Modern Applied Statistics Chap 11: Exploratory Multivariate Analysis

Yongdai Kim

October 27, 2022

Seoul National University

Outline

Introduction

Introduction

Multivariate analysis is concerned with datasets that have more than one response variable for each observational or experimental unit.

Summary

- X : n x p data matrices
  - row : observation
  - columns : variables
  - x : the variables of a case by the row vector

The main division in multivariate methods is between those methods that assume a given structure and those that seek to discover structure from the evidence of the data matrix alone. Methods for known structure are considered in Chapter 12.

Introduction

In pattern-recognition terminology the distinction is between **supervised** and **unsupervised** methods.

One of our examples is the (in)famous iris data collected by Anderson (1935) and given and analysed by Fisher (1936). This has 150 cases, which are stated to be 50 of each of the three species *Iris setosa, virginica and versicolor*. Each case has four measurements on the length and width of its petals and sepals.

A *priori* this seems a supervised problem, and the obvious questions are to use measurements on a future case to classify it, and perhaps to ask how the variables vary among the species. However, the classification of species is uncertain, and similar data have been used to identify species by grouping the cases.

Introduction

Krzanowski (1988) and Mardia, Kent and Bibby (1979) are two general references on multivariate analysis. For pattern recognition we follow Ripley (1996), which also has a computationally-informed account of multivariate analysis.

Most of the emphasis in the literature and in this chapter is on continuous measurements, but we do look briefty at multi-way discrete data in Section 11.4.

Colour can be used very effective1y to differentiate groups in the plots of this chapter, on screen if not on paper. The code given here uses both colours and symbols, but you may prefer to use only one of these to differentiate groups.

Running example 1: *Iris* data

- iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris.
- The species are Iris setosa, versicolor, and virginica.
- Variables : Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species

Running example 2: *Leptograpsus variegatus* crabs data

- Leptograpsus variegatus crabs data set gives body measurements, respectively, for 50 crabs from each of 2 species and 2 sex of crabs
- The species are blue and oragne.
- Variables : sp(B / O), sex(M / F, FL(frontal lobe size), RW(rear width), CL(carapace length), CW(carapace width), BD(body depth)

Visualization Methods

Visualization Methods

The simplest way to examine multivariate data is via a *pairs* or *scatterplot matrix* plot. Pairs plots are a set of two-dimensional projections of a high-dimension point cloud.

However, pairs plots can easily miss interesting structure in the data that depends on three or more of the variables.
Genuinely multivariate methods explore the data in a less coordinate-dependent way.

Visualization Methods

Many of the most effective routes to explore multivariate data use dynamic graphics such as exploratory projection pursuit which chooses 'interesting' rotations of the point cloud.
These are available through interfaces to the package **XGobi** for machines running $X11$. A successor to **XGobi**, **GGobi**, is under development.

Many of the other visualization methods can be viewed as projection methods for particular definitions of 'interestingness'.

Principal Component Analysis

Principal component analysis(PCA)

- PCA has a number of different interpretations.
- The simplest is a projection method finding projections of maximal variability. That is, it seeks linear combinations of the columns of X with maximal (or minimal) variance.

Principal Component Analysis

Let S denote the covariance matrix of the data X

$$nS = (X - n^{-1}\mathbf{1}\mathbf{1}^T X)^T (X - n^{-1}\mathbf{1}\mathbf{1}^T X) = (X^T X - n\bar{x}\bar{x}^T)$$

where $\bar{x} = \mathbf{1}^T X/n$ is the row vector of means of the variables. Then the sample variance of a linear combination $\mathbf{xa}$ of a row vector $\mathbf{x}$ is $\mathbf{a^T \Sigma a}$ and this is to be maximized (or minimized) subject to $\|\mathbf{a}\|^2 = \mathbf{a^T a} = 1$.

Since Σ is a non-negative definite, it has an eigendecomposition

$$\Sigma = C^T \Lambda C$$

where Λ is a diagonal matrix of (non-negative) eigenvalues in decreasing order.

Principal Component Analysis

Let $\mathbf{b} = C\ \mathbf{a}$, then $\|\mathbf{b}\|^2 = \|\mathbf{a}\|^2$ ($\because$ C is orthogonal)

The problem is then equivalent to maximizing $\mathbf{b^T}\Lambda\mathbf{b} = \sum \lambda_i b_i^2$ subject to $\sum b_i^2 = 1$.

Clearly the variance is maximized by taking $\mathbf{b}$ to be the first unit vector, or equivalently taking $\mathbf{a}$ to be the column eigenvector corresponding to the largest eigenvalue of Σ.

The i th principal component is then the i th linear combination picked by this procedure.

Principal Component Analysis

The first k principal components span a subspace containing the 'best' k-dimensional view of the data. It has a maximal covariance matrix.

It also best approximates the original points in the sense of minimizing the sum of squared distances from the points to their projections. The first few principal components are often useful to reveal structure in the data.

The principal components corresponding to the smallest eigenvalues are the most nearly constant combinations of the variables, and can also be of interest.

Principal Component Analysis

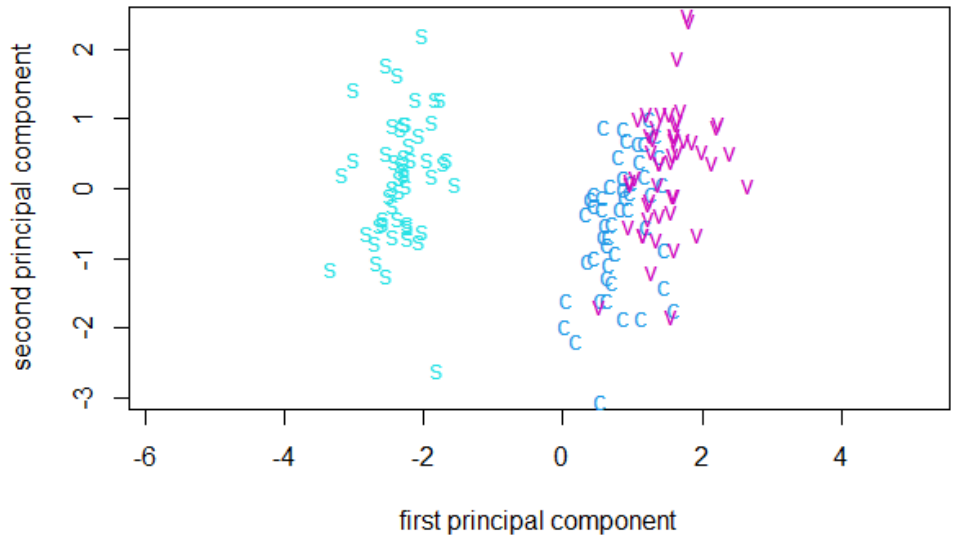

Figure 1: First two principal components for the log-transforrmed iris data.

Figure 1 shows the first two principal components for the *iris* data based on the covariance matrix, revealing the group structure if it had not already been known.

Principal Component Analysis

A *warning*: principal component analysis will reveal the gross features of the data, which may already be known, and is often best applied to residuals after the known structure has been removed.

Cluster Analysis

Factor Analysis

Discrete Multivariate Analysis