Chapter 11 Principal Components Analysis

Applied Multivariate Statistical Analysis

Course Instructor: Shuen-Lin Jeng

Department of Statistics
National Cheng Kung University

March, 10, 2021

11.1 Standard Linear Combination

Principal components analysis has the same objective with the exception that the rows of the data matrix $\mathcal{X}$ will now be considered as observations from a p-variate random variable X.

In the other words, one is searching for linear combinations with the largest variances.

A more flexible approach is to study a weighted average, namely,

$$\delta^T X = \sum_{j=1}^{p} \delta_j X_j, \text{ such that } \sum_{j=1}^{p} \delta_J^2 = 1 \tag{11.1}$$

The weighting vector $\delta = (\delta_1, \dots, \delta_p)^\mathsf{T}$ can then be optimized to investigate and to detect specific features. We call (11.1) a standardized linear combination (SLC).

11.1 Standard Linear Combination

One aim is to maximize the variance of the projection $\delta^\top X$,i.e., to choose δ according to

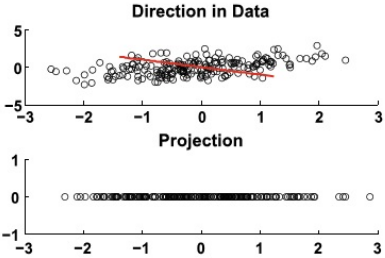
$$max_{\{\delta:\|\delta\|=1\}}\mathbf{Var}\left(\delta^\top X\right) = max_{\{\delta:\|\delta\|=1\}}\delta^\top \mathbf{Var}(X)\delta. \tag{11.2}$$

The interesting "directions" of δ are found through the spectral decomposition of the covariance matrix. Indeed, from Theorem 2.5, the direction δ is given by the eigenvector γ_1 corresponding to the largest eigenvalue λ_1 of the covariance matrix $\sum = \mathbf{Var}(X)$.

11.1 Standard Linear Combination

Fig. 11.1 An arbitrary SLC

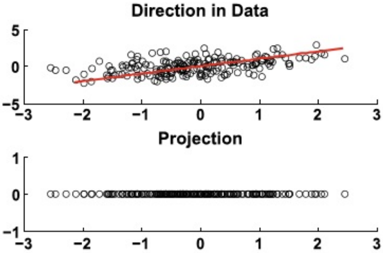
MVApcasimu



Explained variance	**0.50520**
Total variance	**1.96569**
Explained percentage	**0.25701**

11.1 Standard Linear Combination

Fig. 11.2 The most interesting SLC
MVApcasimu



Explained variance	1.46049
Total variance	1.96569
Explained percentage	0.74299

11.1 Standard Linear Combination

The SLC with the highest variance obtained from maximizing (11.2) is the first principal component(PC) $y_1 = \gamma_1^T X$. Orthogonal to the direction γ_1, we find the SLC with the second highest variance: $y_2 = {\gamma_2}^\mathsf{T} X$, the second PC.

Proceeding in this way and writing in matrix notation, the result for a random variable X with $\mathbf{E}(X) = \mu$ and $\mathbf{Var}(X) = \sum = \Gamma\Lambda\Gamma^\mathsf{T}$ is the PC transformation which defined as

$$Y = \Gamma^\mathsf{T}(X - \mu) \tag{11.3}$$

Here we have centered the variable X in order to obtain a zero mean PC variable Y.

11.1 Standard Linear Combination

Example 11.1 Consider a bivariate normal distribution $N(0, \sum)$ with $\sum = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and $\rho > 0$ (see Example 3.13). Recall that the eigenvalues of this matrix are $\lambda_1 = 1 + \rho$ and $\lambda_2 = 1 - \rho$ with corresponding eigenvectors

$$\gamma_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \gamma_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

The PC transformation is thus

$$Y = \Gamma^{\mathsf{T}}(X - \mu) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} X$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}.$$

11.1 Standard Linear Combination

So the first principal component is

$$Y_1 = \frac{1}{\sqrt{2}}(X_1 + X_2)$$

and the second is

$$Y_2 = \frac{1}{\sqrt{2}}(X_1 - X_2)$$

11.1 Standard Linear Combination

$\mathbf{Var}(Y_1) = \mathbf{Var}\left\{\frac{1}{\sqrt{2}}(X_1 + X_2)\right\} = \frac{1}{2}\mathbf{Var}(X_1 + X_2)$
$=\frac{1}{2}\{\mathbf{Var}(X_1) + \mathbf{Var}(X_2) + 2\mathbf{Cov}(X_1, X_2)\} = \frac{1}{2}(1 + 1 + 2\rho) = 1 + \rho = \lambda_1$
Similarly, we find that

$$\mathbf{Var}(Y_2) = \lambda_2$$

11.1 Standard Linear Combination

Theorem 11.1 *For a given $X \sim (\mu, \sum)$ let $Y = \Gamma^T(X-\mu)$ be the PC transformation, Then*

