

MS-E2112 Multivariate Statistical Analysis (5cr)

Lecture 1: Introduction, Multivariate Location and Scatter

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location and Scatter

References

Practical Things

Introduction

Multivariate Location
and Scatter

References

Practical Things

Practical Things

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

- Lecturer: Pauliina Ilmonen, pauliina.ilmonen(a)aalto.fi
- The first lecture is on January 10th at 12.15-14.00
- Exercises: Jaakko Pere, jaakko.pere(a)aalto.fi and Nourhan Shafik, nourhan.shafik(a)aalto.fi
- There are four exercise groups, choose the one that fits best to your schedule

Self Study

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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?

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

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 course examinations is on Thursday 14.4.)
- Participate to weekly exercises (group 1, group 2, group 3 OR group 4) - not compulsory, but highly recommended - max 3 points.
- Be ready to present your homework solutions in the exercise group - 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.

Practical Things

Introduction

Multivariate Location
and Scatter

References

Exercises

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Participate to weekly exercises (group 1, group 2, group 3 OR group 4) - not compulsory, but highly recommended - max 3 points. If you attend 3-5 times, you get 1 point. If you attend 6-8 times, you get 2 points. If you attend at least 9 times (out of 11 times), you get 3 points.

In order to earn the exercise points, you have to arrive on time to the exercise session. The names of the participants are collected at the beginning of each exercise class. You can not get any exercise points without attending the exercises.

Exercise session 11 is reserved for the project work and for summarizing the contents of the course.

Attending all the exercise sessions, including the last one, is highly recommended!

Practical Things

Introduction

Multivariate Location
and Scatter

References

Homework

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

Solve the homework problems and be ready to present your solutions in the exercise group - not compulsory, but highly recommended - max 3 points. **Note that your solution does not have to be perfect or even correct — trying your very best is enough!** If you solve your homework assignments 3-5 times, you get 1 point. If you solve your homework assignments 6-8 times, you get 2 points. If you solve your homework assignments at least 9 times (out of 10 times), you get 3 points.

In order to earn the homework points, you have to arrive on time to the exercise session and write your name to the homework list. You can not get any homework points without attending the exercises.

Project Work

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Find a multivariate (at least 3-variate) dataset (Statistics Finland (=Tilastokeskus), OECD, collect yourself, ...), set a research question, and perform multivariate analysis. Write a report (max 10 pages), and submit it in MyCourses before Monday 11.4.2021 at 12.00 (midday).

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 your chosen 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 submit your project work on time!

Practical Things

Introduction

Multivariate Location
and Scatter

References

How to get a good grade?

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

- 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.

Practical Things

Introduction

Multivariate Location
and Scatter

References

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.

Practical Things

Introduction

Multivariate Location
and Scatter

References

Multivariate Location and Scatter

Let x denote a p -variate random vector with a cumulative distribution function F_x . Let X denote a $n \times p$ data matrix of independent and identically distributed (i.i.d.) observations x_1, x_2, \dots, x_n from the distribution F_x .

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**.

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

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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, \dots, 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

$$T(XA^T + \mathbf{1}_n b^T) = AT(X) + b$$

and

$$S(XA^T + \mathbf{1}_n b^T) = AS(X)A^T$$

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

Scatter Functionals

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

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**.

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

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

The Sample Mean Vector and the Sample Covariance Matrix

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

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

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

Scatter Functionals

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

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

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

and a scatter matrix functional based on fourth moments is

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

Example 1: Bivariate Normal Distribution

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

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

and

$$\mu = [0 \quad 10].$$

Example 1: Bivariate Normal Distribution

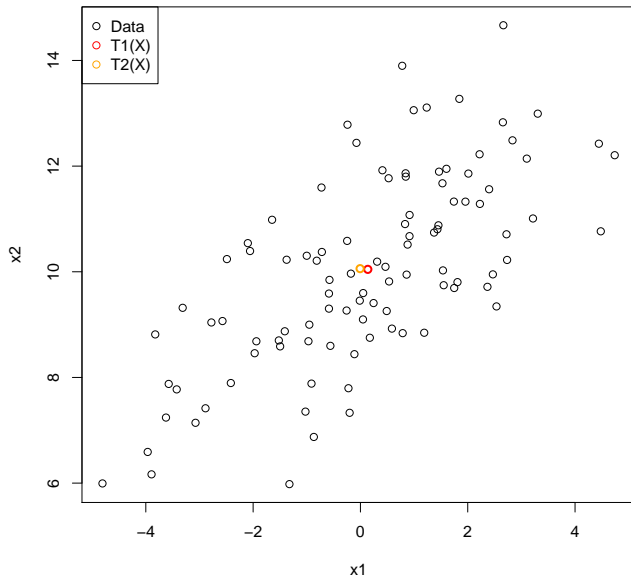
Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References



Example 1: Bivariate Normal Distribution

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

We simulated 100 samples from $N(\mu, A)$ 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.

