Multivariate Statistical Analysis, Exercise 11, Fall 2011, Prof.S.K.Jeng

January 6, 2012 TA: H.C. Cheng

(Principal Components Analysis)

The main objective of principal components analysis (PCA) is to reduce the dimension of the observations. The simplest way of dimension reduction is to take just one element of the observed vector and to discard all others. This is not a very reasonable approach since strength may be lost in interpreting the data. An alternative method is to weight all variables equally, *i.e.*, to consider the simple average $p^{-1}\sum_{j=1}^{p} X_j$ of all the elements in the vector $X = (X_1, \ldots, X_p)^T$. This again is undesirable, since all of the elements of X are considered with equal importance (weight).

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

$$\delta^T X = \sum_{j=1}^{p} \delta_j X_j \text{ so that } \sum_{j=1}^{p} \delta_j^2 = 1.$$

The above equation is also called a standardized linear combination (SLC) and the weighting vector $\delta = (\delta_1, \ldots, \delta_p)^T$ can then be optimized to investigated and to detect specific features. One aim is to maximize the variance of the projection $\delta^T X$, *i.e.*, to choose δ according to

$$\max_{\delta:\|\delta\|=1} \mathrm{Var}(\delta^T X) = \max_{\delta:\|\delta\|=1} \delta^T \mathrm{Var}(X)\delta.$$

The interesting "directions" of $\delta$are found through the spectral decomposition of the covariance matrix (from the Rayleigh quotient theorem). Indeed, the direction δ is given by the eigenvector γ_1 corresponding to the largest eigenvalue λ_1 of the covariance matrix $\Sigma = \mathrm{Var}(X)$. Hence the SLC with the highest variance is the first principal component (PC) $y_1 = \gamma_1^T X$. Orthogonal to the direction γ_1 we find the second highest variance $y_2 = \gamma_2^T X$, the second PC. Proceeding in this way and writing in matrix notation, the result for a random variable X with $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \Sigma = \Gamma\Lambda\Gamma^T$ is the PC transformation which is defined as

$$Y = \Gamma^T(X - \mu).$$

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

In practice the PC transformation has to be replaced by the respective estimators: μ becomes $\overline{x}$, Σ is replaced by S (the empirical covariance matrix). If g_1 denotes the first eigenvector of S, the first principal component is given by $y_1 = (X - 1_n\overline{x}^T)g_1$.

More generally if $S = GLG^T$ is the spectral decomposition of S, then the PCs are obtained by

$$Y = (X - 1_n\overline{x}^T)G.$$

Note that with the centering matrix $H = I - (n^{-1}1_n1_n^T)$ and $H1_n\overline{x}^T = 0$ we can write

$$\begin{aligned} S_Y &= n^{-1}Y^THY = n^{-1}G^T(X - 1_n\overline{x}^T)^TH(X - 1_n\overline{x}^T)G \\ &= n^{-1}G^TX^THXG = G^TSG = L. \end{aligned}$$

where $L = \mathrm{diag}(l_1, \ldots, l_p)$ is the matrix of eigenvalues of S. Hence the variance of y_i equals the eigenvalue l_i.

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:

$$\Psi_q = \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} = \frac{\sum_{j=1}^q \mathrm{Var}(Y_j)}{\sum_{j=1}^p \mathrm{Var}(Y_j)}.$$

The covariance between the PC vector Y and the original vector X is calculated as

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathbb{E}(XY^T) - \mathbb{E}(X)\mathbb{E}(Y^T) = \mathbb{E}(XY^T) \\ &= \mathbb{E}(XX^T\Gamma) - \mu\mu^T\Gamma = \mathrm{Var}(X)\Gamma \\ &= \Sigma\Gamma = \Gamma\Lambda\Gamma^T\Gamma = \Gamma\Lambda. \end{aligned}$$

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

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

Using actual data, this can be translated into

$$r_{X_iY_j} = g_{ij}\Big(\frac{l_j}{s_{X_iX_i}}\Big)^{1/2}.$$

The correlations can be used to evaluate the relations between the PCs Y_j where $j = 1, \ldots, q$, and the original variables X_i, where $i = 1, \ldots, p$. Note that

$$\sum_{j=1}^p r^2_{X_iY_j} = \frac{\sum_{j=1}^p l_jg_{ij}^2}{s_{X_iX_i}} = \frac{s_{X_iX_i}}{s_{X_iX_i}} = 1.$$

So the $r^2_{X_iY_j}$ may be seen as the proportion of variance of X_i explained by Y_j.