$$\mathbf{E}Y_j = 0,\ \mathrm{j} = 1, \dots, p \tag{11.4}$$

$$\mathbf{Var}(Y_j) = \lambda_j,\ \mathrm{j} = 1, \dots, p \tag{11.5}$$

$$\mathbf{Cov}(Y_i, Y_j) = 0,\ \mathrm{i} \neq j \tag{11.6}$$

$$\mathbf{Var}(Y_1) \geq \mathbf{Var}(Y_2) \geq \dots \geq \mathbf{Var}(Y_p) \geq 0 \tag{11.7}$$

$$\sum_{j=1}^{p} \mathbf{Var}(Y_j) = \mathrm{tr}\left(\sum\right) \tag{11.8}$$

$$\prod_{j=1}^{p} \mathbf{Var}(Y_j) = \left|\sum\right| \tag{11.9}$$

11.1 Standard Linear Combination

Theorem 11.2 *There exists no SLC that has larger variance than* $\lambda_1 = \mathbf{Var}(Y_1)$.

Theorem 11.3 *If* $Y = a^T X$ *is an SLC that is not correlated with the first* k *PCs of* X *, then the variance of* Y *is maximized by choosing it to be the* $(k+1)$*-st PC.*

11.2 Principal Components in Practice

If g_1 denotes the first eigenvector of $\mathcal{S}$, the first principal component is given by $y_1 = \left(\mathcal{X} - 1_n \bar{x}^{\top}\right) g_1$. More generally if $\mathcal{S} = \mathcal{G}\mathcal{L}\mathcal{G}^{\top}$ is the spectral decomposition of $\mathcal{S}$, then the PCs are obtained by

$$\mathcal{Y} = (\mathcal{X} - 1_n \bar{x}^{\top})\mathcal{G}. \tag{11.10}$$

11.2 Principal Components in Practice

Note that with the centering matrix $\mathcal{H} = \mathcal{I} - (n^{-1}1_n1_n^\top)$ and $\mathcal{H}1_n\bar{x}^\top = 0$, we can write

$\mathcal{S}_y = n^{-1}\mathcal{Y}^\top\mathcal{H}\mathcal{Y}$

$$= n^{-1}\mathcal{G}^\top(\mathcal{X}-1_n\bar{x}^\top)^\top\mathcal{H}(\mathcal{X}-1_n\bar{x}^\top)\mathcal{G} = n^{-1}\mathcal{G}^\top\mathcal{X}^\top\mathcal{H}\mathcal{X}\mathcal{G} = \mathcal{G}^\top\mathcal{S}\mathcal{G} = \mathcal{L} \quad (11.11)$$

where $\mathcal{L} = \mathrm{diag}(\ell_1, \dots, \ell_p)$ is the matrix of eigenvalues of $\mathcal{S}$.

Hence, the variance of y_i equals the eigenvalue ℓ_i!

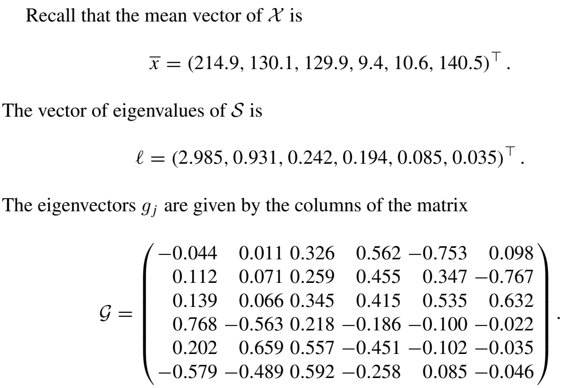
11.2 Principal Components in Practice

The PC techniques is sensitive to scale changes. If we multiply one variable by a scalar we obtain different eigenvalues and eigenvectors.

※ The PC transformation should be applied to data that have approximately the same scale in each variable.

Example 11.2 Let us apply this technique to the bank data set.
In this example, we do not standardize the data. Figure 11.3 shows some PC plots of the bank data set. The genuine and counterfeit bank notes are marked by "∘" and "+", respectively.

11.2 Principal Components in Practice



Recall that the mean vector of $\mathcal{X}$ is

$$\overline{x} = (214.9, 130.1, 129.9, 9.4, 10.6, 140.5)^\top .$$

The vector of eigenvalues of $\mathcal{S}$ is

$$\ell = (2.985, 0.931, 0.242, 0.194, 0.085, 0.035)^\top .$$

The eigenvectors g_j are given by the columns of the matrix

$$\mathcal{G} = \begin{pmatrix} -0.044 & 0.011 & 0.326 & 0.562 & -0.753 & 0.098 \\ 0.112 & 0.071 & 0.259 & 0.455 & 0.347 & -0.767 \\ 0.139 & 0.066 & 0.345 & 0.415 & 0.535 & 0.632 \\ 0.768 & -0.563 & 0.218 & -0.186 & -0.100 & -0.022 \\ 0.202 & 0.659 & 0.557 & -0.451 & -0.102 & -0.035 \\ -0.579 & -0.489 & 0.592 & -0.258 & 0.085 & -0.046 \end{pmatrix} .$$

The first column of $\mathcal{G}$ is the first eigenvector and gives the weights used in the linear combination of the original data in the first PC.

11.2 Principal Components in Practice

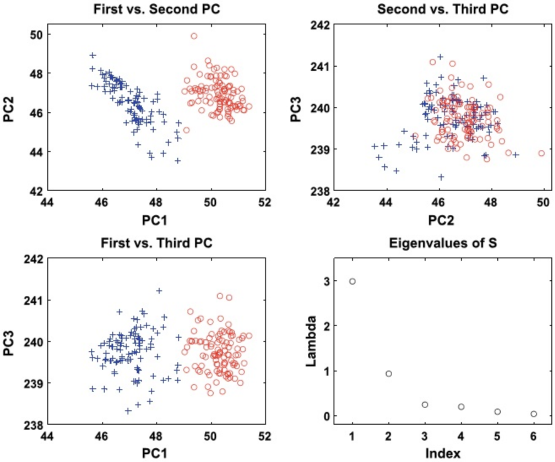


Fig. 11.3 Principal components of the bank data MVApcabank

11.2 Principal Components in Practice

Example 11.3 To see how sensitive the PCs are to a change in the scale of the variables, assume that X_1, X_2, X_3 ad X_6 are measured in *cm* and that X_4 and X_5 remain in *mm* in the bank data set.

11.2 Principal Components in Practice

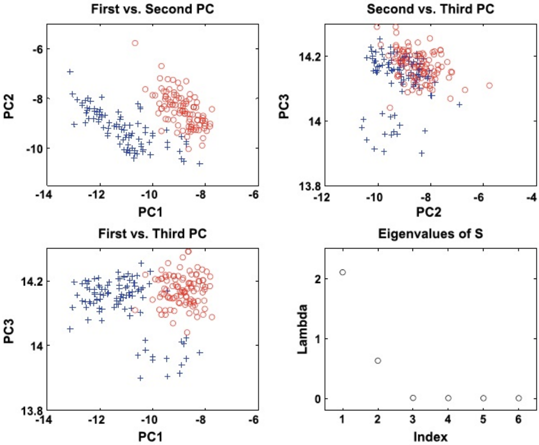


Fig. 11.4 Principal components of the rescaled bank data MVApcabankr

11.3 Interpretation of the PCs

In Sect.11.2, the eigenvectors were calculated for the bank data. In particular, with centered x's, we had

$$y_1 = -0.044x_1 + 0.112x_2 + 0.139x_3 + 0.768x_4 + 0.202x_5 - 0.579x_6$$

$$y_2 = 0.011x_1 + 0.071x_2 + 0.066x_3 - 0.563x_4 + 0.659x_5 - 0.489x_6$$

and

x_1 = length

x_2 = left height

x_3 = right height

x_4 = bottom frame

x_5 = top frame

x_6 = diagonal.

11.3 Interpretation of the PCs

Hence, the first PC is essentially the difference between the bottom frame variable and the diagonal. The second PC is best described by the difference between the top frame variable and the sum of bottom frame and diagonal variables.

The weighting of the PCs tells us in which directions, expressed in original coordinates, the best variance explanation is obtained. A measure of how well the first q PCs explain variation is given by the relative proportion:

$$\psi_q = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\sum_{j=1}^{q} \mathbf{Var}(Y_j)}{\sum_{j=1}^{p} \mathbf{Var}(Y_j)} \tag{11.12}$$

11.3 Interpretation of the PCs

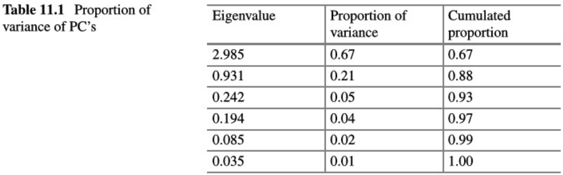


Table 11.1 Proportion of variance of PC's

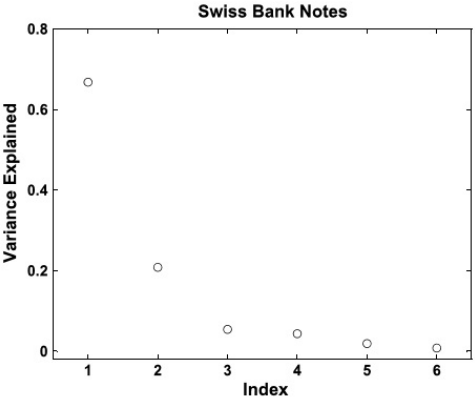
Eigenvalue	Proportion of variance	Cumulated proportion
2.985	0.67	0.67
0.931	0.21	0.88
0.242	0.05	0.93
0.194	0.04	0.97
0.085	0.02	0.99
0.035	0.01	1.00

The first PC ($q = 1$) already explains 67% of the variation.

