find $u_1 \in \mathbb{R}^p$ that maximises $\sum_{i=1}^{n} \|p_{x_i}\|^2$ under the constraint $\|u_1\| = 1$. With (10.1) we can write

$$\begin{pmatrix} p_{x_1} \\ p_{x_2} \\ \vdots \\ p_{x_n} \end{pmatrix} = \begin{pmatrix} x_1^\top u_1 \\ x_2^\top u_1 \\ \vdots \\ x_n^\top u_1 \end{pmatrix} = \mathcal{X} u_1$$

and the problem can finally be reformulated as: find $u_1 \in \mathbb{R}^p$ with $\|u_1\| = 1$ that maximises the quadratic form $(\mathcal{X}u_1)^\top (\mathcal{X}u_1)$ or

$$\max_{u_1^\top u_1 = 1} u_1^\top (\mathcal{X}^\top \mathcal{X}) u_1. \tag{10.3}$$

The solution is given by Theorem 2.5 (using $\mathcal{A} = \mathcal{X}^\top \mathcal{X}$ and $\mathcal{B} = \mathcal{I}$ in the theorem).

Theorem 10.1 *The vector u_1 which minimises (10.2) is the eigenvector of $\mathcal{X}^\top \mathcal{X}$ associated with the largest eigenvalue λ_1 of $\mathcal{X}^\top \mathcal{X}$.*

Note that if the data have been centred, i.e. $\overline{x} = 0$, then $\mathcal{X} = \mathcal{X}_c$, where $\mathcal{X}_c$ is the centred data matrix, and $\frac{1}{n}\mathcal{X}^\top \mathcal{X}$ is the covariance matrix. Thus Theorem 10.1 says that we are searching for a maximum of the quadratic form (10.3) w.r.t. the covariance matrix $\mathcal{S}_\mathcal{X} = n^{-1}\mathcal{X}^\top \mathcal{X}$.

Representation of the Cloud on F_1

The coordinates of the n individuals on F_1 are given by $\mathcal{X}u_1$. $\mathcal{X}u_1$ is called the *first factorial variable* or the *first factor* and u_1 the *first factorial axis*. The n individuals, x_i, are now represented by a new factorial variable $z_1 = \mathcal{X}u_1$. This factorial variable is a linear combination of the original variables $(x_{[1]}, \ldots, x_{[p]})$ whose coefficients are given by the vector u_1, i.e.

$$z_1 = u_{11}x_{[1]} + \cdots + u_{p1}x_{[p]}. \tag{10.4}$$

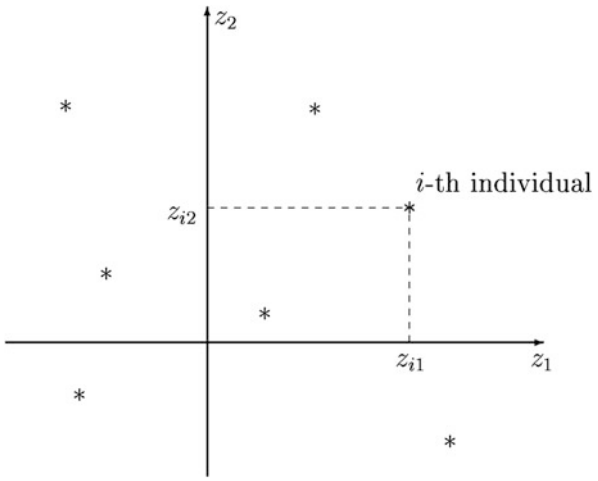


Fig. 10.4 Representation of the individuals $x_1, \ldots, x_n$ as a two-dimensional point cloud

