STAT 802 Multivariate Analysis Oct 23, 2003

What do we know about Principal Components Analysis (PCA)?

They are vectors which define indexes - linear combinations of the p variables, such that:

1. The linear combinations are uncorrelated
2. The order of the linear combinations produced is in order of decreasing variance
3. The variance of the i^{th} principal component is the i^{th} eigenvalue of the covariance matrix. The mean of each principal component is 0.
4. The linear combinations can be considered to be resultant vectors formed by the linear combination of the basis vectors corresponding to the original variables. The directions of these resultant vectors (i.e. principal components) are the directions of the eigenvectors. In fact the coefficients of the linear combinations are the components of the eigenvectors.
5. The p principal component scores are the p index values produced for each observation vector (each row of the data matrix). The representation in p-space of the data using the principal component scores has the same configuration as the original data space, but represented relative to a new basis.
6. Principal Components can be extracted from the covariance matrix of the correlation matrix. The latter is recommended except when the variables are commensurate.
7. The variance of the p-principal components sums to the trace of the covariance matrix, for principal components extracted from the covariance matrix. When the correlation matrix is used, this sum is p.
8. A subset of the first k principal components is often used to represent the data more simply. Often k=2 or 3 is used for graphical presentation of the principal components. The sum of the first k eigenvalues can be used to compute the proportion of the summed p data variances accounted for by these k principal components.
9. A scree plot shows how the eigenvalues decrease across their order. When the amount of decrease becomes linear in the order number, the components are being extracted from a spherical configuration of points, and these components are considered to be expendable in reproducing the information in the data. In other words, the scree plot can suggest a value for k.
10. Principal components are especially useful when $k \leq 2$, for then the data can be plotted in a one-or two-dimensional scatter. Outliers, clusters, and any categorical features not in the quantitative data set, can be examined visually in this case.
11. Variables amenable to PCA are usually quantitative – qualitative variables can be made quantitative with dummy variables, but this is not often helpful unless the categorical variables are naturally dichotomous.
12. Principal component scores will usually be approximately normally distributed, due to the CLT phenomenon. A normal Q-Q plot can therefore be used to study anomalies, and this is much better than looking at the original variables for two reasons – the original data variables may have unknown distribution shapes, and also the confusion of univariate vs multivariate outliers is reduced.

13. If p variables are to be used as independent variables in a regression, and if there is a possibility of numerical problems due to collinearity, replacing the original variables with their principal components may be advisable. The betas expressed on the original variables will still be unstable, but the predictions using the regression fit will not be affected by matrix inversion problems.

14. Mathematically, PCA is just eigenanalysis!

Examples:

Men Data
Track Records – Women
Track Records – Men

Exercise: This should be handed in on Thursday, Oct 30.

Table 1.9 and Table 8.6 in the text give record times for men and woman for the 1984 Olympics in Los Angeles. (If you want to use updated data and have time to spare, feel free to use it instead!). I will e-mail these data sets to you in case you wish to use my format, but the data is also on the CD that comes with the text.

Explore these two data sets using PCA. Make use of the country name variable as well – it helps to interpret the output of the PCA. Note any clusters or anomalies that show up graphically. Write up your report in no more than 2 pages, including any graphs you wish to include – you can describe briefly what you did and what you found, but you need not print out everything! The 2 page limitation is important for your mark on this assignment.