The first three ($q = 3$) PCs explain 93% of the variation.

11.3 Interpretation of the PCs

Fig. 11.5 Relative proportion of variance explained by PCs MVApcabanki



11.3 Interpretation of the PCs

Hence, the correlation, $\rho_{X_iY_j}$, between variable X_i and the PC Y_j is

$$\rho_{X_iY_j} = \frac{\gamma_{ij}\lambda_j}{(\sigma_{X_iX_i}\lambda_j)^{1/2}} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma_{X_iX_i}}\right)^{1/2}. \tag{11.14}$$

$$r_{X_iY_j} = g_{ij}\left(\frac{\ell_j}{s_{X_iX_i}}\right)^{1/2} \tag{11.15}$$

11.3 Interpretation of the PCs

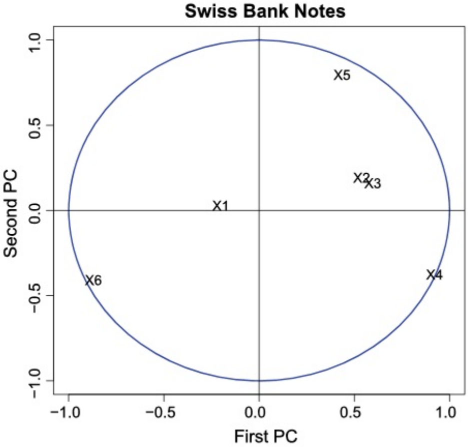
Note that

$$\sum_{j=1}^{p} r_{X_iY_j}^2 = \frac{\sum_{j=1}^{p} \ell_j g_{ij}^2}{s_{X_iX_i}} = \frac{s_{X_iX_i}}{s_{X_iX_i}} = 1 \tag{11.16}$$

Indeed, $\sum_{j=1}^{p} \ell_j g_{ij}^2 = g_i^{\top} \mathcal{L} g_i$ is the (i,i)-element of the matrix $\mathcal{G}\mathcal{L}\mathcal{G}^{\top} = \mathcal{S}$, so that $r_{X_iY_j}^2$ may be seen as the proportion of variance of X_i explained by Y_j.

11.3 Interpretation of the PCs

Fig. 11.6 The correlation of the original variable with the PCs MVApcabanki



11.3 Interpretation of the PCs

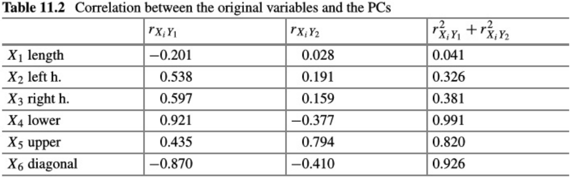


Table 11.2 Correlation between the original variables and the PCs

	$r_{X_i Y_1}$	$r_{X_i Y_2}$	$r^2_{X_i Y_1} + r^2_{X_i Y_2}$
X_1 length	−0.201	0.028	0.041
X_2 left h.	0.538	0.191	0.326
X_3 right h.	0.597	0.159	0.381
X_4 lower	0.921	−0.377	0.991
X_5 upper	0.435	0.794	0.820
X_6 diagonal	−0.870	−0.410	0.926

The correlations of the original variables X_i and the first two PCs are given in Table 11.2 along with the cumulated percentage of variance of each variable explained by Y_1 and Y_2.

11.4 Asymptotic Properties of the PCs

Theorem 11.4 *Let* $\sum > 0$ *with distinct eigenvalues, and let* $\mathcal{S} \sim n^{-1} W_p(\sum, n-1)$ *with spectral decompositions* $\sum = \Gamma\Lambda\Gamma^T$ *, and* $\mathcal{S} = \mathcal{G}\mathcal{L}\mathcal{G}^T$. *Then*

a. $\sqrt{n-1}(\ell - \lambda) \overset{\mathcal{L}}{\rightarrow} N_p(0, 2\Lambda^2)$,
where $\ell = (\ell_1, \ldots, \ell_p)^T$ *and* $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ *as the diagonals of* $\mathcal{L}$ *and* Λ,

b. $\sqrt{n-1}(g_j - \gamma_j) \overset{\mathcal{L}}{\rightarrow} N_p(0, \mathcal{V}_j)$, *with* $\mathcal{V}_j = \lambda_j \sum_{k \neq j} \frac{\lambda_k}{(\lambda_k - \lambda_j)^2} \gamma_k \gamma_k^T$,

c. the elements in ℓ *are asymptotically independent of the elements in* $\mathcal{G}$.

11.5 Normalized Principal Components Analysis

When the variables are measured in heterogeneous scales (such as years, kilograms, dollars, etc.). A standardization of the variables, namely,

$$\mathcal{X}_s = \mathcal{H}\mathcal{X}\mathcal{D}^{-1/2} \tag{11.19}$$

where $\mathcal{D} = \text{diag}(s_{X_1X_1}, \ldots, s_{X_pX_p})$. Note that $\bar{x}_S = 0$ and $\mathcal{S}_{\mathcal{X}_s} = \mathcal{R}$, the correlation matrix of $\mathcal{X}$.