Subspaces of Dimension 2

If we approximate the n individuals by a plane (dimension 2), it can be shown via Theorem 2.5 that this space contains u_1. The plane is determined by the best linear fit (u_1) and a unit vector u_2 orthogonal to u_1 which maximises the quadratic form $u_2^\top(\mathcal{X}^\top\mathcal{X})u_2$ under the constraints

$$\|u_2\| = 1, \text{ and } u_1^\top u_2 = 0.$$

Theorem 10.2 *The second factorial axis, u_2, is the eigenvector of $\mathcal{X}^\top\mathcal{X}$ corresponding to the second largest eigenvalue λ_2 of $\mathcal{X}^\top\mathcal{X}$.*

The unit vector u_2 characterises a second line, F_2, on which the points are projected. The coordinates of the n individuals on F_2 are given by $z_2 = \mathcal{X}u_2$. The variable z_2 is called the *second factorial variable* or *the second factor*. The representation of the n individuals in two-dimensional space ($z_1 = \mathcal{X}u_1$ vs. $z_2 = \mathcal{X}u_2$) is shown in Fig. 10.4.

Subspaces of Dimension q ($q \leq p$)

In the case of q dimensions the task is again to minimise (10.2) but with projection points in a q-dimensional subspace. Following the same argument as above, it can be shown via Theorem 2.5 that this best subspace is generated by $u_1, u_2, \ldots, u_q$, the orthonormal eigenvectors of $\mathcal{X}^\top\mathcal{X}$ associated with the corresponding eigenvalues

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$. The coordinates of the n individuals on the kth factorial axis, u_k, are given by the kth factorial variable $z_k = \mathcal{X}u_k$ for $k = 1, \ldots, q$. Each factorial variable $z_k = (z_{1k}, z_{2k}, \ldots, z_{nk})^\top$ is a linear combination of the original variables $x_{[1]}, x_{[2]}, \ldots, x_{[p]}$ whose coefficients are given by the elements of the kth vector $u_k : z_{ik} = \sum_{m=1}^{p} x_{im}u_{mk}$.

Summary

- ↪ The p-dimensional point cloud of individuals can be graphically represented by projecting each element into spaces of smaller dimensions.
- ↪ The first factorial axis is u_1 and defines a line F_1 through the origin. This line is found by minimising the orthogonal distances (10.2). The factor u_1 equals the eigenvector of $\mathcal{X}^\top\mathcal{X}$ corresponding to its largest eigenvalue. The coordinates for representing the point cloud on a straight line are given by $z_1 = \mathcal{X}u_1$.
- ↪ The second factorial axis is u_2, where u_2 denotes the eigenvector of $\mathcal{X}^\top\mathcal{X}$ corresponding to its second largest eigenvalue. The coordinates for representing the point cloud on a plane are given by $z_1 = \mathcal{X}u_1$ and $z_2 = \mathcal{X}u_2$.
- ↪ The factor directions $1, \ldots, q$ are $u_1, \ldots, u_q$, which denote the eigenvectors of $\mathcal{X}^\top\mathcal{X}$ corresponding to the q largest eigenvalues. The coordinates for representing the point cloud of individuals on a q-dimensional subspace are given by $z_1 = \mathcal{X}u_1, \ldots, z_q = \mathcal{X}u_q$.

10.3 Fitting the n-Dimensional Point Cloud

Subspaces of Dimension 1

Suppose that $\mathcal{X}$ is represented by a cloud of p points (variables) in $\mathbb{R}^n$ (considering each column). How can this cloud be projected into a lower dimensional space? We start as before with one dimension. In other words, we have to find a straight line G_1, which is defined by the unit vector $v_1 \in \mathbb{R}^n$, and which gives the best fit of the initial cloud of p points.

Algebraically, this is the same problem as above (replace $\mathcal{X}$ by $\mathcal{X}^\top$ and follow Sect. 10.2): the representation of the jth variable $x_{[j]} \in \mathbb{R}^n$ is obtained by the projection of the corresponding point onto the straight line G_1 or the direction v_1. Hence we have to find v_1 such that $\sum_{j=1}^{p} \|p_{x_{[j]}}\|^2$ is maximised, or equivalently, we have to find the unit vector v_1 which maximises $(\mathcal{X}^\top v_1)^\top(\mathcal{X}v_1) = v_1^\top(\mathcal{X}\mathcal{X}^\top)v_1$. The solution is given by Theorem 2.5.

