MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 1: Introduction, Multivariate Location and Scatter

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Practical Things

Practical Things

- Lecturer: Pauliina Ilmonen, pauliina.ilmonen(a)aalto.fi
- Lectures on Mondays 12.15-14.00 (2.1. - 6.2., 20.2. - 27.3.), U147 (U5)
- Exercises by: Niko Lietzen, niko.lietzen(a)aalto.fi and Sami Helander, sami.helander(a)aalto.fi
- There are two exercise groups, choose the one that fits better to your schedule
- Weekly exercises Group 1 on Thursdays 16.15-18.00 (5.1. - 9.2., 23.2. - 30.3.), U344 (Sami)
- Weekly exercises Group 2 on Fridays 10.15-12.00 (6.1. - 10.2., 24.2. - 31.3.), U257 (Niko)
- Book: K. V. Mardia, J. T. Kent, J. M. Bibby: Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979)

Self Study

Make sure that you know how to calculate the univariate means, medians, variances, and max and min values. Familiarize yourself with the correlation coefficients and common graphical presentations (boxplots, scatter plots, histograms, bar plots, pie charts) of data. Make sure that you know what is a cumulative distribution function, a probability density function, and a probability mass function. Make sure that you know what is the expected value of a random variable. Read about univariate and multivariate normal distributions and elliptical distributions. Make sure that you know what is meant by central symmetric distributions and skew distributions.

How to pass this course?

You are expected to

- Attend the lectures and be active - not compulsory, no points, but highly recommended.
- Submit your project work on time - THIS IS COMPULSORY - max 6 points.
- Take the exam - max 24 points. (The next examinations are on Thursday 6.4. at 9-12 and on Wednesday 24.5. at 16.30-19.30.)
- Submit your homework exercises on time - not compulsory, but highly recommended - max 3 points.
- Participate to weekly exercises (group 1 OR group 2) - not compulsory, but highly recommended - max 3 points.

Max total points $= 6 + 24 + 3 + 3 = 36$. You need at least 16 points in order to pass the course.

How to get a good grade?

- Attend the lectures and be active!
- Work hard on your project work.
- Be active in exercises!
- Study for the exam!

Grading is based on the total points as follows: 16p -> 1, 20p -> 2, 24p -> 3, 28p -> 4, 32p -> 5.

Project Work

Find a multivariate (at least 3-variate) dataset (tilastokeskus, OECD, collect yourself, ...), set a research question, and perform multivariate analysis. Write a report (max 10 pages), print it and submit it to Sami, Niko or Pauliina on Wed 5.4.2016 the latest!

Goals of the project work:

- Description of the research questions
- Description of the dataset
- Univariate and bivariate statistical analysis to present the variables
- Application of multivariate statistical methods to answer research questions (justification and output)
- Conclusions and answers to the question raised at the beginning
- Critical evaluation of the analysis

Remember that No findings is a finding! Note that you will automatically get 0 points from the exam if you will not return your project work on time!

Introduction

Introduction

The first step of all statistical analysis is the univariate and bivariate analysis. First calculate the univariate means, medians, variances, max and min values. Then calculate the correlation coefficients. And take a look at your data — literally! Make histograms of continuous variables and pie charts of categorical variables. Make boxplots to detect univariate outliers, and make scatter plots to detect bivariate structures.

Note that visualization is not always easy when the data contains a large number of individuals, but do not skip plotting your data! It is very important that you get familiar with your data before you conduct any large multivariate analysis.

Multivariate Location and Scatter

Multivariate Data

Let x denote a p-variate random vector with a cumulative distribution function F_x and let $X = [x_1 .... x_n]$, where $x_1, ..., x_n$ are i.i.d. observations from the distribution F_x.

Location Functionals

Definition

A $p \times 1$ vector-valued functional $T(F_x)$, which is affine equivariant in the sense that

$$T(F_{Ax+b}) = AT(F_x) + b$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b, is called a location functional.

Scatter Functionals

Definition

A $p \times p$ matrix-valued functional $S(F_x)$ which is positive definite and affine equivariant in the sense that

$$S(F_{Ax+b}) = AS(F_x)A^T$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b, is called a scatter functional.

Location and Scatter Estimates

The corresponding sample statistics are obtained if the functionals are applied to the empirical cumulative distribution F_n based on a sample $x_1, x_2, \ldots, x_n$. Notation $T(F_n)$ and $S(F_n)$ or $T(X)$ and $S(X)$ is used for the sample statistics. The location and scatter sample statistics then also satisfy

$$T(AX + b1_n^T) = AT(X) + b$$

and

$$S(AX + b1_n^T) = AS(X)A^T$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b.

Scatter Functionals

Scatter matrix functionals are usually standardized such that in the case of standard multivariate normal distribution $S(F_x) = I$.

Shape Functionals

Definition

If a positive definite $p \times p$ matrix-valued functional $S(F_x)$ satisfies that $S(F_{Ax+b})$ is proportional to $AS(F_x)A^T$ for all nonsingular $p \times p$ matrices A and for all p-vectors b, then $S(F_x)$ is called a shape functional.

Traditional Functionals

The first examples of location and scatter functionals are the mean vector and the regular covariance matrix:

$$T_1(F_x) = E(x) \text{ and } S_1(F_x) = Cov(F_x) = E\left((x - E(x))(x - E(x))^T\right).$$

The Sample Mean Vector and the Sample Covariance Matrix

Traditional estimates of the mean vector and the covariance matrix are calculated as follows:

$$T_1(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$S_1(X) = Cov(X) = \frac{1}{n-1}\sum_{i=1}^{n}\left((x_i - T_1(X))(x_i - T_1(X))^T\right).$$

Why do we need other location and scatter measures???

Scatter Functionals

There are several other location and scatter functionals, even families of them, having different desirable properties (robustness, efficiency, limiting multivariate normality, fast computations, etc).

Moment Based Functionals

Location and scatter functionals can be based on the third and fourth moments as well. A location functional based on third moments is

$$T_2(F_x) = \frac{1}{p} E\left((x - E(x))^T Cov(F_x)^{-1}(x - E(x))x\right)$$

and a scatter matrix functional based on fourth moments is

$$S_2(F_x) = \frac{1}{p+2} E\left((x - E(x))(x - E(x))^T Cov(F_x)^{-1}(x - E(x))(x - E(x))^T\right).$$

Example 1: Bivariate Normal Distribution

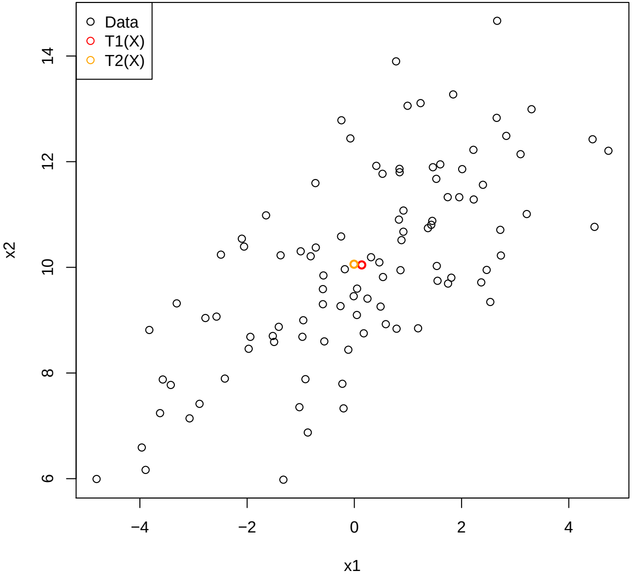
In this example we consider bivariate normal distribution $N(\mu, \mathrm{A})$, where

$$\mathrm{A} = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$$

and

$$\mu = \begin{bmatrix} 0 & 10 \end{bmatrix}.$$

Example 1: Bivariate Normal Distribution



Example 1: Bivariate Normal Distribution

We simulated 100 samples from $N(\mu, \mathrm{A})$ and we then calculated the sample mean vector $T_1(X)$, the location vector based on third moments $T_2(X)$, the sample covariance matrix $S_1(X)$ and the scatter matrix based on fourth moments $S_2(X)$ of each sample. In order to compare $T_1(X)$, $T_2(X)$, $S_1(X)$, and $S_2(X)$, we calculated the means of the estimates.

