Chapter 16 Notes

Principal Components Analysis

The General Idea and Theory

It's common to have data sources with highly related variables, for instance:

- Databases with many variables on similar product characteristics
- Survey results where questions are highly related
- Patient health data

Highly correlated variables can be a problem for modeling (e.g. issues related to multicollinearity) and large numbers of variables can be problematic for interpretation (having more predictor variables can make interpretation and visualization much more complicated).

One technique that can alleviate these issues is principal components analysis. In principal components analysis, we start with data containing variables x_1,…, x_p which have some significant correlations between them. Often *p* is large. We can get rid of the correlation by transforming the x_i variables to new variables z_i called principal components where

$$z_i = a_{i1}\, x_1 + a_{i2}\, x_2 + \ldots + a_{ip}\, x_p$$

There are just as many z variables as we had x variables. The vector $a_i = (a_{i1}, a_{i2}, \ldots, a_{ip})'$ is the i^{th} eigenvector of the correlation (or covariance) matrix for the data and the eigenvector a_i has an associated eigenvalue λ_i. The eigenvalues are ordered from largest to smallest. The correlation matrix is usually preferable to the covariance matrix because scale of the original variables is removed (large x variables with large variations are not given more importance than small variables with small variations).

The i^{th} eigenvalue tells us about the portion of the overall variation in the original data captured by the i^{th} principal component, so z_1 tells us the most about the overall variation in the original data and z_p tells us the least, but together the z variables contain all of the variation that was in the original data. Specifically, the proportion of variation from the original variables described by z_i is

$$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

If we take the first k z variables, that will give us

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

of the overall variation in the original data. So we can approximate the original data by a subset of the new variables and the new variables are uncorrelated. Our goals in such analyses are:

- to remove the correlations
- reduce the number of dimensions we have to work with (hopefully a small number of z variables will tell us about a large percentage of the variation in the original data)
- potentially get a rough understand of underlying features in the data

The original variables with large positive or negative coefficients in a principal component z_i are highly correlated with what that principal component represents and sometimes we can get a rough interpretation of what that principal component tells us (maybe it's like a product quality feature or a contrast between quality and price).

Note that we lose some interpretability because we are no longer looking at our original variables (variables that we know the meaning of). Instead, we have aggregate measures which may represent underlying features in the data that we can't concretely define (and we may not be able to even describe what those features are). The first few principal components will be the most likely to be interpretable because they pick up the more prominent features in the data.

Part of the goal is to work with fewer dimensions, so we would want to choose some number of the first principal components.

There are several rules of thumb about how many principal components to keep, including:

- describing at least some certain percentage (say 70% or 90%) of the variation in the original variables
- keeping components whose eigenvalues are greater than the average of the eigenvalues
- looking at a scree plot of the eigenvalues and looking for an elbow in the curve (where the curve starts to become very horizontal)

The Princomp Procedure

The procedure of interest in SAS is **proc princomp**. It will generate the eigen-information and various useful graphics for us. We can also output the principal components to an output data set to visualize and use the principal components after the analysis.

Examples and Exercises

U.S. Crimes Data

As a concrete example, let's start with the Crimes data set from the **Getting Started** example in **The Princomp Procedure** documentation and see what that tells us about crime rates in each state in 1977.

- Look at scatter plots and correlations to see that there is significant correlation in the data.

- We can look at the principal components analysis to see how much variation is described by the first few principal components and what features those principal components might represent about crime.
- For the first few principal components (which we might be able to interpret somewhat) we can also look at where the states fall on those principal axes. This might tell us something about the general types of crimes that tend to happen in the various states.
- If we are willing to assume the components are normally distributed, we could also include confidence ellipses to see which states might be outliers (atypical for the features we are picking out of the data, assuming the data in the principal components is normally distributed).

Decathlon Data

In the decathlon data we have measures for several athletes in the 10 events of a decathlon along with an overall score for the athletes. We could perform a pca to see if we could pick features of athleticism out of the individual event measurements, and we could see how those features match up with the score (which we will call **dscore** to avoid confusion with the concept of scores for principal components) in the data set.

Some initial processing:

- First look for and remove extreme overall scores. These athletes may be outliers and might have too much of an impact on the correlation matrix and the principal component transformation. We can check for outliers using univariate statistics and box plots.
- After removing outliers, we may also want to change signs for events where smaller measurements are better. Then increases in all variables will be indicators of better scores.

Finally, we can perform a principal components analysis.

- Perform a principal components analysis on the processed event variables (everything but **dscore**).
- What features can we identify and how do (or don't) those correspond to the **dscore** variable in the data set?
- Look at plots and at the correlation of the principal components with the **dscore** variable.
- What do the plots and correlations suggest about relationships between the principal components and the decathlon score?

Pain Data

The pain data set provides a correlation matrix for 9 questions patients who were asked about source of pain. The questions are given on pages 295 and 299 of the text.

We can perform a principal components analysis on this correlation matrix using a **type** option to the **proc** statement.

- Perform a principal components analysis on this data and see what features you can identify in the data.
- Can you infer what any of the first few features might represent?

Principal Component Regression

One use for principal components is in regression. This example is intended to introduce the idea. There is a procedure in SAS for doing principal component regression, and additional diagnostics should be checked to avoid overfitting as can happen following the direct route used in the following.

If we fit a linear regression model to principal components, this is called principal component regression. If a few of the principal components represent features which are closely related to a chosen response, those components may be good explanatory variables for the response and we can potentially model the response as a linear function of some of the principal components. We are then predicting the response as a function of the features those principal components are picking out of the original explanatory variables.

As an example, let's take the output data set **pcout** from the **decathlon** principal components analysis. We could model **score** using linear regression to see how well the first two principal components (the "overall athleticism" and "power vs. stamina" contrast components) could be used to predict the assigned **score**.

- Use **proc reg** to fit a linear regression of **dscore** as a function of **prin1** and **prin2** from that output data.
- Take a look at the diagnostics and results. How well do these first 2 components predict decathlon score?
- Are there any issues that show up in the diagnostics?
- If we choose all 10 principal components as our possible explanatory variables and use a forward selection method with entry significance level .05, which principal components would we keep as predictors?
- What are the benefits and drawbacks of this model compared to the 2 component model? (Consider interpretability and amount of variation described by each model).