Theorem 10.3 *v_1 is the eigenvector of $\mathcal{X}\mathcal{X}^\top$ corresponding to the largest eigenvalue μ_1 of $\mathcal{X}\mathcal{X}^\top$.*

Representation of the Cloud on G_1

The coordinates of the p variables on G_1 are given by $w_1 = \mathcal{X}^\top v_1$, the first factorial axis. The p variables are now represented by a linear combination of the original individuals $x_1, \ldots, x_n$, whose coefficients are given by the vector v_1, i.e. for $j = 1, \ldots, p$

$$w_{1j} = v_{11}x_{1j} + \cdots + v_{1n}x_{nj}. \tag{10.5}$$

Subspaces of Dimension q ($q \leq n$)

The representation of the p variables in a subspace of dimension q is done in the same manner as for the n individuals above. The best subspace is generated by the orthonormal eigenvectors $v_1, v_2, \ldots, v_q$ of $\mathcal{X}\mathcal{X}^\top$ associated with the eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_q$. The coordinates of the p variables on the kth factorial axis are given by the factorial variables $w_k = \mathcal{X}^\top v_k$, $k = 1, \ldots, q$. Each factorial variable $w_k = (w_{k1}, w_{k2}, \ldots, w_{kp})^\top$ is a linear combination of the original individuals $x_1, x_2, \ldots, x_n$ whose coefficients are given by the elements of the kth vector v_k : $w_{kj} = \sum_{m=1}^{n} v_{km}x_{mj}$. The representation in a subspace of dimension $q = 2$ is depicted in Fig. 10.5.

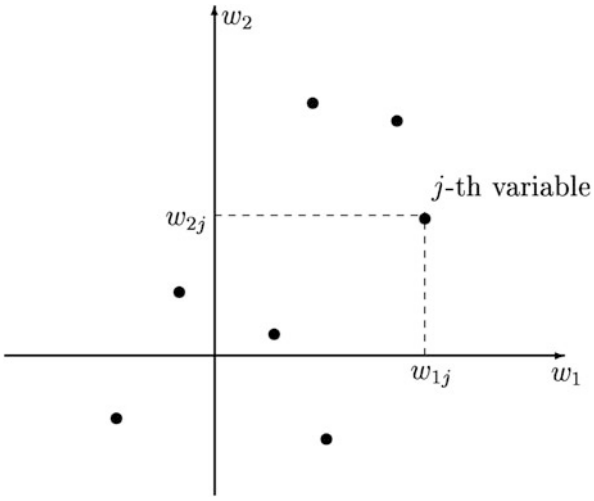


Fig. 10.5 Representation of the variables $x_{[1]}, \ldots, x_{[p]}$ as a two-dimensional point cloud

Summary

- The n-dimensional point cloud of variables can be graphically represented by projecting each element into spaces of smaller dimensions.
- The first factor direction is v_1 and defines a line G_1 through the origin. The vector v_1 equals the eigenvector of $\mathcal{X}\mathcal{X}^\top$ corresponding to the largest eigenvalue of $\mathcal{X}\mathcal{X}^\top$. The coordinates for representing the point cloud on a straight line are $w_1 = \mathcal{X}^\top v_1$.
- The second factor direction is v_2, where v_2 denotes the eigenvector of $\mathcal{X}\mathcal{X}^\top$ corresponding to its second largest eigenvalue. The coordinates for representing the point cloud on a plane are given by $w_1 = \mathcal{X}^\top v_1$ and $w_2 = \mathcal{X}^\top v_2$.
- The factor directions $1, \ldots, q$ are $v_1, \ldots, v_q$, which denote the eigenvectors of $\mathcal{X}\mathcal{X}^\top$ corresponding to the q largest eigenvalues. The coordinates for representing the point cloud of variables on a q-dimensional subspace are given by $w_1 = \mathcal{X}^\top v_1, \ldots, w_q = \mathcal{X}^\top v_q$.

