

Day 9: Unsupervised learning and dimensional reduction

Kenneth Benoit and Slava Mikhaylov

Introduction to Data Science and Big Data Analytics

26 August 2015

Day 9 Outline

Dimensional reduction methods

Exam (p)Review

Parametric v. non-parametric methods

- ▶ **Parametric methods** model feature occurrence according to some stochastic distribution, typically in the form of a measurement model
 - ▶ for instance, model words as a multi-level Bernoulli distribution, or a Poisson distribution
 - ▶ feature effects and “positional” effects are unobserved parameters to be estimated
- ▶ **Non-parametric methods** typically based on the Singular Value Decomposition of a matrix
 - ▶ principal components analysis
 - ▶ correspondence analysis
 - ▶ other (multi)dimensional scaling methods

Dimensional reduction methods

Non-parametric dimensional reduction methods

- ▶ Non-parametric methods are algorithmic, involving no “parameters” in the procedure that are estimated
- ▶ Hence there is no uncertainty accounting given distributional theory
- ▶ Advantage: don't have to make assumptions
- ▶ Disadvantages:
 - ▶ cannot leverage probability conclusions given distributional assumptions and statistical theory
 - ▶ results highly fit to the data
 - ▶ not really assumption-free (if we are honest)

Principal Components Analysis

- ▶ For a set of features X_1, X_2, \dots, X_p , typically centred (to have mean 0)
- ▶ the **first principal component** is the normalized linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$$

that has the largest variance

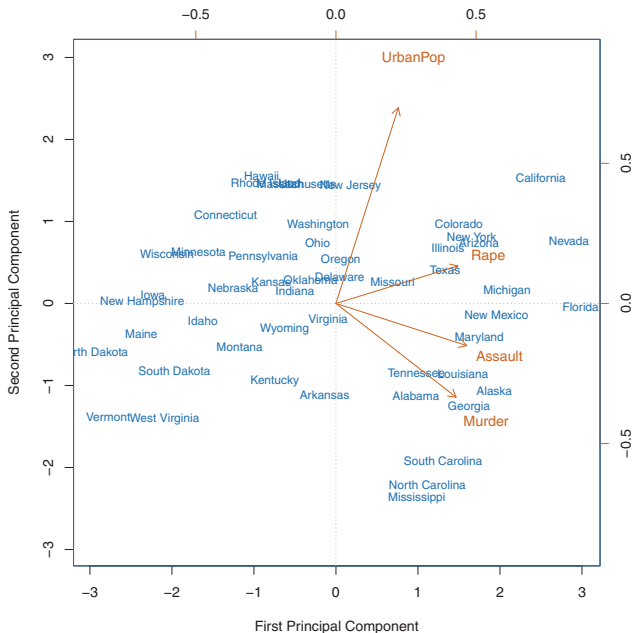
- ▶ **normalized** means that $\sum_{j=1}^p \phi_{j1}^2 = 1$
- ▶ the elements $\phi_{11}, \dots, \phi_{p1}$ are the **loadings** of the first principal component
- ▶ the second principal component is the linear combination Z_2 of X_1, X_2, \dots, X_p that has maximal variance out of all linear combinations that are *uncorrelated* with Z_1

PCA factor loadings example

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

TABLE 10.1. *The principal component loading vectors, ϕ_1 and ϕ_2 , for the USArrests data. These are also displayed in Figure 10.1.*

PCA factor loadings biplot



PCA projection illustrated

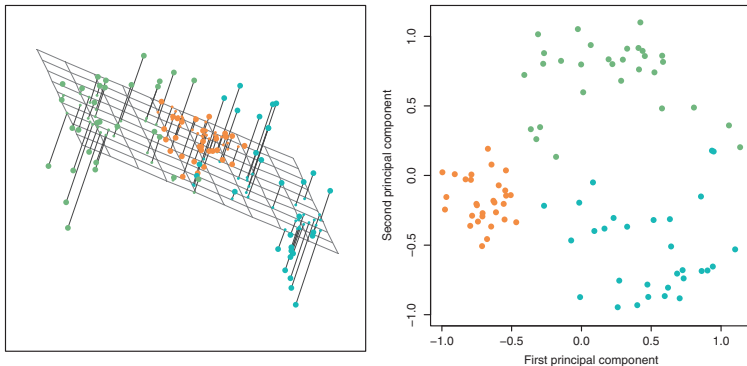


FIGURE 10.2. *Ninety observations simulated in three dimensions. Left: the first two principal component directions span the plane that best fits the data. It minimizes the sum of squared distances from each point to the plane. Right: the first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane. The variance in the plane is maximized.*

Correspondence Analysis

- ▶ CA is like factor analysis for categorical data
- ▶ Following normalization of the marginals, it uses Singular Value Decomposition to reduce the dimensionality of the word-by-text matrix
- ▶ This allows projection of the positioning of the words as well as the texts into multi-dimensional space
- ▶ The number of dimensions – as in factor analysis – can be decided based on the eigenvalues from the SVD

Exam (p)Review

Exam hints

- ▶ Structure
- ▶ Topic scope
- ▶ Format