7. Variable extraction and dimensionality reduction

The goal of the variable selection in the preceding chapter was to find least useful variables so that it would be possible to reduce the dimensionality or sometimes even to eliminate distracting variables. Now another approach is presented for *dimensionality reduction* which is on the basis of mathematical transforms for a data matrix. The goal is to obtain such a transformed matrix in another space that enables to find a smaller dimension q than p in the original one, i.e., $q<p$.

Generating new ”artificial” variables from the original data set is called variable or feature extraction. *Principal component analysis* (PCA) to be briefly considered is perhaps the best known of such techniques. It as well as *independent component analysis* (ICA) are unsupervised methods. They are linear transformations that optimally reduce dimensionality, in terms of the number of variables, of the original unsupervised data set.

While PCA and ICA are mainly used for numerical (time-independent) data, there are also methods such as *Fourier transform* and *wavelets* for one-dimensional time-series data and their two-dimensional versions for image data.

7.1. Variable extraction in general

Data preprocessing may include transformation (projection) of the original cases (also called objects, patterns or examples) into the transformed variable space, frequently along with *reduction of dimensionality* of a case by extraction of only the most informative variables.

The transformation and reduction of dimensionality may improve the process through consideration of only most important data representation retaining maximum information about the original data.

Reduction of the original dimensionality denotes a transformation of original p-dimensional cases into other q-dimensional variable cases, $q \leq p$. The transformation and dimensionality reduction can be considered as a transformation (mapping)

$$\mathbf{y} = \mathbf{F}(\mathbf{x})$$

of p-dimensional original cases $\mathbf{x}$ (vectors) into q-dimensional transformed vectors $\mathbf{y}$. $\mathbf{F}(\mathbf{x})$ is the q-dimensional transforming function.

The projection and reduction of the variable space may depend on the goal of processing which is, for instance, the classification of cases.

The reasons for data transformation and dimensionality reduction are as follows:

- Removing redundancy in data
- Compression of data sets
- Obtaining transformed and reduced cases containing only relevant variables that aid to design classifiers with better generalization capabilities

- Discovering the intrinsic variables of data that aid to design a data model and improving understanding of phenomena that generate cases
- Projecting high-dimensional data (preserving intrinsic data topology) onto low-dimensional space to visually show clusters and other relationships in data

7.2 Principal component analysis

Obviously the best known method of linear transformation and variable extraction is *Principal component analysis* (PCA) based on the statistical characteristics of a given data set represented by *covariance matrix* of data cases, its *eigenvalues* and the corresponding *eigenvectors*.

Principal component analysis represents a linear regression analysis as fitting planes to data in the way of least-squares errors.

It determines an optimal linear transformation

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

of a real-valued p-dimensional data vector **x** into another q-dimensional transformed vector **y**. The $p \times q$ linear transformation matrix **W** is optimal from the viewpoint of obtaining the maximal information retention. PCA is realized calculating correlations among elements of the original data vectors and finding (possibly reduced) representation retaining the maximum nonredundant and uncorrelated intrinsic information of the original data.

The original data set of matrix **X** is used to compute its covariance matrix, its eigenvalues and the corresponding eigenvectors arranged in descending order. The arrangement of subsequent rows of a transformation matrix **W** as the normalized eigenvectors, corresponding to the subsequent largest eigenvalues of the covariance matrix, will result in an optimal linear transformation matrix.

The elements of the *q*-dimensional transformed vector **y** will be uncorrelated and arranged in descending order according to decreasing information content. This allows for a straightforward reduction of dimensionality - and thus data compression – by discarding trailing variable elements with the lowest information content.

Depending on the nature of an original case, one can obtain a substantial reduction of vector dimensionality $q << p$ compared with the dimensionality of the original data.

Having determined the optimal transformation matrix $\hat{\mathbf{W}}$, one can reduce the decorrelated vector dimension and use reduced vectors for classification. In addition, all original p-dimensional cases can be optimally transformed to cases in the lower dimensionality space. The original data is then compressed with the minimal information loss when the data are reconstructed.

A PCA-based linear transformation of original data can also be interpreted as a projection of original cases into q-dimensional variable space with orthonormal bases guaranteeing that one obtains decorrelation of elements. We can see PCA as unsupervised learning from data. It does not employ class information, but only discovers correlation among cases and their elements, as well as ordered intrinsic directions where the cases change most (with maximum variance) as in Fig. 7.1. Despite this, it can also be utilized in classifier design for the projection and reduction of cases.

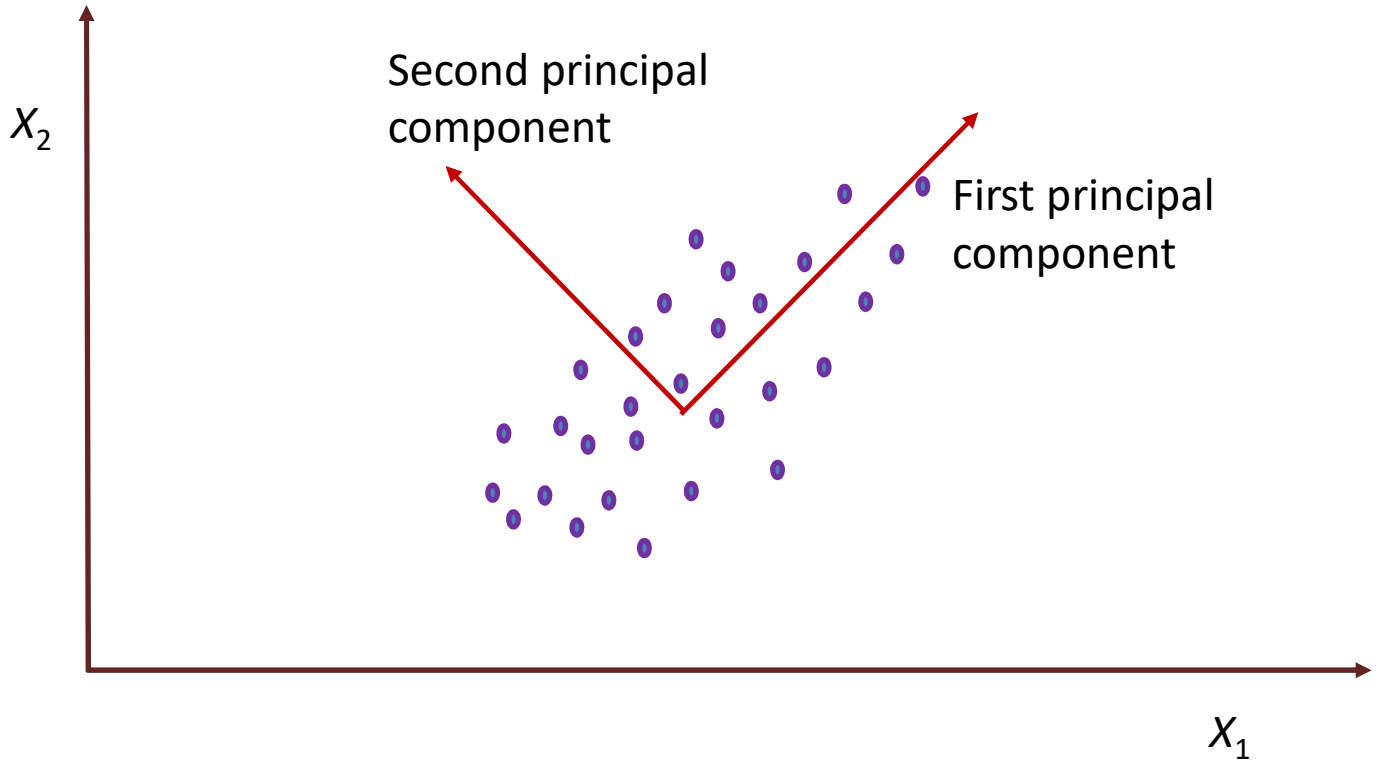


Fig. 7.1 Principal components.

We look at a 2-dimensional plane, but generally that of q < p is equally usable. The plane is spanned by the linear combination of variables that has maximum sample variance. Thus ”interesting” is defined in terms of the *maximum variability* in the data.

