Topic 1: Exploratory data analysis

Welcome to Week 1

Dr Pavel Krivitsky gives you a brief overview of topics and concepts we'll be covering in this week.

Transcript

Weekly learning outcomes

- Describe the basic properties of multivariate normal distribution.
- Diagnose deviations from normality in real-world datasets.
- Obtain and plot point estimates and confidence intervals for vector means and differences of vector means.
- Test linear hypotheses about vector means and differences of vector means.
- Explain the assumptions underlying the above inferential procedures.
- Use R packages to perform analyses.

Topics we will cover are:

- Topic 1: Exploratory data analysis of multivariate data
- Topic 2: The multivariate normal distribution
- Topic 3: Estimation of vector mean and of variance matrices: point estimates
- Topic 4: Confidence intervals and hypothesis tests for the mean vector

Optional readings

An alternative presentation of the concepts for this week can be found in:

Johnson, R. A., & Wichern, D. (2008). *Applied Multivariate Statistical Analysis* (6th ed.). Pearson Prentice Hall.

- 1.3–1.4
- 4.1–4.2, 4.6
- 5.1–5.5 and 6.1–6.3

Härdle, W. K., & Simar, L. (2014). *Applied Multivariate Statistical Analysis* (4th ed.). Springer.

- 1.4–1.7
- 3.1–3.3

- 4.4, 5.1–5.3, 6.1
- 7.1–7.2

All readings are available from the course Leganto reading list. Please keep in mind that you will need to be logged into Moodle to access the Leganto reading list.

Questions about this week's topics?

This week's topics were prepared by Dr P. Krivitsky. If you have any questions or comments, please post them under Discussion or email directly: p.krivitsky@unsw.edu.au

Exploratory data analysis of multivariate data

Introduction

We begin by taking a look at how to summarise multivariate data—with a focus on quantitative data—and to visualise it.

Data organisation

Assume, we are dealing with $p \geq 1$ *variables*. The values of these variables are all recorded for each distinct *item*, *individual*, or *experimental trial*. Each of these three words will be substituted sometimes by the word "case". We will use the notation x_{ij} to indicate a particular value of the ith variable that is observed on the jth case. Consequently, n measurements on p variables can be represented in a form of a matrix

$$\overset{p\times n}{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pj} & \cdots & x_{pn} \end{pmatrix} \tag{1.1}$$

The matrix X above contains the data consisting of all the observations on all the variables. This way of representing the data allows easy manipulations to be performed in order to obtain some easy descriptive statistics for each of the variables.

Basic summaries

For example, the *sample mean* of the second variable is just $\bar{x}_2 = \frac{1}{n}\sum_{j=1}^{n} x_{2j}$, the *sample variance* of the second variable is just $s_2^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2$ (Note that for the sample variance we shall sometimes use the divisor of $n-1$ rather than n and each time this will be differentiated by displaying the appropriate expression).

The *sample covariance* (the simple measure of linear association between variables 1 and 2) is given by $s_{12} = \frac{1}{n}\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2)$ and one can understand easily how $s_{ik}, i = 1, 2, \ldots, p$, $k = 1, 2, \ldots, p$ can be defined. Finally, the *sample correlation coefficient* (the measure of linear association between two variables that does not depend on the units of measurement) can be defined. The sample correlation coefficient of the ith and kth variables is defined by $r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$. Because of the well-known Cauchy–Bunyakovsky–Schwartz Inequality, $|r_{ik}| \leq 1$ holds. Note also that $r_{ik} = r_{ki}$ for all $i = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$ holds.

It should be repeatedly noted that the sample correlations and covariance are useful only when trying to measure the *linear* association between two variables. Their value is less informative and is misleading in cases of *nonlinear* association. In this case one needs to invoke the *correlation quotient* instead.

Since covariance and correlation coefficients are routinely calculated and analysed they are very widely used and provide useful numerical summaries of association when the data do not exhibit obvious nonlinear patterns of association.

The descriptive statistics that we discussed until now are usually organised into arrays, namely:

- **Vector of sample means** $\bar{x} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{pmatrix}^\top$
- **Matrix of sample variances and covariances**

$$\overset{p\times p}{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix} \tag{1.2}$$

- **Matrix of sample correlations**

$$\overset{p\times p}{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} \tag{1.3}$$

Visualisation

Initial graphical representation of the data

Some simple characteristics of the data are worth studying before the actual multivariate analysis would begin:

- drawing a scatterplot of the data
- calculating simple univariate descriptive statistics for each variable
- calculating sample correlation and covariance coefficients
- linking multiple two-dimensional scatterplots

R

In R, these are implemented in `base::rowMeans`, `base::colMeans`, `stats::cor`, `graphics::plot`, `graphics::pairs`, `GGally::ggpairs`. Here, the format is `PACKAGE::FUNCTION`, and you can learn more by running

```
library(PACKAGE)
? FUNCTION
```

Next, work through the basic multivariate summaries and visualisation example in the next section.

Example: R activity on basic multivariate summaries and visualisation

The first practice example will demonstrate basic multivariate summary statistics and graphics. You can use the RStudio Console here or complete the exercise in your own RStudio.

Transcript

To complete this task select the 'Multivariate_exploration_visualisation.Rmd' in the 'Files' section of RStudio. Follow the example contained within the RMD file.

If you choose to complete the example in your own RStudio, upload the following file:

Multivariate_exploration_visualisation.Rmd

The output of the RMD file is also displayed below: