Principal components analysis I

Patrick Breheny

November 17

Introduction

- At the end of the previous lecture, I suggested that the singular values decomposition could be used to reduce the dimensionality of the data set
- This is the key idea behind principal components analysis – to reduce the dimension of $\mathbf{X}$ while accounting for as much of the information in $\mathbf{X}$ as possible
- This aim is achieved by transforming to a new set of variables (the principal components) that are linear combinations of the original variables
- The new set of variables have lower dimension and are uncorrelated, both of which greatly simplify description, summarization, analysis, and model fitting

Principal components in terms of SVD components

- Suppose that we have centered and scaled $\mathbf{X}$ so that all of its columns have mean zero and variance 1, and let $\mathbf{X} = \mathbf{UDV}^T$ be the SVD of this centered and scaled matrix
- By convention, the singular values $\{d_j\}$ and their associated vectors $\{\mathbf{u}_j\}$ and $\{\mathbf{v}_j\}$ are ordered, so that $d_1 \geq d_2 \geq \cdots \geq d_p$
- Now, the variables $d_j\mathbf{u}_j$ are called the *principal components* of the original data $\mathbf{X}$, for reasons that we will now describe

Properties of principal components

- First, note that the principal components are linear combinations of the original variables:

$$\mathbf{X}\mathbf{v}_j = d_j \mathbf{u}_j$$

- Furthermore, $\mathrm{Var}(d_1\mathbf{u}_1) \geq \mathrm{Var}(d_2\mathbf{u}_2) \geq \cdots \geq \mathrm{Var}(d_p\mathbf{u}_p)$
- Indeed, out of all possible vectors $\mathbf{z}$ that can be formed from a normalized linear combination of the original explanatory variables (*i.e.*, such that $\mathbf{z} = \mathbf{X}\mathbf{a}$ where $\mathbf{a}^T\mathbf{a} = 1$), the variable with the largest variance is $d_1\mathbf{u}_1$
- Out of all possible normalized linear combinations $\mathbf{z}$, the one that has the largest variance and is orthogonal to the first combination (*i.e.*, such that $\mathbf{z}^T\mathbf{u}_1 = 0$) is $d_2\mathbf{u}_2$, and so on

Properties of principal components (cont'd)

- Recall that $\{d_j^2\}$ are the eigenvalues of $\mathbf{X}^T\mathbf{X}$, and that $\sum_j \lambda_j = \mathrm{tr}(\mathbf{X}^T\mathbf{X})$
- Thus, the jth principal component accounts for a proportion p_j of the total variation in the original data:

$$p_j = \frac{d_j^2}{\mathrm{tr}(\mathbf{X}^T\mathbf{X})}$$

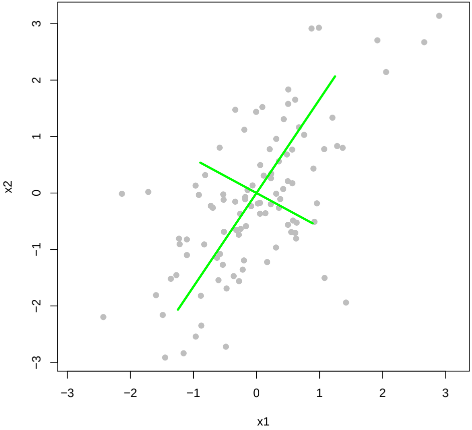
- Note that, as $\mathbf{X}$ has been centered, the elements on the diagonal of $\mathbf{X}^T\mathbf{X}$ are proportional to the variance of each of the original explanatory variables

More terminology

To summarize,

- The vectors $\mathbf{v}_j$ (the columns of $\mathbf{V}$) are the principal component directions, or *loadings*, and they describe the transformation process by which the new variables are created out of the old
- The vectors $\mathbf{u}_j$ (the columns of $\mathbf{U}$) are the normalized principal components (sometimes called the *principal component scores*)
- The singular values d_j are used to rank the principal components in term of importance