Let **X** be an $n \times p$ data matrix, where rows are cases or vectors $\mathbf{x}(i)$) and columns variables. To be exact, row i is the transpose $\mathbf{x}^T$ of the ith data vector $\mathbf{x}(i)$, because in mathematics these are given as column vectors. We assume that the mean estimated from the data set is first subtracted from the value of the variable in question. (Thereafter its mean is equal to 0 for every variable.)

Let **a** be the p×1 column vector of projection weights (still unknown) that result in the largest variance when data **X** are projected along **a**. The projection of any particular vector **x** is is the linear combination $\mathbf{a}^T\mathbf{x}$. We can express the projected values onto **a** of all data vectors in **X** as **Xa**, yielding an n×1 column vector of projected values. We can define the *variance* along **a** as

$$\sigma_a^2 = (\mathbf{Xa})^T(\mathbf{Xa}) = \mathbf{a}^T\mathbf{X}^T\mathbf{Xa} = \mathbf{a}^T\mathbf{Va},$$

where $\mathbf{V}=\mathbf{X}^T\mathbf{X}$ is the covariance matrix of p×p (**X** with mean equal to 0 mentioned above). Thus, variance (scalar to be maximized) can be as a function of **a** and covariance matrix **V**.

Maximizing variance directly is not well-defined, since we can increase it without limit simply by increasing the size of the components of **a**. So we impose a normalization constraint on **a** vectors such that $\mathbf{a}^T\mathbf{a}=1$.

With the normalization constraint it is possible to rewrite the optimization problem as that of maximizing quantity

$$u = \mathbf{a}^T \mathbf{V} \mathbf{a} - \lambda(\mathbf{a}^T \mathbf{a} - 1),$$

where λ is a Lagrange multiplier. Differentiating with respect to **a** gives

$$\frac{\partial u}{\partial \mathbf{a}} = 2\mathbf{V}\mathbf{a} - 2\lambda\,\mathbf{a} = \mathbf{0},$$

which reduces to the eigenvalue form with identity matrix **I** as follows.

$$(\mathbf{V} - \lambda\,\mathbf{I})\mathbf{a} = \mathbf{0}.$$

The first principal component of **a** is the eigenvector associated with the largest eigenvalue of covariance matrix **V**. Further, the second principal component (the direction orthogonal to the first component that has the greatest projected variance) is the eigenvector corresponding to the second greatest eigenvalue of **V** and more generally the eigenvector for the *k*th greatest eigenvalue to the *k*th principal component.

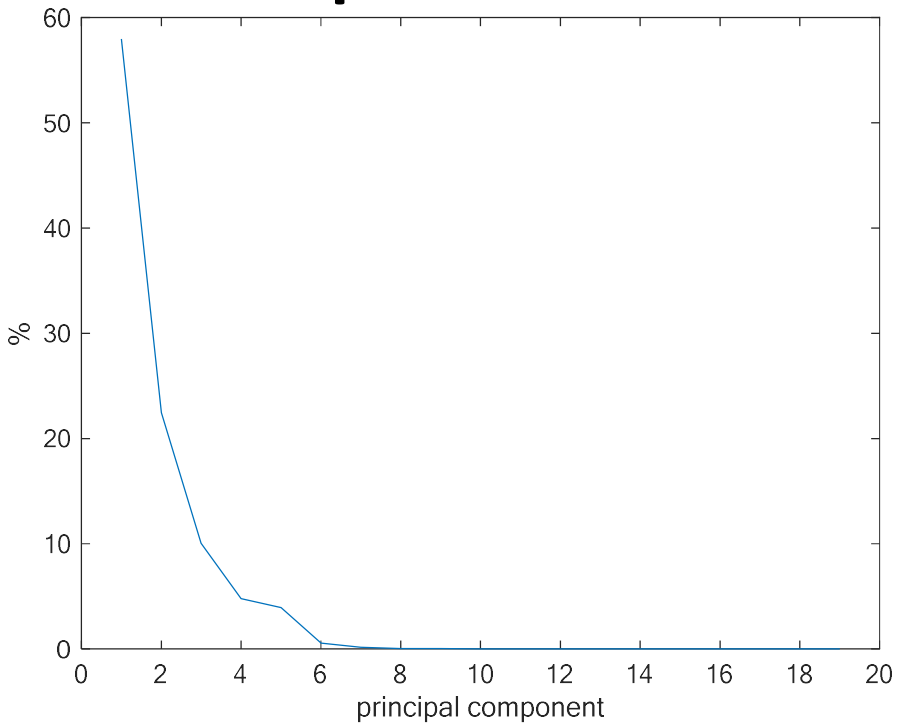
When the data are projected into the first k eigenvectors, the variance of the projected data can be expressed as $\Sigma^{k}_{j=1}\lambda_j$, where λ_j is the jth eigenvalue. Equivalently, the squared error in terms of approximating the true data matrix $\mathbf{X}$ using only the first k eigenvalues can be expressed as follows.

$$\frac{\sum_{j=k+1}^{p} \lambda_j}{\sum_{l=1}^{p} \lambda_l}.$$

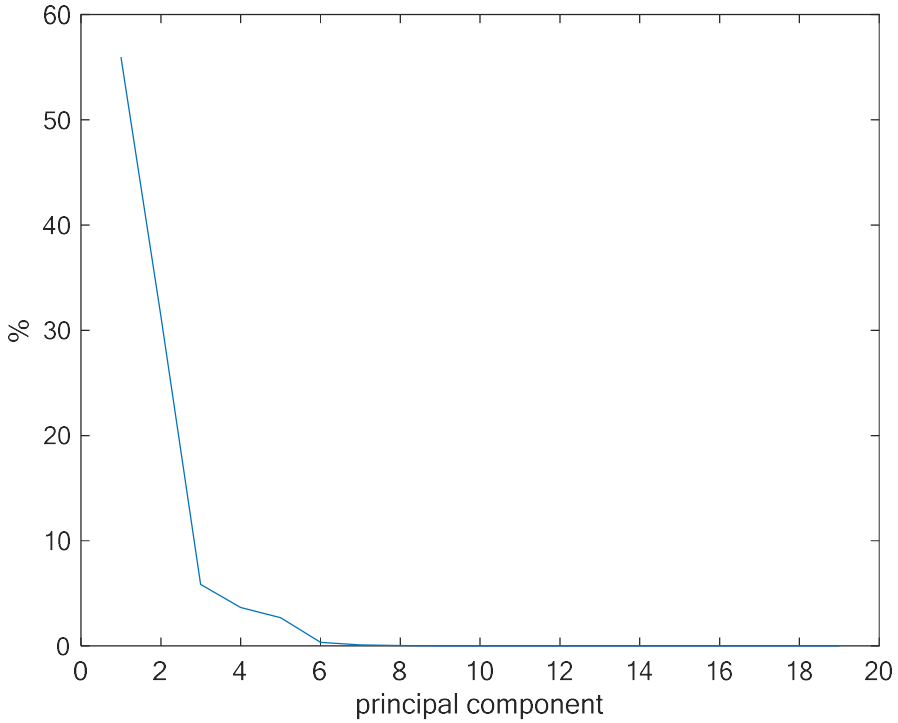
To select an appropriate number $k=q$ of principal components, k is increased until the squared error quantity above is less than some acceptable degree of squared error.

Three examples are shown in Figs. 7.2, 7.3 and 7.4.