10.4 Relations Between Subspaces

The aim of this section is to present a duality relationship between the two approaches shown in Sects. 10.2 and 10.3. Consider the eigenvector equations in $\mathbb{R}^n$

$$(\mathcal{X}\mathcal{X}^\top)v_k = \mu_k v_k \tag{10.6}$$

for $k \leq r$, where $r = \operatorname{rank}(\mathcal{X}\mathcal{X}^\top) = \operatorname{rank}(\mathcal{X}) \leq \min(p, n)$. Multiplying by $\mathcal{X}^\top$, we have

$$\mathcal{X}^\top(\mathcal{X}\mathcal{X}^\top)v_k = \mu_k \mathcal{X}^\top v_k \tag{10.7}$$

$$\text{or} \quad (\mathcal{X}^\top\mathcal{X})(\mathcal{X}^\top v_k) = \mu_k(\mathcal{X}^\top v_k) \tag{10.8}$$

so that each eigenvector v_k of $\mathcal{X}\mathcal{X}^\top$ corresponds to an eigenvector $(\mathcal{X}^\top v_k)$ of $\mathcal{X}^\top\mathcal{X}$ associated with the same eigenvalue μ_k. This means that every nonzero eigenvalue of $\mathcal{X}\mathcal{X}^\top$ is an eigenvalue of $\mathcal{X}^\top\mathcal{X}$. The corresponding eigenvectors are related by

$$u_k = c_k \mathcal{X}^\top v_k,$$

where c_k is some constant.

Now consider the eigenvector equations in $\mathbb{R}^p$:

$$(\mathcal{X}^\top \mathcal{X})u_k = \lambda_k u_k \tag{10.9}$$

for $k \leq r$. Multiplying by $\mathcal{X}$, we have

$$(\mathcal{X}\mathcal{X}^\top)(\mathcal{X}u_k) = \lambda_k(\mathcal{X}u_k), \tag{10.10}$$

i.e. each eigenvector u_k of $\mathcal{X}^\top\mathcal{X}$ corresponds to an eigenvector $\mathcal{X}u_k$ of $\mathcal{X}\mathcal{X}^\top$ associated with the same eigenvalue λ_k. Therefore, every nonzero eigenvalue of $(\mathcal{X}^\top\mathcal{X})$ is an eigenvalue of $\mathcal{X}\mathcal{X}^\top$. The corresponding eigenvectors are related by

$$v_k = d_k \mathcal{X}u_k,$$

where d_k is some constant. Now, since $u_k^\top u_k = v_k^\top v_k = 1$ we have $c_k = d_k = \frac{1}{\sqrt{\lambda_k}}$. This lead to the following result:

Theorem 10.4 (Duality Relations) *Let r be the rank of $\mathcal{X}$. For $k \leq r$, the eigenvalues λ_k of $\mathcal{X}^\top\mathcal{X}$ and $\mathcal{X}\mathcal{X}^\top$ are the same and the eigenvectors (u_k and v_k, respectively) are related by*

$$u_k = \frac{1}{\sqrt{\lambda_k}}\mathcal{X}^\top v_k \tag{10.11}$$

$$v_k = \frac{1}{\sqrt{\lambda_k}}\mathcal{X}u_k. \tag{10.12}$$

Note that the projection of the p variables on the factorial axis v_k is given by

$$w_k = \mathcal{X}^\top v_k = \frac{1}{\sqrt{\lambda_k}}\mathcal{X}^\top\mathcal{X}u_k = \sqrt{\lambda_k}\, u_k. \tag{10.13}$$

Therefore, the eigenvectors v_k do not have to be explicitly recomputed to get w_k.

Note that u_k and v_k provide the SVD of $\mathcal{X}$ (see Theorem 2.2). Letting $U = [u_1\ u_2\ \ldots\ u_r]$, $V = [v_1\ v_2\ \ldots\ v_r]$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$ we have

$$\mathcal{X} = V\ \Lambda^{1/2}\ U^\top$$

so that

$$x_{ij} = \sum_{k=1}^{r} \lambda_k^{1/2}\, v_{ik}\, u_{jk}. \tag{10.14}$$

In the following section this method is applied in analysing consumption behaviour across different household types.

Summary

↪ The matrices $\mathcal{X}^\top\mathcal{X}$ and $\mathcal{X}\mathcal{X}^\top$ have the same nonzero eigenvalues $\lambda_1, \ldots, \lambda_r$, where $r = \text{rank}(\mathcal{X})$.

↪ The eigenvectors of $\mathcal{X}^\top\mathcal{X}$ can be calculated from the eigenvectors of $\mathcal{X}\mathcal{X}^\top$ and vice versa:

$$u_k = \frac{1}{\sqrt{\lambda_k}}\mathcal{X}^\top v_k \quad \text{and} \quad v_k = \frac{1}{\sqrt{\lambda_k}}\mathcal{X}u_k.$$

↪ The coordinates representing the variables (columns) of $\mathcal{X}$ in a q-dimensional subspace can be easily calculated by $w_k = \sqrt{\lambda_k}u_k$.

10.5 Practical Computation

The practical implementation of the techniques introduced begins with the computation of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and the corresponding eigenvectors $u_1, \ldots, u_p$ of $\mathcal{X}^\top\mathcal{X}$. (Since p is usually less than n, this is numerically less involved than computing v_k directly for $k = 1, \ldots, p$.) The representation of the n individuals on a plane is then obtained by plotting $z_1 = \mathcal{X}u_1$ versus $z_2 = \mathcal{X}u_2$ ($z_3 = \mathcal{X}u_3$ may eventually be added if a third dimension is helpful). Using the Duality Relation (10.13) representations for the p variables can easily be obtained. These representations can be visualised in a scatterplot of $w_1 = \sqrt{\lambda_1}\,u_1$ against $w_2 = \sqrt{\lambda_2}u_2$ (and eventually against $w_3 = \sqrt{\lambda_3}\,u_3$). Higher dimensional factorial resolutions can be obtained (by computing z_k and w_k for $k > 3$) but, of course, cannot be plotted.

A standard way of evaluating the quality of the factorial representations in a subspace of dimension q is given by the ratio

$$\tau_q = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}, \tag{10.15}$$

where $0 \leq \tau_q \leq 1$. In general, the scalar product $y^\top y$ is called the inertia of $y \in \mathbb{R}^n$ w.r.t. the origin. Therefore, the ratio τ_q is usually interpreted as the percentage of the inertia explained by the first q factors. Note that $\lambda_j = (\mathcal{X}u_j)^\top(\mathcal{X}u_j) = z_j^\top z_j$. Thus, λ_j is the inertia of the jth factorial variable w.r.t. the origin. The denominator in (10.15) is a measure of the total inertia of the p variables, $x_{[j]}$. Indeed, by (2.3)

$$\sum_{j=1}^{p}\lambda_j = \text{tr}(\mathcal{X}^\top\mathcal{X}) = \sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2 = \sum_{j=1}^{p}x_{[j]}^\top x_{[j]}.$$

Remark 10.1 It is clear that the sum $\sum_{j=1}^{q} \lambda_j$ is the sum of the inertia of the first q factorial variables $z_1, z_2, \ldots, z_q$.

Example 10.1 We consider the data set in Table 22.6 which gives the food expenditures of various French families (manual workers = MA, employees = EM, managers = CA) with varying numbers of children (2, 3, 4 or 5 children). We are interested in investigating whether certain household types prefer certain food types. We can answer this question using the factorial approximations developed here.