11.5 Normalized Principal Components Analysis

The PC transformations of the matrix $\mathcal{X}_s$ are referred to as the *Normalized Principal Components*(NPCs). The spectral decomposition of $\mathcal{R}$ is

$$\mathcal{R} = \mathcal{G}_{\mathcal{R}} \mathcal{L}_{\mathcal{R}} \mathcal{G}_{\mathcal{R}}^{\top} \tag{11.20}$$

where $\mathcal{L}_R = \text{diag}(\ell_1^{\mathcal{R}}, \dots, \ell_p^{\mathcal{R}})$ and $\ell_1^{\mathcal{R}} \geq \dots \geq \ell_p^{\mathcal{R}}$ are eigenvalues of $\mathcal{R}$ with corresponding eigenvectors $g_1^{\mathcal{R}}, \dots, g_p^{\mathcal{R}}$ (note that here $\sum_{j=1}^{p} \ell_j^R = \text{tr}(\mathcal{R}) = p$). The NPCs, Z_j, provide a representation of each individual and are given by

$$\mathcal{Z} = \mathcal{X}_{\mathcal{S}} \mathcal{G}_{\mathcal{R}} = (z_1, \dots, z_p) \tag{11.21}$$

11.5 Normalized Principal Components Analysis

After transforming the variables, once again, we have that

$$\bar{z} = 0 \tag{11.22}$$

$$\mathcal{S}_{\mathcal{Z}} = \mathcal{G}_{\mathcal{R}}^{\top} \mathcal{S}_{\mathcal{X}_S} \mathcal{G}_{\mathcal{R}} = \mathcal{G}_{\mathcal{R}}^{\top} \mathcal{R} \mathcal{G}_{\mathcal{R}} = \mathcal{L}_{\mathcal{R}} \tag{11.23}$$

The NPCs provide a perspective similar to that of the PCs, but in terms of the relative position of individuals NPC gives each variable the same weight (with the PCs the variable with the largest variance received the largest weight).

11.6 Principal Components as a Factorial Method

The empirical PCs (normalizes or not) turn out to be equivalent to the factors that one would obtain by decomposing the appropriate data matrix into its factors(see Chap.10). It will be shown that the PCs are the factors representing the rows of the centered data matrix and that the NPCs correspond to the factors of the standardized data matrix.

Assume, as in Chap 10, that we want to obtain representations of the individuals (the rows of $\mathcal{X}$) and of the variables (the column of $\mathcal{X}$) in spaces of smaller dimension.

11.6 Principal Components as a Factorial Method

We will first shift the origin to the center of gravity, $\bar{x}$, of the point cloud. This is the same as analyzing the centered data matrix $\mathcal{X}_c = \mathcal{H}\mathcal{X}$. Note that the spectral decomposition of $\mathcal{X}_{\mathcal{C}}^{\top}\mathcal{X}_{\mathcal{C}}$ is related to that of $\mathcal{S}_X$. namely,

$$\mathcal{X}_{\mathcal{C}}^{\top}\mathcal{X}_{\mathcal{C}} = \mathcal{X}^{\top}\mathcal{H}^{\top}\mathcal{H}\mathcal{X} = n\mathcal{S}_X = n\mathcal{G}\mathcal{L}\mathcal{G}^{\top} \tag{11.28}$$

The factorial variables are obtained by projecting $\mathcal{X}_C$ in $\mathcal{G}$,

$$\mathcal{Y} = \mathcal{X}_{\mathcal{C}}\mathcal{G} = (y_1, \dots, y_p) \tag{11.29}$$

11.6 Principal Components as a Factorial Method

(Note that the y's here correspond to the z's in Sect.10.2.) Since $\mathcal{H}\mathcal{X}_{\mathcal{C}} = \mathcal{X}_{\mathcal{C}}$, it immediately follows that

$$\bar{y} = 0 \tag{11.30}$$

$$\mathcal{S}_Y = \mathcal{G}^{\top}\mathcal{S}_{\mathcal{X}}\mathcal{G} = \mathcal{L} = \mathrm{diag}(\ell_1, \ldots, \ell_p) \tag{11.31}$$

The scatterplot of the individuals on the factorial axes are thus centered around the origin and are more spread out in the first direction (first PC has variance ℓ_1) than in the second direction (second PC has variance ℓ_2)

11.6 Principal Components as a Factorial Method

Considering the geometric representation, there is a nice statistical interpretation of the angle between two columns of $\mathcal{X}_C$. Given that

$$x_{C[j]}^{\top} x_{C[k]} = n s_{X_j X_k} \tag{11.34}$$

$$\left\| x_{C[j]} \right\|^2 = n s_{X_j X_j} \tag{11.35}$$

where $x_{C[j]}$ and $x_{C[k]}$ denote the j-th and k-th column of $\mathcal{X}_C$.

11.6 Principal Components as a Factorial Method

It holds that in the full space of the variables , if θ_{jk} is the angle between two variables, $x_{C[j]}$ and $x_{C[k]}$, then

$$\cos\theta_{jk} = \frac{x_{C[j]}^{\top} x_{C[k]}}{\left\| x_{C[j]} \right\| \left\| x_{C[k]} \right\|} = r_{X_j X_k} \tag{11.36}$$

The NPCs can also be viewed as a factorial method for reducing the dimension.

Quality of the Representations

As said before, an overall measure of the quality of the representation is given by

$$\psi = \frac{\ell_1 + \ell_2 + \ldots + \ell_q}{\sum_{j=1}^{p} \ell_j}$$

11.6 Principal Components as a Factorial Method

The values $\cos^2 \vartheta_{ik}$ are sometimes called relative contributions of the k-th axis to the representation of the i-th individual, e.g., if $\cos^2 \vartheta_{i1} + \cos^2 \vartheta_{i2}$ is large (near one), we know that the individual i is well represented on the plane of the first two principal axes since its corresponding angle with the plane is close to zero.

Example 11.6

Example 11.6 Let us return to the French food expenditure example, see Appendix B.6. This yields a two-dimensional representation of the individuals as shown in Fig. 11.7. Calculating the matrix $\mathcal{G}_{\mathcal{R}}$ which gives the weights of the variables (milk, vegetables, etc.).

$$\mathcal{G}_{\mathcal{R}} = \begin{pmatrix} -0.240 & 0.622 & -0.011 & -0.544 & 0.036 & 0.508 \\ -0.466 & 0.098 & -0.062 & -0.023 & -0.809 & -0.301 \\ -0.446 & -0.205 & 0.145 & 0.548 & -0.067 & 0.625 \\ -0.462 & -0.141 & 0.207 & -0.053 & 0.411 & -0.093 \\ -0.438 & -0.197 & 0.356 & -0.324 & 0.224 & -0.350 \\ -0.281 & 0.523 & -0.444 & 0.450 & 0.341 & -0.332 \\ 0.206 & 0.479 & 0.780 & 0.306 & -0.069 & -0.138 \end{pmatrix},$$

Example 11.6

The interpretation of the principal components is best understood when looking at the correlations between the original X_i 's and the PCs. Since the first two PCs explain 88.1% of the variance, we limit ourselves to the first two PCs. The results are shown in Table 11.3.

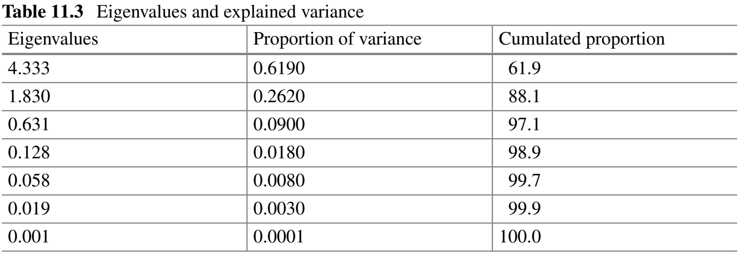
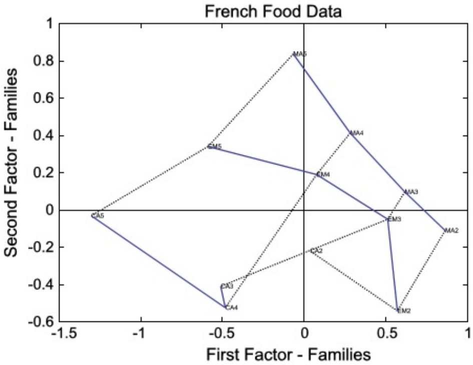


Table 11.3 Eigenvalues and explained variance

Eigenvalues	Proportion of variance	Cumulated proportion
4.333	0.6190	61.9
1.830	0.2620	88.1
0.631	0.0900	97.1
0.128	0.0180	98.9
0.058	0.0080	99.7
0.019	0.0030	99.9
0.001	0.0001	100.0

Example 11.6

Fig. 11.7 Representation of the individuals MVAnpcafood



Example 11.6

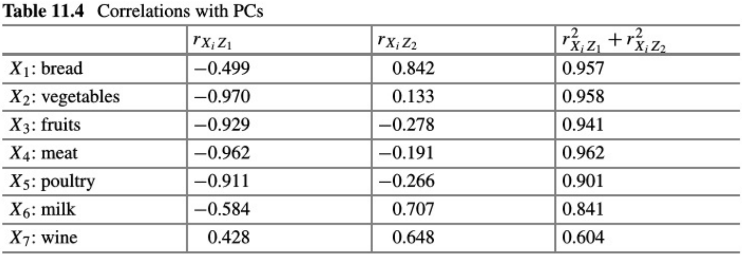
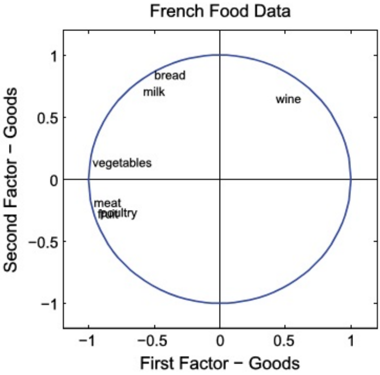