Example 1



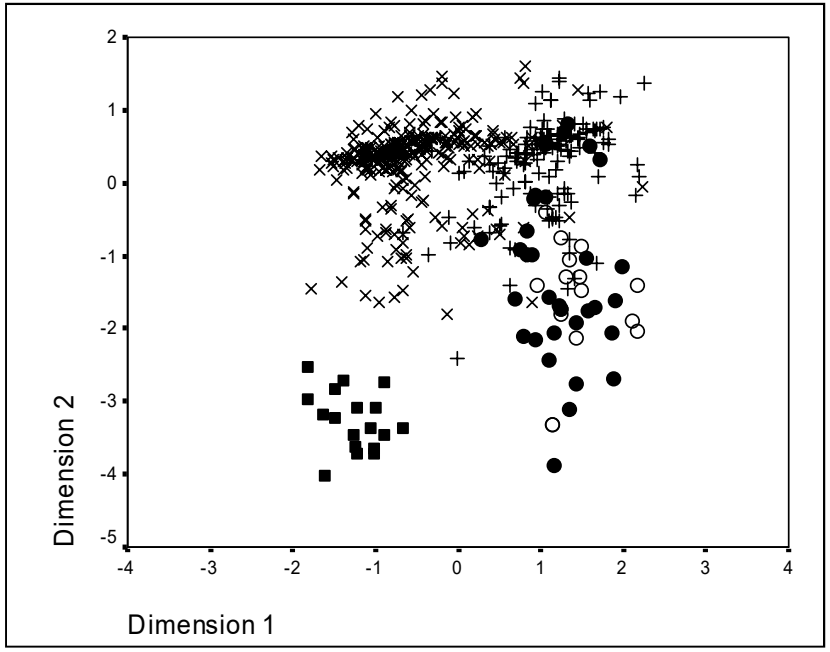
(a)



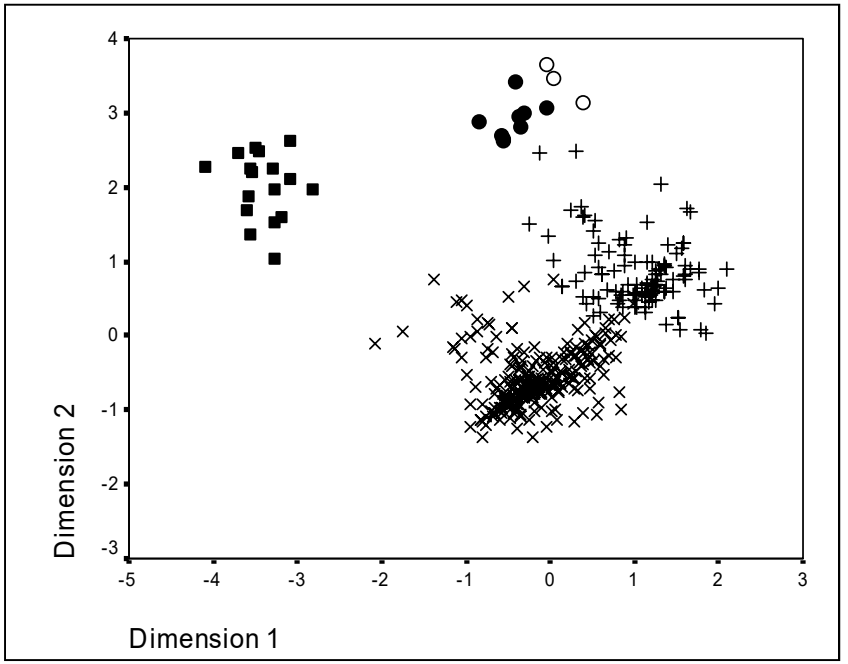
(b)

Fig. 7.2 Looking back at Fig. 4.11 (p. 161), there are so-called Scree plots computed on the basis of their data: (a) before cleaning outliers and (b) after cleaning. The first five components out of 19 explained (% of the vertical axis) virtually all variance in the data.

