4.4. Principal component analysis

1. Estimated principal components:

We first estimate the theoretical variance-covariance matrix Σ of the random vector

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix}$$

by the sample variance-covariance S,

$$S = \begin{bmatrix} \hat{V}(Z_1) & \hat{C}(Z_1, Z_2) & \cdots & \hat{C}(Z_1, Z_p) \\ \hat{C}(Z_2, Z_1) & \hat{V}(Z_2) & \cdots & \hat{C}(Z_2, Z_p) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{C}(Z_p, Z_1) & \hat{C}(Z_p, Z_2) & \cdots & \hat{V}(Z_p) \end{bmatrix},$$

Where

$$\hat{V}(Z_j) = \frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}{n-1}, \hat{C}(Z_j, Z_k) = \frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{n-1}, j, k = 1, \cdots, p,$$

and where $\bar{X}_j = \frac{\sum_{i=1}^{n} X_{ij}}{n}$. Suppose $e_1, e_2, \cdots, e_p$ are orthonormal eigenvectors of S corresponding to the eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Then, the i'th estimated principal component is

$$\hat{Y}_i = e_i^t Z, \hat{V}(\hat{Y}_i) = \hat{\lambda}_i, i = 1, \cdots, p.$$

Example (Splus):

varir=var(ir) ### ir=as.matrix(iris[,2:5]) in R

eigenvarir=eigen(varir)

eigenvarir$vectors

$\hat{Y}_1 = 0.3613Z_1 - 0.0845Z_2 + 0.8566Z_3 + 0.3582Z_4$

eigenvarir$vectors[,1]

$\hat{Y}_4 = -0.3154Z_1 + 0.3197Z_2 + 0.4783Z_3 - 0.7536Z_4$

eigenvarir$vectors[,4]

eigenvarir$values[1] # $\hat{V}(\hat{Y}_1) = 4.228$

irprin=princomp(ir)

irprin

irprin$loadings # e_1, e_2, e_3, e_4

irprin$sdev^2 # $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4$

2. Excluding principal components:

The purpose of the principal component analysis is to reduce the complexity of multivariate data by transforming the data into the principal component space, and then choosing the first p^* principal components that explain "most" of the variation in the original variables. 3 criteria can be used to decide how many principal components to retain:

(a) Plot the eigenvalues $\hat{\lambda}_i$ versus i, The resulting plot is called a screeplot.

Example (Splus):

plot(irprin)

plot(irprin,style="lines") # the screeplot, not generated in R

It seems that the first one or the first two principal components are "mountainside" and the others are screes. Therefor, either $\hat{Y}_1$ or ($\hat{Y}_1$,$\hat{Y}_2$) should be retained.

(b) Include just enough components to explain some amount (typically 90%) of variance.

Example (Splus):

summary(irprin)

(c) Excluding those principal components with eigenvalues below the average.

Example (Splus):

mean(irprin$sdev^2) # $\frac{\hat{\lambda}_1+\hat{\lambda}_2+\hat{\lambda}_3+\hat{\lambda}_4}{4}$ = average

irprin$sdev^2 # $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4$

Note:

The first criteria usually results in too many included components while the third criteria typically includes too few. The 90% criteria is often a useful compromise.