Table 11.4 Correlations with PCs

	$r_{X_i Z_1}$	$r_{X_i Z_2}$	$r^2_{X_i Z_1} + r^2_{X_i Z_2}$
X_1: bread	−0.499	0.842	0.957
X_2: vegetables	−0.970	0.133	0.958
X_3: fruits	−0.929	−0.278	0.941
X_4: meat	−0.962	−0.191	0.962
X_5: poultry	−0.911	−0.266	0.901
X_6: milk	−0.584	0.707	0.841
X_7: wine	0.428	0.648	0.604

Example 11.6

Fig. 11.8 Representation of the variables MVAnpcafood



Example 11.6

- Since the quality of the representation in Figure 11.8 is good for all the variables (except maybe X_7), their relative angles give a picture of their original correlation.
- wine is negatively correlated with the vegetables, fruits, meat, and poultry groups ($\theta > 90$), whereas taken individually this latter grouping of variables are highly positively correlated with each other($\theta \approx 0$).
- Bread and milk are positively correlated but poorly correlated with meat, fruits and poultry ($\theta \approx 90$).

Example 11.6

- Now the representation of the individuals in Fig. 11.7 can be interpreted better. From Fig. 11.8 and Table 11.3, we can see that the first factor Z_1 is a vegetable–meat– poultry–fruit factor (with a negative sign), whereas the second factor is a milk–bread– wine factor (with a positive sign).
- Note that this corresponds to the most important weights in the first columns of $\mathcal{G}_{\mathcal{R}}$.
- In Fig. 11.7, lines were drawn to connect families of the same size and families of the same professional types. A grid can clearly be seen (with a slight deformation by the manager families) that shows the families with higher expenditures (higher number of children) on the left.

Example 11.6

- Considering both figures together explains what types of expenditures are responsible for similarities in food expenditures.
- Bread, milk, and wine expenditures are similar for manual workers and employees.
- Families of managers are characterized by higher expenditures on vegetables, fruits, meat, and poultry.
- Very often when analyzing NPCs (and PCs), it is illuminating to use such a device to introduce qualitative aspects of individuals in order to enrich the interpretations of the graphs.

11.8 Boston Housing

The variable X_4 is dropped because it is a discrete 0-1 variable.
The scale difference of the remaining 13 variables motivates an NPCA based on the correlation matrix.

11.8 Boston Housing

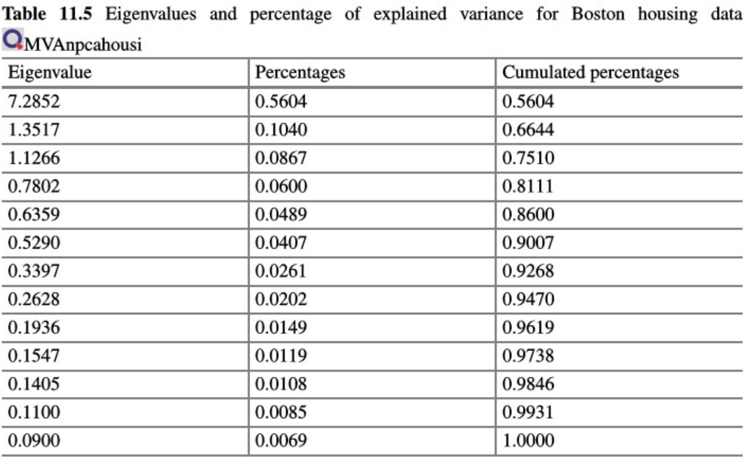


Table 11.5 Eigenvalues and percentage of explained variance for Boston housing data MVAnpcahousi

Eigenvalue	Percentages	Cumulated percentages
7.2852	0.5604	0.5604
1.3517	0.1040	0.6644
1.1266	0.0867	0.7510
0.7802	0.0600	0.8111
0.6359	0.0489	0.8600
0.5290	0.0407	0.9007
0.3397	0.0261	0.9268
0.2628	0.0202	0.9470
0.1936	0.0149	0.9619
0.1547	0.0119	0.9738
0.1405	0.0108	0.9846
0.1100	0.0085	0.9931
0.0900	0.0069	1.0000