Example 2



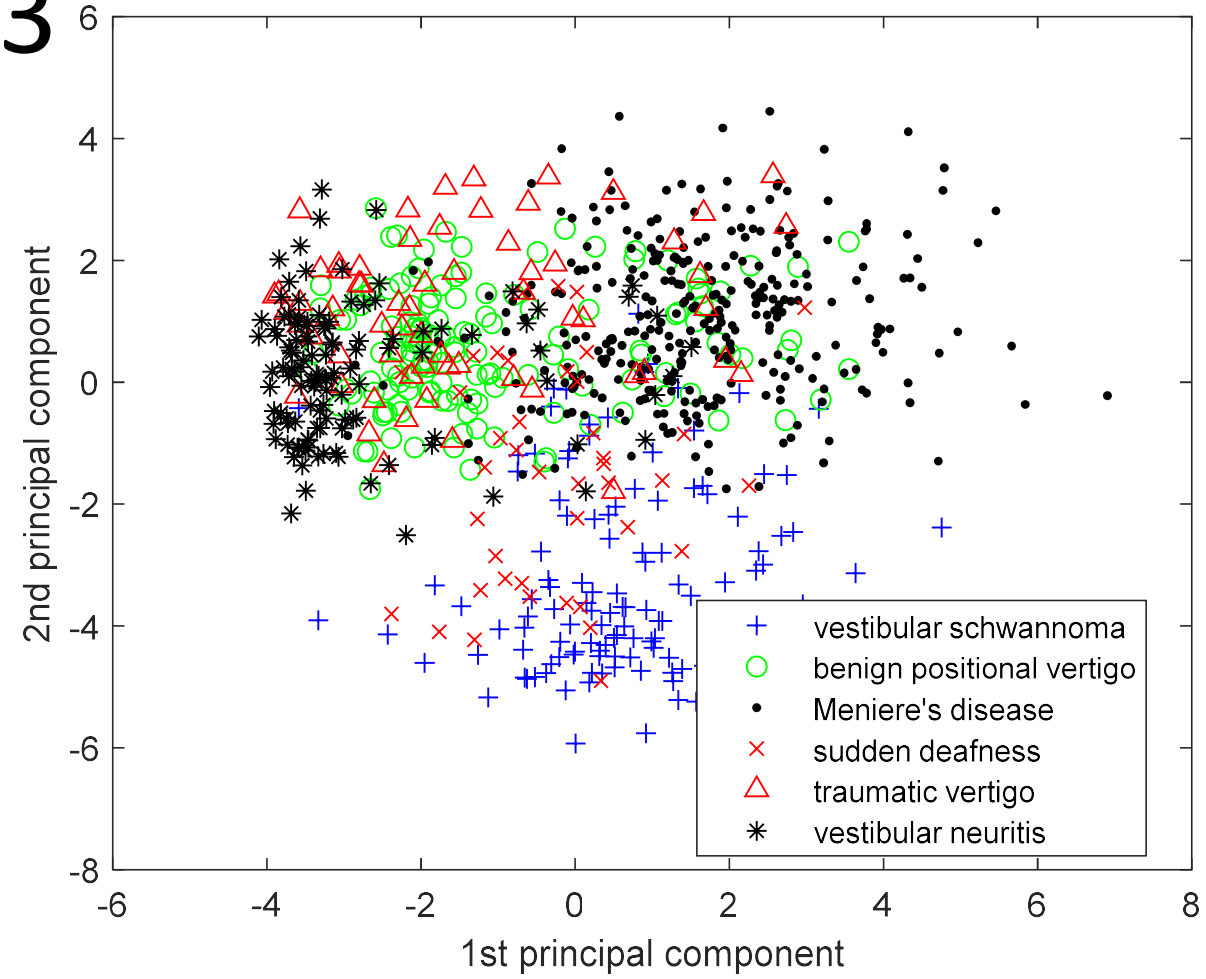
(a)



(b)

