STAT 8250 — Applied Multivariate Methods
Lab – Due Friday, Nov. 10

Example:

Jeffers (1967) described a study carried out on the strength of wooden props used in mining. A number of variables were measured on each pitprop and the prop was compressed in a vertical position until failure occurred. The variables measured were the following: Topdiam(x_1) = the top diameter of the prop, in inches; Length(x_2) = the length of the prop, in inches; Moist(x_3) = the moisture content of the prop, expressed as a percentage of dry weight; Testsg(x_4) = the specific gravity of the timber at the time of the test; Ovensg(x_5) = the oven-dry specific gravity of the timber; Ringtop(x_6) = the number of annual rings at the top of the prop; Ringbut(x_7) = the number of annual rings at the base of the prop; Bowmax(x_8) = the maximum bow, in inches; Bowdist(x_9) = the distance of the point of maximum bow from the top of the prop, in inches; Whorls(x_{10}) = the number of knot whorls; Clear(x_{11}) = the length of the clear prop from the top of the prop, in inches; Knots(x_{12}) = the average number of knots per whorl; Diaknot(x_{13}) = the average diameter of the knots, in inches.

Although one of the main aims of the study was to relate the maximum compressive strength to the variables, an initial attempt was made to reduce the dimension of the problem using principal components. The table below gives the upper triangle of the correlation matrix for the 13 variables measured on $n = 180$ props cut from Corsican pine.

Copy the file pitprop1.sas from the course's home directory to your directory, run the program and look at the program and output. I did not have the original data for this example, only the correlation matrix. Therefore, I entered the data in data set pitprop, which is of the special type corr, a correlation matrix data set. To avoid having to enter each off-diagonal element of this 13×13 matrix twice, I just entered the upper-triangle of the matrix and I entered missing values (periods) for the lower triangle. To fill in these lower triangle elements I went into PROC IML and replaced all of the missing values. The final correlation matrix, ready for analysis, is printed on p.1 of the output.

Using PROC PRINCOMP, I obtained the eigenvalues and vectors for this correlation matrix. The eigenvalues appear on p.2 of the output. Notice that SAS prints the erroneous information: 10000 observations. The first step is to decide on the number of principal components, k. Bearing in mind the sampling variability of the eigenvalues, it is not easy in this example to choose a cutting point. The rather arbitrary rule of considering components with eigenvalues greater than unity would lead to $k = 4$, thus accounting for 74% of the sum of the eigenvalues. This rule often leads to too few components and does not seem appropriate here. A figure approaching 90% would often be regarded as adequate and Jeffers settled for $k = 6$, giving a total contribution of about 87%.

The corresponding eigenvectors for the first six PCs are given at the bottom of p. 2 of the output. Interpretations of these PCs can be derived by examining the size and sign of the elements of the eigenvectors. These elements give the relative weights given to the variables in each component. Important variables are those with large negative or positive weights.

The eigenvalues on p.2 have been scaled to have length 1. A technique which can be useful in interpreting PCs is to re-scale the eigenvectors, so that the largest element in each vector is 1. That is, we take each eigenvector and multiply it by 1 over its largest element. Jeffers suggests this approach in this example and interprets the PCs based on "large" elements of the associated eigenvectors where large is defined (somewhat arbitrarily) to be greater than .7 (in absolute value). The first 6 re-scaled eigenvectors appear below.

Based on these eigenvectors, we can derive some interpretations for the first 6 PCs. For example, the first PC gives high positive weights to top diameter, the length, the number of rings at the top and base, the bow, and the number of whorls. This PC may be interpretted as a general index of the size of the prop. The second PC, giving high weight only to moisture content and the specific gravity of green timber, is a measure of the degree of seasoning. The third component is a measure of the rate of growth and strength of the timber, while the fourth component is a contrast between the length of the clear prop from the top and the number of rings at the base. The fifth PC is a direct measure of the number of knots per whorl, and the sixth component is a combined index of the average diameter of the knots and the strength of the timber, that is, a combined strength index. These interpretations can also be obtained by looking at the eigenvectors on the scale on which they're given on p.2, but perhaps Jeffers' method makes it a bit easier.

Exercise:

Data on the national women's track records for seven events are contained in the file wtrack.dat in the public data directory. A partial listing of the data appears below.

Write a SAS program to read the data in, and perform 2 principal components analyses: 1 using the correlation matrix and 1 using the covariance matrix. Answer the following questions.

1. What percentage of the total variance is explaind by the first principal component for the original variable analysis? What percentage is explained by the first PC for the standardized variable analysis?

2. Look at the first eigenvector in each analysis. Interpret the first PC based on the covariance matrix. Interpret the first PC based on the correlation matrix.

3. Which analysis is appropriate? Why?

4. Interpret the second PC from the correlation matrix analysis.