The data set of this exercise comes from n random sample with variate 6. Please answer the following questions

(a) Please find the principal directions with respect to each eigenvalues

Solution: The vector of eigenvalues of S is

$$l = (2.985, 0.931, 0.242, 0.194, 0.085, 0.035)^T.$$

The eigenvectors g_j are given by the columns of the matrix

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

(b) Please plot the principal components 1 vs. 2, 2 vs. 3, 1 vs. 3 of the data and mark the first 100 by "o" and others "+" respectively.

Solution: See figure 1.

(c) Please find the proportion of variance and the cumulated proportion of each eigenvalue.

Solution:

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

(d) Please find the correlation coefficient $r_{X_i Y_j}$, $i = 1, \ldots, 6$, $j = 1, 2$ and plot it out. Which of the original variables are most strongly correlated with the principal component Y_1 and Y_2?

Solution: See figure 2. The variables X_4, X_5 and X_6 correspond to correlations near the periphery of the circle and are thus well explained by the first two principal components.

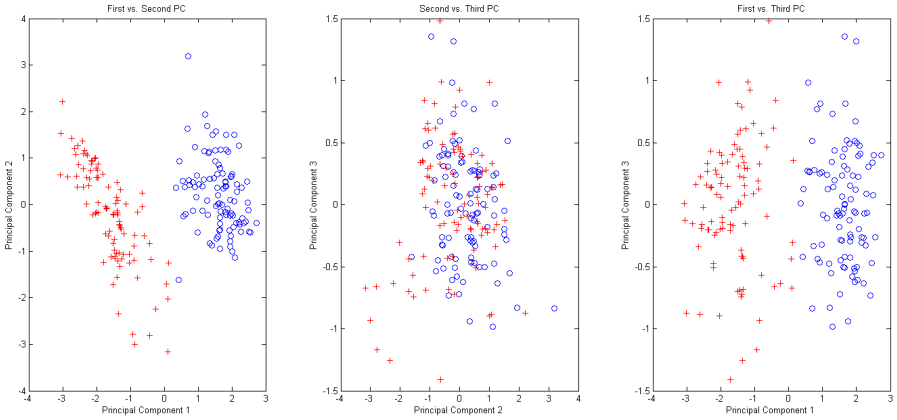

Figure 1:

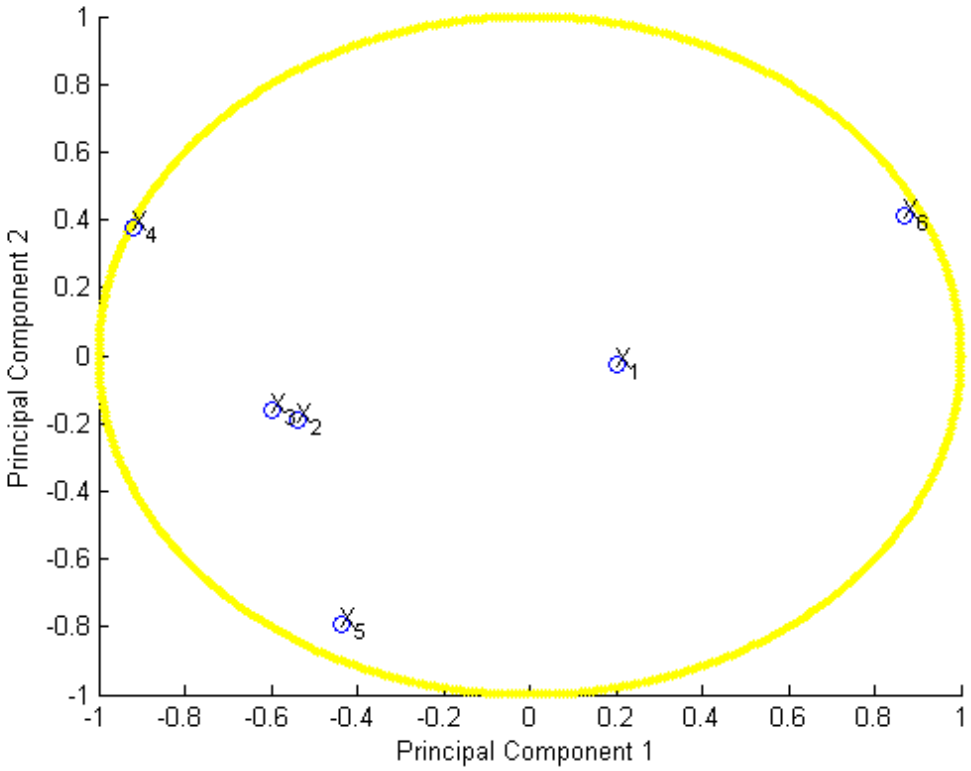

Figure 2: