Section on PCA and Biplots

PCA is a multivariate technique for data reduction.

Biplots are graphical techniques for expressing results of PCA and other analyses.

PCA (Principal Components Analysis)

1) Each point can be described as a vector. Recall that we can plot observations on axes representing variables and we can plot variables on axes representing observations.

   The objective of PCA is to create a new set of axes which describe greater variance than the original axes. From the original data (X) we will produce a set linear combinations resulting in new variables (Y_i) such that the values of the new variables will describe the maximum possible variance. The first new axis is the first principal component and describes the most variance. Subsequent components will have successively smaller variance and will have no covariance ($Cov(Y_i,Y_j = 0)$ between the component scores.

   Observations will be projected on this axis (perpendicularly) to get scores for each observation.

2) The locations of the point is given by the Euclidean distance from the origin O to a point P, where the length of the vector is $L(P) = \sqrt{x_1^2 + x_2^2 + x_3^2 + ... + x_k^2}$

   Frequently this distance measure will be

   $$\text{standardized } L(P) = \sqrt{\left(\frac{x_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2}{\sqrt{\sigma_{22}}}\right)^2 + \left(\frac{x_3}{\sqrt{\sigma_{33}}}\right)^2 + ... + \left(\frac{x_k}{\sqrt{\sigma_{kk}}}\right)^2}$$, but not today.

   These points can be projected on the principal component axes.

3) The principal components scores are sets of values defined by linear combinations of the eigenvectors and the original variables (which can be standardized, but not today). For the usual case with *n* observations and *k* variables where $n > k$,

$$Y_{1i} = e_{11}X_{1i} + e_{12}X_{2i} + e_{13}X_{3i} + \ldots + e_{1k}X_{ki}$$

$$Y_{2i} = e_{21}X_{1i} + e_{22}X_{2i} + e_{23}X_{3i} + \ldots + e_{2k}X_{ki}$$

$$Y_{3i} = e_{31}X_{1i} + e_{32}X_{2i} + e_{33}X_{3i} + \ldots + e_{3k}X_{ki}$$

$$\vdots \qquad \vdots$$

$$Y_{ki} = e_{k1}X_{1i} + e_{k2}X_{2i} + e_{k3}X_{3i} + \ldots + e_{kk}X_{ki}$$

   There will be a score for each observation in X. There will be a set of scores for each principal component.

   Note that the X variables are centered on the mean and scaled by the standard deviation.

   In many cases the analysis is run on a correlation matrix. In these cases the variances are all "1" and the eigenvalues sum to "n" instead of the total variance. .

5) $\hat{V}(Y_i) = \hat{V}(e_i' X_i) = e_i' \hat{V}(X_i) e_i = e_i' S e_i = \lambda_i$, giving the variance of the principal component, where $S = \lambda_1 e_1' e_1 + \lambda_2 e_2' e_2 + \lambda_3 e_3' e_3 + ... + \lambda_p e_p' e_p$. Recall that among the principal components the covariance is 0.

6) The proportion of variation described by the i[th] principal component is given by

$p_i = \dfrac{\lambda_i}{\lambda_1 + \lambda_2 + \lambda_3 + ... + \lambda_p}$ = $p_i = \dfrac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$ and the proportion of variance accounted for by the first "k" principal components ($k<p$) is $p_i = \dfrac{\lambda_1 + \lambda_2 + \lambda_3 + ... + \lambda_k}{\lambda_1 + \lambda_2 + \lambda_3 + ... + \lambda_p}$ = $p_i = \dfrac{\lambda_1 + \lambda_2 + \lambda_3 + ... + \lambda_k}{\sum_{i=1}^{p} \lambda_i}$

7) The correlation between the i[th] principal component score and the k[th] variable is $r_{x_k, \hat{y}_i} = \dfrac{\hat{e}_{ik}\sqrt{\lambda_i}}{\sqrt{S_{ik}}}$

Biplots – created using a macro by Michael Friendly (requires IML)

Simultaneously shows the relative positions of the rows (observations) and columns (variables) of a dataset.

The position of the observations will be determined by the Principal Component Scores (Y_i above)

The position of the variables is based on the value of the eigenvector value for the variable or the correlation between the principal component and the variable.

The observation axes may be scaled by a value referred to as alpha (α), such that the coordinates are multiplied by this value. The most common values are given below.

$\alpha = 0$, the biplot is called a **GH biplot** and results in a scale factor to equate the maximum distance from the origin of the variable and observation markers. In this type of biplot the column (variable) correlations are more closely approximated.

$\alpha = 1$, the biplot is called a **JK biplot** and is the usual PCA scale for the observations and eigenvectors for the variables. In this type of biplot the row (observations) values are more closely approximated.

$\alpha = 0.5$ is another scale value that is commonly used and is called the **SQ biplot** or **symmetric biplot** since the coordinates tend to be more similar between observations and variables.

Variable vectors may also be scaled separately