Chapter 16

Principal Components Analysis

Review: Previous Models

- Response y and a bunch of predictors x_j
- Strong correlations between predictors are a problem
- Hope to use a small number of fairly unrelated predictor variables

Highly Correlated Data

- Common to have highly related variables in data
- Examples:
  - Databases with many variables on similar product characteristics
  - Survey results where questions are highly related
  - Patient health data

PCA: Motivation

Start with large number of correlated variables

- Want to reduce to fewer dimensions (or variables)
- Want to retain large amount of the information from the original variables
- Would be nice to be able to pick features out of the original data

PCA: Methodology

- Start with continuous variables $x_1, \ldots, x_p$ containing some large correlations
- Perform eigen decomposition on correlation (or covariance) matrix of the x_i's
- Construct principal components z_j with
  - $z_j = a_{j1}\, x_1 + a_{j2}\, x_2 + \ldots + a_{jp}\, x_p$
  - (a_{j1},…, a_{jp}) is j^{th} eigenvector
  - z_j's are uncorrelated

Methodology Continued

- Have principal components $z_1, \dots, z_p$
- z_j's have same total variation as original x_i's
- Ordered from most variation to least
- Eigenvalue λ_j tells us about variation in the x_i's described by z_j
- Percentage of variation in z_j is $\frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k}$

Some Benefits of PCA

- Remove correlation
- Can describe large amount of original variation in fewer new “variables”
- Makes visualization and modeling easier
- Relationships between original variables and principal components may indicate underlying features

Some Trade-Offs

- Knew precisely what the original variables measured
- Don’t really know what the principal components represent
- Will throw away some percentage of the original information

Choosing Components to Keep

A few rules of thumb:

- Describe at least some certain percentage (say 70% or 90%) of total variation
- Keep components with larger than average eigenvalues
- Look for elbow in scree plot of eigenvalues

The Princomp Procedure

- Will generate eigen information
- Can be used for generating scree plots
- Can be used to generate score plots (show original data points on axes defined by principal components)
- Can create output data set containing principal component values

Example: US Crime Data

- From **Getting Started** example in **The Princomp Procedure** documentation
- Crime rates by type in each state in 1977
- High correlation between rates for different crimes shouldn’t be too surprising…

US Crime Data Analysis

- Correlation checks via scatter plots and correlation matrix
- Perform principal components analysis
- See how much variation described by first few components
- Try to interpret first few components from eigenvectors
- Obtain confidence ellipses to look for possible outlier states (assuming approx. normal z_j's)

Example: Decathlon Data

- Start with **olympic.dat** from text
- Contains:
  - Athlete's name
  - Times or distances for decathlon events
  - Overall score for decathlon
- Will look for underlying features in time and distance measurements
- Then look at relationships with **dscore**

Decathlon: Initial Processing

- Look for and remove extreme overall decathlon scores
- Change signs for events where smaller measurements are better (timed events)
- Then increases in all variables will be indicators of better performance

Decathlon: PCA

- Perform PCA on everything but **name** and **dscore**
- What features can we identify?
- Look at plots and at the correlation of the principal components with the **dscore** variable
- Relationships between the principal components and the **dscore**?

Exercise: Pain Survey

- 123 patients with extreme pain
- Patients rate 9 statements about their pain from 1 to 6 (disagreement to agreement)
- Data is a correlation matrix for their responses defined as a special data type

Exercise: PCA on Pain Data

- Perform a principal components analysis on this data
- How many components would we want to keep?
- What might the first few components represent?

Principal Component Regression

- Fit a linear regression model as a function of a few principal components
- Example is intended to introduce idea
- In practice need to beware of over-fitting (describing too much of the variation)
- Additional diagnostics may be needed to guard against that

Exercise: Decathlon PCA

- Use principal components as predictors for **dscore**
- Linear regression of **dscore** on **prin1** and **prin2**
- Comment on model (model fit, diagnostics,…)
- Repeat using all 10 principal components and forward selection with entry level .05
- Benefits and drawbacks of two models? (Consider interpretability and variation described)