$$T_1(X): \begin{bmatrix} 0.006703295 \\ 10.001765054 \end{bmatrix} \qquad T_2(X): \begin{bmatrix} 0.01626947 \\ 9.99082058 \end{bmatrix}$$

$$S_1(X): \begin{bmatrix} 4.029396 & 2.034711 \\ 2.034711 & 2.968536 \end{bmatrix} \qquad S_2(X): \begin{bmatrix} 3.9197916 & 2.003406 \\ 2.003406 & 2.924344 \end{bmatrix}$$

Both location estimates seem to estimate the parameter μ and both scatter estimates seem to estimate the parameter A.

Example 2: Independent Components, Skewed Distributions

In this example we consider $\mathrm{Gamma}(\alpha, \beta)$ and $\chi^2(k)$ distributions, where

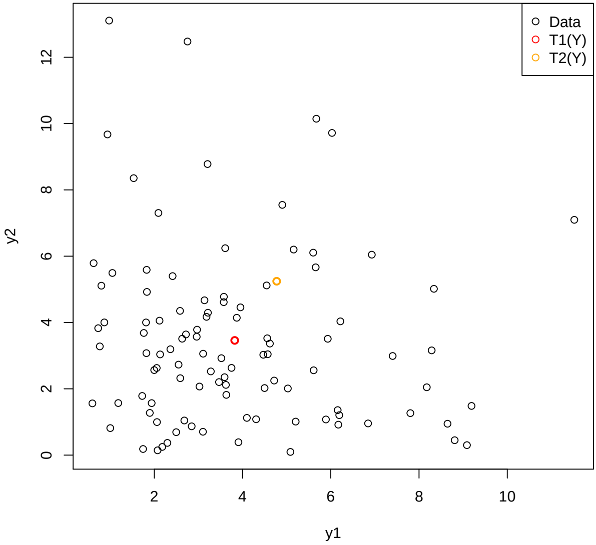
$$\alpha = 2,$$

$$\beta = 0.5$$

and

$$k = 3.$$

Example 2: Independent Components, Skewed Distributions



Example 2: Independent Components, Skewed Distribution

As in Example 1, we ran the simulation 100 times and calculated the means.

$$T_1(Y): \begin{bmatrix} 4.031022 \\ 2.964918 \end{bmatrix} \qquad T_2(Y): \begin{bmatrix} 5.944029 \\ 4.740199 \end{bmatrix}$$

$$S_1(Y): \begin{bmatrix} 8.16111692 & 0.04234064 \\ 0.04234064 & 5.76640662 \end{bmatrix} \qquad S_2(Y): \begin{bmatrix} 13.4080726 & 0.1142734 \\ 0.1142734 & 9.8194396 \end{bmatrix}$$

Here the location estimates differ significantly from each other. Also the scatter estimates differ significantly from each other. Note also that the off-diagonal elements of both scatter estimates are small.

Location and Scatter Functionals Under Symmetry Assumptions

We now consider the behavior of scatter and location functionals under some symmetry assumptions.

Theorem

Under the assumption of central symmetry, all location functionals are equal to the center of symmetry.

Proof:...

Theorem

Under the assumption of multivariate elliptical distribution, all scatter functionals are proportional.

Proof:...

Note that in general different location functionals do not measure the same population quantities. That is true also for scatter functionals — different scatter functional do not necessarily measure the same population quantities!

Next Week

Next week we will talk about principal component analysis (PCA).

References

References I

- K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).
- H. Oja, Multivariate Nonparametric Methods With R, Springer-Verlag, New York, 2010.

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.