**Week 8 Discussion Board:**

What are two benefits of using Principal Components Analysis (PCA)?  
Are there any limitations to using PCA?

One of the benefits of using Principle Components Analysis (PCA) is that the principle components will look for a few linear combinations of the original variables that can be used to summarize a data set while losing as little information as possible in the process. The second benefit is that the method consists of transforming a set of correlated variables to a new set of variables which are uncorrelated, which assist in dealing with a multicollinearity.

One of the limitations of PCA is that if the original variables are almost uncorrelated, then there is little purpose in performing PCA. The method will merely find components which are close to the original variables but will arrange them in decreasing order of variance.

Per Michael's point, PCA is a very useful technique for variable reduction.  For example, I recently worked with a high end retailer that wanted to better understand the relationship between customer satisfaction and sales.  They provided data for each of their 60+ United States based stores which included quarterly sales figures and quarterly results of two different customer satisfaction surveys.  Between the two different surveys their were over one hundred different individual questions asked of customers.  During data prep, we were able to run PCA and determined that the over 100 potential input variables could be reduced into fewer than 10 principal components.  This occurred because many individual questions were actually measuring the same construct.  For example, one of our components was composed of questions that were all closely related to the customer's in-store experience and wait time while another included questions related to the follow up actions taken by the sales associate (did the client receive a personalized thank you note, etc).  Reducing the 100 different potential input variables into a small subset of components which explained almost all of the variance in the responses provided a better performing (more accurate) model which ran more efficiently.  As described above, most of the components clearly measured an easily identifiable construct (e.g. "In Store Experience") and were thus easy to label and explain  however, some of the components were more difficult to interpret and label and this led to some complications in communicating model results to management. However the benefits outweighed these challenges.

Given a set of eigenvalues: 7.2 6.1 4.3 2.2 1.4 1.3 0.8 0.1, how many principal components should we keep and what is the cumulative percent of the total variation captured by the eight principal components?

The Paper discusses various approaches to the problem including:  1) The Eigenvalue 1 criterion in which you retain any component with an eigenvalue greater than 1  (Use PROC Factor with MINEIGEN=1); 2) Generate a "Scree Plot" and "look for a break between the components with relatively large eigenvalues and those with small eigenvalues"; 3) Retain a component if it explains a specified proportion of variance in the data and 4)using the "Intepretability Criterion".    
  
Each of these four approaches is further described in the paper and the authors recommend using a combination of the four approaches.  I found the use of the "Interpretability Criterion" especially interesting because, as seen in my post to the this week's other thread, I described a problem I had encountered interpreting some components that had resulted from a recent application of PCA which led to some challenge in communicating results.