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

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

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

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

$$\alpha = 2,$$

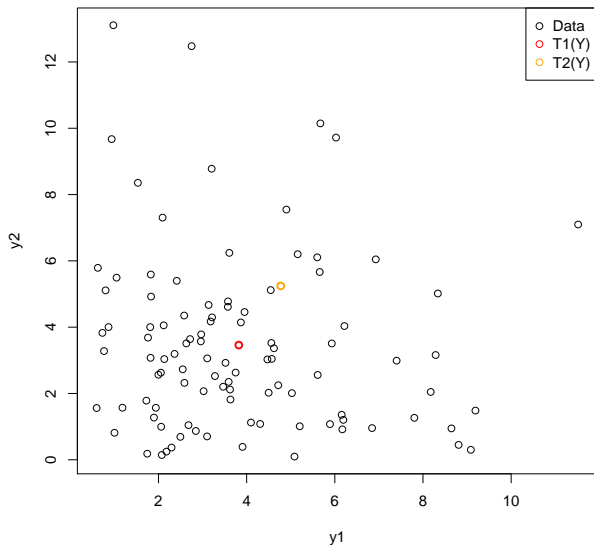
$$\beta = 0.5$$

and

$$k = 3.$$

Example 2: Independent Components, Skewed Distributions

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala



Practical Things

Introduction

Multivariate Location
and Scatter

References

Example 2: Independent Components, Skewed Distribution

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

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

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

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

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

Theorem

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

Let x denote a p -variate random vector with a cumulative distribution function F_x . Let $\theta \in \mathbb{R}^p$ and assume that $x - \theta \sim -(x - \theta)$. Let T be an affine equivariant location functional and assume that $T(F_x)$ exists as finite quantity.

Since T is affine equivariance and since x is symmetric about θ , we have that

$$T(F_x) - \theta = T(F_{x-\theta}) = T(F_{-(x-\theta)}) = T(F_{-x+\theta}) = -T(F_x) + \theta.$$

Thus

$$2T(F_x) = 2\theta$$

and it follows that

$$T(F_x) = \theta.$$

Since T was an arbitrarily chosen location functional, this completes the proof.

Theorem

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

Let x denote a p -variate random vector with a cumulative distribution function F_x . Assume that

$$x = \Omega z + \mu,$$

where $\mu \in \mathbb{R}^p$, $\Omega \in \mathbb{R}^{p \times p}$, Ω is full rank, and $z \sim Oz$ for all orthogonal $O \in \mathbb{R}^{p \times p}$. Let S be an affine equivariant scatter functional and assume that $S(F_x)$ exists as finite quantity.

Proof (2/3)

Since $z \sim Oz$ for all orthogonal $O \in \mathbb{R}^{p \times p}$, it holds that $z \sim PJz$ for all permutation matrices $P \in \mathbb{R}^{p \times p}$ and for all sign change matrices $J \in \mathbb{R}^{p \times p}$. Now it follows from affine equivariance of S that

$$S(F_{PJz}) = PJS(F_z)(PJ)^T$$

and from the property $PJz \sim z$ that

$$S(F_{PJz}) = S(F_z).$$

Thus

$$S(F_z) = PJS(F_z)(PJ)^T.$$

As $S(F_z) = PJS(F_z)(PJ)^T$ holds for all permutation matrices $P \in \mathbb{R}^{p \times p}$ and for all sign change matrices $J \in \mathbb{R}^{p \times p}$, we have that

$$(S(F_z))_{ij} = -(S(F_z))_{ji}, \quad i \neq j$$

and

$$(S(F_z))_{ii} = (S(F_z))_{jj}.$$

Thus

$$S(F_z) \propto I.$$

Proof (3/3)

It now follows from above and from affine equivariance of S that

$$S(F_x) = S(F_{\Omega Z + \mu}) = \Omega S(F_z) \Omega^T = \Omega c \cdot I \Omega^T = c \Omega \Omega^T,$$

where c is a constant that may depend on S .

Since S was an arbitrarily chosen scatter functional, this completes the 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

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

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

Practical Things

Introduction

Multivariate Location
and Scatter

References

References

References I



Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

-  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

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Practical Things

Introduction

Multivariate Location
and Scatter

References

-  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.