The correlation matrix corresponding to the data is

$$\mathcal{R} = \begin{pmatrix} 1.00 & 0.59 & 0.20 & 0.32 & 0.25 & 0.86 & 0.30 \\ 0.59 & 1.00 & 0.86 & 0.88 & 0.83 & 0.66 & -0.36 \\ 0.20 & 0.86 & 1.00 & 0.96 & 0.93 & 0.33 & -0.49 \\ 0.32 & 0.88 & 0.96 & 1.00 & 0.98 & 0.37 & -0.44 \\ 0.25 & 0.83 & 0.93 & 0.98 & 1.00 & 0.23 & -0.40 \\ 0.86 & 0.66 & 0.33 & 0.37 & 0.23 & 1.00 & 0.01 \\ 0.30 & -0.36 & -0.49 & -0.44 & -0.40 & 0.01 & 1.00 \end{pmatrix}.$$

We observe a rather high correlation (0.98) between meat and poultry, whereas the correlation for expenditure for milk and wine (0.01) is rather small. Are there household types that prefer, say, meat over bread?

We shall now represent food expenditures and households simultaneously using two factors. First, note that in this particular problem the origin has no specific meaning (it represents a "zero" consumer). So it makes sense to compare the consumption of any family to that of an "average family" rather than to the origin. Therefore, the data is first centred (the origin is translated to the centre of gravity, $\overline{x}$). Furthermore, since the dispersions of the seven variables are quite different each variable is standardised so that each has the same weight in the analysis (mean 0 and variance 1). Finally, for convenience, we divide each element in the matrix by $\sqrt{n} = \sqrt{12}$. (This will only change the scaling of the plots in the graphical representation.)

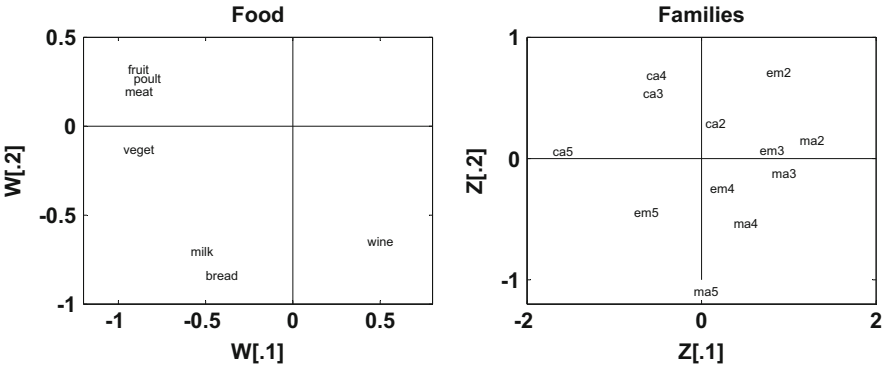
The data matrix to be analysed is

$$\mathcal{X}_* = \frac{1}{\sqrt{n}} \mathcal{H} \mathcal{X} \mathcal{D}^{-1/2},$$

where $\mathcal{H}$ is the centering matrix and $\mathcal{D} = \operatorname{diag}(s_{X_i X_i})$ (see Sect. 3.3). Note that by standardising by $\sqrt{n}$, it follows that $\mathcal{X}_*^\top \mathcal{X}_* = \mathcal{R}$ where $\mathcal{R}$ is the correlation matrix of the original data. Calculating

$$\lambda = (4.33, 1.83, 0.63, 0.13, 0.06, 0.02, 0.00)^\top$$

shows that the directions of the first two eigenvectors play a dominant role ($\tau_2 = 88\,\%$), whereas the other directions contribute less than 15 % of inertia. A two-dimensional plot should suffice for interpreting this data set.



bread	-0.49873	-0.84162
veget	-0.96975	-0.13310
fruit	-0.92913	0.27791
meat	-0.96210	0.19107
poult	-0.91125	0.26590
milk	-0.58434	-0.70690
wine	0.42820	-0.64815

ma2	1.12853	0.14353
em2	0.74582	0.70752
ca2	0.04654	0.28641
ma3	0.80563	-0.12763
em3	0.66887	0.06423
ca3	-0.66886	0.53487
ma4	0.36826	-0.54151
em4	0.09955	-0.24969
ca4	-0.63184	0.68533
ma5	-0.08726	-1.09621
em5	-0.77026	-0.44656
ca5	-1.70498	0.03971