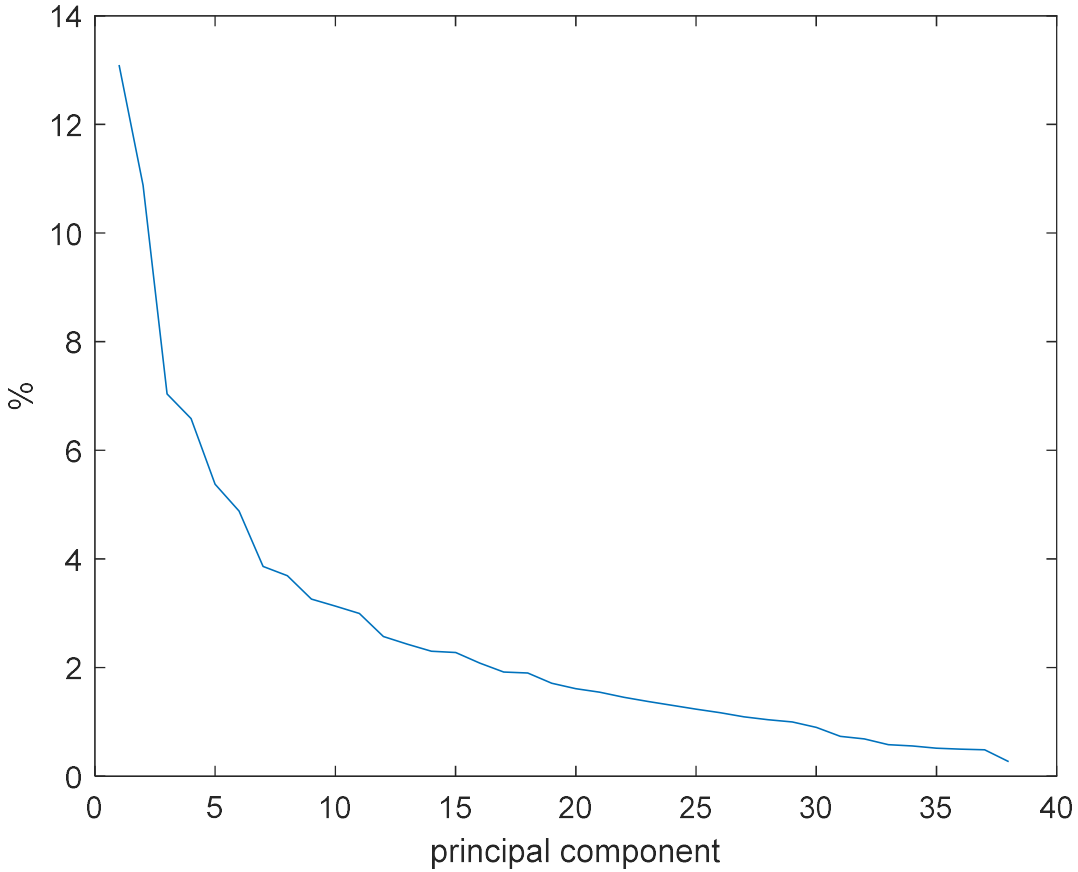
Fig. 7.3 Returning back to Table 3.1 (p. 84), the data of eight most important variables were used for principal components: (a) The first and second principal components (Dimensions 1 and 2) before cleaning and (b) after cleaning of noisy data. There are four disease classes and, outside them, the group of the normal (black squares).

Example 3



(a)

Fig. 7.4 (a) The two most important (largest) principal components of Vertigo data set from Table 2.1 (p. 63).



(b)

Fig. 7.4 (b) Eigenvalue spectrum (%) of principal components. The two largest components explained around 24%. Altogether, there were 38 components computed from original 38 variables.

Let the eigenvalues of the covariance matrix be arranged in decreasing order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Assume that the corresponding orthonormal eigenvectors (unit length) $\mathbf{e}^1$, $\mathbf{e}^2$,…, $\mathbf{e}^p$ compose $p \times p$ orthonormal matrix

$$\mathbf{E} = [\mathbf{e}^1, \mathbf{e}^2, \ldots, \mathbf{e}^p]$$

with columns being orthonormal eigenvectors. Then the optimal linear transformation

$$\hat{\mathbf{y}} = \hat{\mathbf{W}}\mathbf{x}$$

transforms the original p-dimensional $\mathbf{x}$ into q-dimensional case, maximizing the variance of the projected cases and provides $q \times p$ optimal transformation matrix $\hat{\mathbf{W}}$.

The resulting optimal linear transformation with the optimal transformation matrix $\hat{\mathbf{W}}$ is called the *Karhunen-Loéve* (KLT) or *Hotelling transformation*.

PCA can be effectively used for variable extraction and dimensionality reduction. Instead of the entire p-dimensional original data cases $\mathbf{x}$, one can form q-dimensional, $q \leq p$, vectors $\mathbf{y}$ containing only the first q most dominant principal components.