Lecture 07
Dimensionality Reduction with PCA

10 February 2016

Taylor B. Arnold
Yale Statistics
STAT 365/665

Yale

As we have started to see, the curse of dimensionality stops us from being able to fit arbitrarily complex models in high dimensional spaces.

Additive models try to avoid this by fixing the structure of the learned models to limit interactions between the input variables.

Tree-based models attempt to use the data itself to greedily learn which interactions are actually important.

Today we are going to look at another technique called principal components (PCs), or principal component analysis (PCA), a specific example of dimensionality reduction.

Like trees, these use the data to find lower dimensional structures hidden in higher dimensional space. They differ from trees, however, because principal components use only the predictor variables (not the responses) and attempt to capture global and linear structure, rather than local ones.

Motivating example

Say that we have a dataset of the following measurements from a large set of human volunteers with the following variables:

- height
- weight
- waist size
- shoe size
- length of right arm
- length of left arm
- length of torso
- pant inseam length
- hat size
- left hand ring size
- right hand ring size

Technically we have 10 variables, though most of the variation in the dataset can probably be summarized by at most 2-3 summary variables.

Motivating example, cont.

In decreasing order of variation, consider the following measurements that can be derived from these 10 variables

1. height
2. body mass index
3. ratio of torso length to total height

Overall height captures a large amount of the variation in the total dataset. Accounting separately for BMI, which in theory should be relatively uncorrelated with overall height, captures much of the next largest variation in the data. The final measurement attempts to capture the remaining variation based on how height is distributed over a given individuals frame.

Motivating example, cont.

Conceptually, these values mimic what principal components attempt to do: describe the maximum amount of variation in the data with a smaller number of variables.

Each principal component, however, must be a linear function of the input variables (so BMI would not be allowed). We also want them to be defined mathematically rather than requiring us to hand construct them for each dataset.

Principal components

Formally, the principal components of the matrix X are a linear reparameterization $T = XW$ of the matrix X. The first column of T is the first principal component, the second column is the second principal component, and so on.

Specifically, the matrix W is defined uniquely by the following conditions:

1. Each column of T must be uncorrelated with the others; specifically, W is an orthogonal matrix called the *loadings*
2. The first column of T has the largest variance of all linear combinations of the columns of X, the second column has the highest variance conditioned on being uncorrelated with the first, and so forth.

Principal components, cont.

It can be shown that the matrix W is equal to the eigenvectors of the Gram matrix X^tX. From this relationship, there are many results from numerical linear algebra that can be used to develop theoretical results about principal components.

For the purposes of this course, however, we will be more concerned with how they can actually be of use in data analysis, visualizations, and predictive modeling. We will do that for the remainder of today's class by applying them to several example datasets.

A look ahead

The main shortcoming of principal components are that they only capture global linear structures in the data. This tends to be a larger problem for prediction than it is for visualization.

Figuring out how to get non-linear extensions of principal components is a wide open problem in statistic and machine learning. Some avenues of research include:

- locally linear embedding
- factor models
- diffusion maps
- mixture models

We will touch on some, though certainly not all, of these in the upcoming weeks.