Fig. 10.6 Representation of food expenditures and family types in two dimensions `MVAdecofood`

The coordinates of the projected data points are given in the two lower windows of Fig. 10.6. Let us first examine the food expenditure window. In this window we see the representation of the $p = 7$ variables given by the first two factors. The plot shows the factorial variables w_1 and w_2 in the same fashion as Fig. 10.4. We see that the points for meat, poultry, vegetables and fruits are close to each other in the lower left of the graph. The expenditures for bread and milk can be found in the upper left, whereas wine stands alone in the upper right. The first factor, w_1, may be interpreted as the meat/fruit factor of consumption, the second factor, w_2, as the bread/wine component.

In the lower window on the right-hand side, we show the factorial variables z_1 and z_2 from the fit of the $n = 12$ household types. Note that by the Duality Relations of Theorem 10.4, the factorial variables z_j are linear combinations of the factors w_k from the left window. The points displayed in the consumer window (graph on the right) are plotted relative to an average consumer represented by the origin. The manager families are located in the lower left corner of the graph whereas the manual workers and employees tend to be in the upper right. The factorial variables for CA5 (managers with five children) lie close to the meat/fruit factor. Relative to the average consumer this household type is a large consumer of meat/poultry and

fruits/vegetables. In Chap. 11, we will return to these plots interpreting them in a much deeper way. At this stage, it suffices to notice that the plots provide a graphical representation in $\mathbb{R}^2$ of the information contained in the original, high-dimensional (12×7) data matrix.

Summary

- The practical implementation of factor decomposition of matrices consists of computing the eigenvalues $\lambda_1, \ldots, \lambda_p$ and the eigenvectors $u_1, \ldots, u_p$ of $\mathcal{X}^\top \mathcal{X}$. The representation of the n individuals is obtained by plotting $z_1 = \mathcal{X}u_1$ vs. $z_2 = \mathcal{X}u_2$ (and, if necessary, vs. $z_3 = \mathcal{X}u_3$). The representation of the p variables is obtained by plotting $w_1 = \sqrt{\lambda_1}u_1$ vs. $w_2 = \sqrt{\lambda_2}u_2$ (and, if necessary, vs. $w_3 = \sqrt{\lambda_3}u_3$).
- The quality of the factorial representation can be evaluated using τ_q which is the percentage of inertia explained by the first q factors.

10.6 Exercises

Exercise 10.1 *Prove that $n^{-1}\mathcal{Z}^\top \mathcal{Z}$ is the covariance of the centred data matrix, where $\mathcal{Z}$ is the matrix formed by the columns $z_k = \mathcal{X}u_k$.*

Exercise 10.2 *Compute the SVD of the French food data (Table 22.6).*

Exercise 10.3 *Compute $\tau_3, \tau_4, \ldots$ for the French food data (Table 22.6).*

Exercise 10.4 *Apply the factorial techniques to the Swiss bank notes (Sect. 22.2).*

Exercise 10.5 *Apply the factorial techniques to the time budget data (Table 22.14).*

Exercise 10.6 *Assume that you wish to analyse p independent identically distributed random variables. What is the percentage of the inertia explained by the first factor? What is the percentage of the inertia explained by the first q factors?*

Exercise 10.7 *Assume that you have p i.i.d. r.v.'s. What does the eigenvector, corresponding to the first factor, look like.*

Exercise 10.8 *Assume that you have two random variables, X_1 and $X_2 = 2X_1$. What do the eigenvalues and eigenvectors of their correlation matrix look like? How many eigenvalues are nonzero?*

Exercise 10.9 *What percentage of inertia is explained by the first factor in the previous exercise?*

Exercise 10.10 *How do the eigenvalues and eigenvectors in Example 10.1 change if we take the prices in USD instead of in EUR? Does it make a difference if some of the prices are in EUR and others in USD?*