An illustration



Uses for principal component analysis

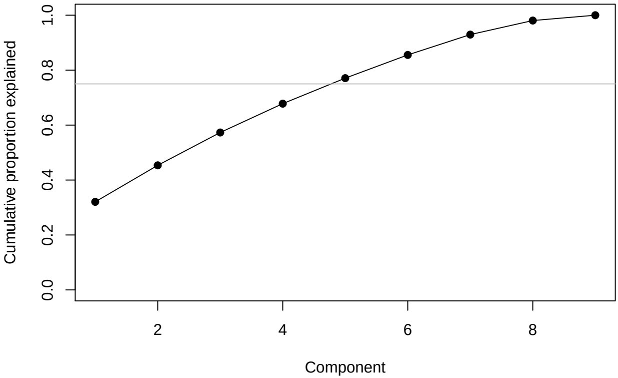
- The general hope of principal components analysis is that the first few principal components will contain almost all the relevant information for the problem, and can therefore provide a convenient lower-dimensional summary of the data
- The two most common uses for principal components are:
  - As a means of constructing informative graphical summaries of multivariate data
  - As inputs for regression
- Occasionally, principal components are also of interest directly, as a way of extracting underlying, latent factors from data

How many components?

- Given that we would like to reduce the dimension of the problem by selecting some smaller number of principal components, an obvious next question is: how many components?
- A number of approaches have been proposed, both formal and informal
- We will present the two most common informal approaches here, and illustrate their use on our heart disease data set
- Recall that this data set had a number of highly correlated variables (cholesterol, body fat %, BMI, etc.)

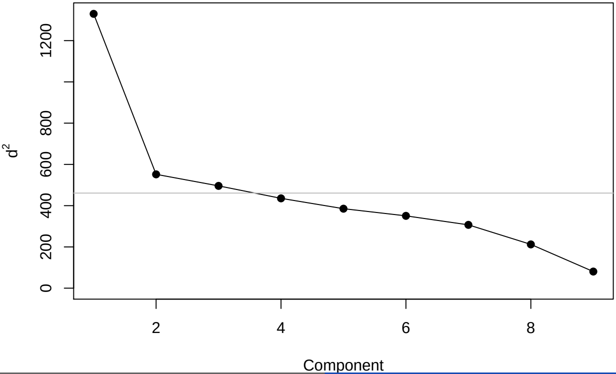
Cumulative proportion explained

One possibility is to base the decision on the cumulative proportion explained, and when it crosses some threshold; say, 75%



Scree plot

Another approach is to plot is to plot $\{d_j^2\}$ and base the decision on when the proportion of variance explained falls below $1/p$ (this plot is known as a *scree plot*:



Scree plot (cont'd)

- Alternatively, one can take the more subjective approach of looking at the scree plot and subjectively identifying an "elbow" where the slope changes from steep to shallow
- This rule would suggest retaining only the first principal component; the other two rules suggest retaining three or four

Heart disease data: Loadings

Let's continue with our heart disease example, and inspect the first four principal components; below are the loadings:

	PC1	PC2	PC3	PC4
sbp	-0.32	0.24	-0.13	-0.20
tobacco	-0.30	0.46	0.07	0.01
ldl	-0.33	-0.36	0.00	0.14
adiposity	-0.52	-0.19	-0.08	-0.14
typea	0.02	-0.28	0.79	-0.21
obesity	-0.40	-0.39	0.04	-0.31
alcohol	-0.12	0.54	0.46	-0.26
age	-0.46	0.19	-0.14	0.16
famhist	-0.20	0.00	0.34	0.83

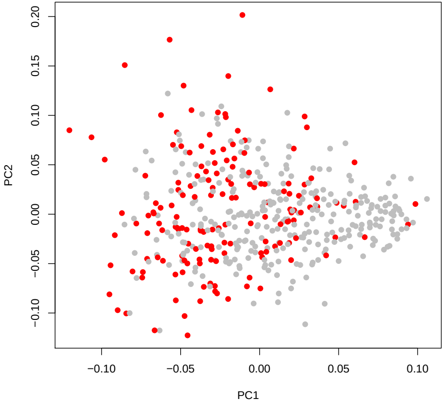
Note: the direction ($\pm$) of each vector is arbitrary

Interpretation

Thus, with some oversimplification:

- The first principal component distinguishes young, slim people with low cholesterol and blood pressure from old, overweight people with high blood pressure and cholesterol
- The second principal component primary reflects smoking and drinking
- The third principal component predominantly reflects stress
- The fourth principal component predominantly reflects family history

Plotting the principal components



Gray=Healthy, Red=CHD

Plotting the principal components (cont'd)