11.8 Boston Housing

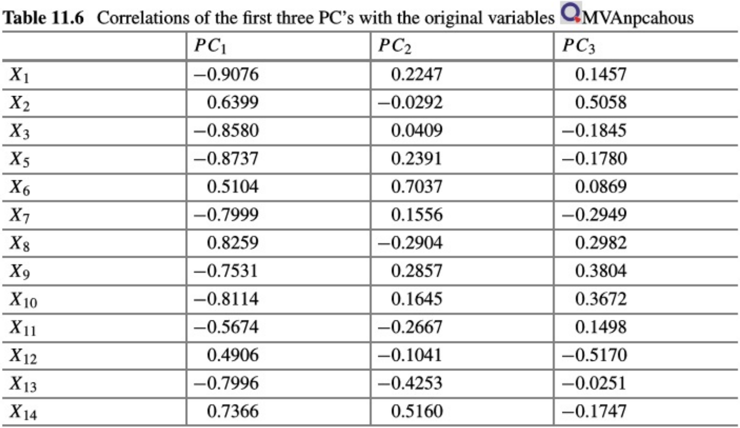


Table 11.6 Correlations of the first three PC's with the original variables MVAnpcahous

	PC_1	PC_2	PC_3
X_1	−0.9076	0.2247	0.1457
X_2	0.6399	−0.0292	0.5058
X_3	−0.8580	0.0409	−0.1845
X_5	−0.8737	0.2391	−0.1780
X_6	0.5104	0.7037	0.0869
X_7	−0.7999	0.1556	−0.2949
X_8	0.8259	−0.2904	0.2982
X_9	−0.7531	0.2857	0.3804
X_{10}	−0.8114	0.1645	0.3672
X_{11}	−0.5674	−0.2667	0.1498
X_{12}	0.4906	−0.1041	−0.5170
X_{13}	−0.7996	−0.4253	−0.0251
X_{14}	0.7366	0.5160	−0.1747

11.8 Boston Housing

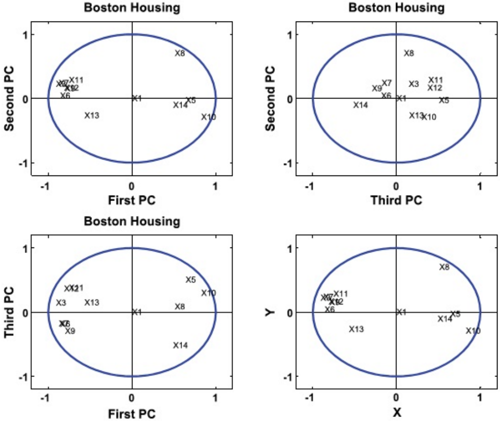


Fig. 11.10 NPCA for the Boston housing data, correlations of first three PCs with the original variables MVAnpcahousi

11.8 Boston Housing

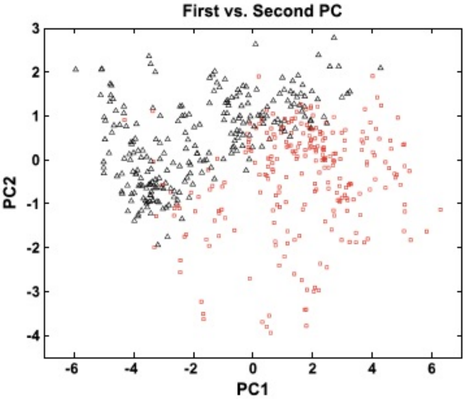
- The correlations with the first PC show a very clear pattern.
- The variables X_2, X_6, X_8, X_{12},and X_{14} are strongly positively correlated with the first PC, whereas the remaining variables are highly negatively correlated.
- The first PC axis could be interpreted as a quality of life and house indicator.
- The second axis, given the polarities of X_{11} and X_{13} and of X_6 and X_{14}, can be interpreted as a social factor explaining only 10% of the total variance.
- The third axis is dominated by a polarity between X_2 and X_{12}.

11.8 Boston Housing

- The set of individuals from the first two PCs can be graphically interpreted if the plots are color coded with respect to some particular variable of interest.
- Figure11.11 color codes $X_{14} >$ median as red points. Clearly, the first and second PCs are related to house value.
- The situation is less clear in Fig.11.12 where the color code corresponds to X_4, the Charles River indicator, i.e., houses near the river are colored red.

11.8 Boston Housing

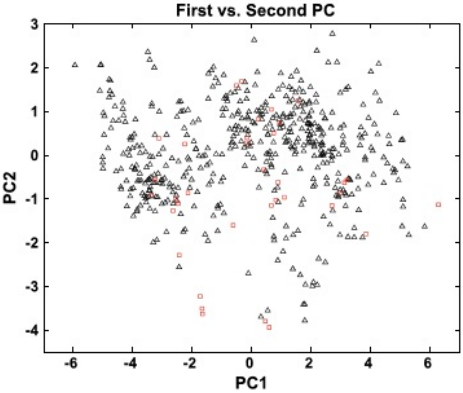
Fig. 11.11 NPC analysis for the Boston housing data, scatterplot of the first two PCs. More expensive houses are marked with red color MVAnpcahous



11.8 Boston Housing

Fig. 11.12 NPC analysis for the Boston housing data, scatterplot of the first two PCs. Houses close to the Charles River are indicated with red squares

MVAnpcahous


First vs. Second PC
PC2
PC1