Course guide
200606 - AMD - Multivariate Data Analysis

Last modified: 01/06/2023

Unit in charge: School of Mathematics and Statistics
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.
1004 - UB - (ENG)Universitat de Barcelona.

Degree: MASTER'S DEGREE IN STATISTICS AND OPERATIONS RESEARCH (Syllabus 2013). (Optional subject).

Academic year: 2023 **ECTS Credits:** 5.0 **Languages:** Spanish, English

LECTURER

Coordinating lecturer: MIQUEL SALICRÚ PAGES

Others: Segon quadrimestre:
JAMIE ARJONA MARTINEZ - A
FERRAN REVERTER COMES - A
MIQUEL SALICRÚ PAGES - A

PRIOR SKILLS

1. This course presupposes knowledge of linear algebra: diagonalization of a symmetric matrix, vector projection, vector derivation of linear and quadratic functions.
2. It is also necessary to have successfully completed a course on statistical inference covering the classical univariate tests (Student's t test, Fisher's F test).

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
1. CE-3. Ability to formulate, analyze and validate models applicable to practical problems. Ability to select the method and / or statistical or operations research technique more appropriate to apply this model to the situation or problem.
2. CE-6. Ability to use appropriate software to perform the necessary calculations in solving a problem.
3. CE-9. Ability to implement statistical and operations research algorithms.
5. CE-2. Ability to master the proper terminology in a field that is necessary to apply statistical or operations research models and methods to solve real problems.
6. CE-4. Ability to use different inference procedures to answer questions, identifying the properties of different estimation methods and their advantages and disadvantages, tailored to a specific situation and a specific context.

Transversal:
4. TEAMWORK: Being able to work in an interdisciplinary team, whether as a member or as a leader, with the aim of contributing to projects pragmatically and responsibly and making commitments in view of the resources that are available.

7. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

TEACHING METHODOLOGY

Language: the first part of the course (50%) will be taught in English, and the second part (50%) will be taught in Spanish.
Theoretical sessions: conventional lecture classes according to the schedule made known at the start of the course.
Problems: problems serve to underpin the theoretical concepts addressed in the theory sessions. Students are asked to hand in some problems during the course.
Practicals: the facilities of matrix programming are employed to carry out a multivariate analysis. Practical work is assessed. The R programming language is used. Practicals are done in groups of two students.
Project: students work on the multivariate analysis of a particular database using the methods taught in this course. The project is carried out by groups of two students. Each groups writes a report about their project and hands this in.

LEARNING OBJECTIVES OF THE SUBJECT

A student that has successfully completed the course will be able to:

1. Recognize the multivariate nature of a particular database.
2. Explain the advantage of a multivariate approach over a traditional univariate approach.
3. Explain the aims of the most commonly used multivariate methods (principal component analysis, correspondence analysis, factor analysis, multidimensional scaling, MANOVA, discriminant analysis, cluster analysis, etc.).
4. Identify the most appropriate multivariate method for the analysis of a particular database.
5. Implement the most basic multivariate methods using matrix calculations in the R environment.
6. Apply multivariate descriptive statistics to a set of variables.
7. Apply the basic principles of dimension reduction.
8. Apply the necessary transformation for a particular analysis (selection of the metric).
9. Perform multivariate visualization of data sets on the computer.
10. Interpret visual representations (biplots) of multivariate data sets.
11. Explain the multivariate normal distribution and its properties.
12. Give the definition of the most basic multivariate statistical tests.
13. Apply the most common multivariate hypothesis tests regarding mean vectors and covariance matrices.
14. Apply linear and quadratic discriminant analysis to data stemming from different populations, obtaining the discriminant functions under the assumption of multivariate normality, and classify the individuals of unknown group status.
15. Enumerate the basic clustering methods.
16. Apply different algorithms for creating clusters.
17. Interpret the results of the most commonly used multivariate methods.
18. Apply factor analysis and extract the common dimensions of a set of variables.
19. Apply repeated measurement analysis, profile analysis, and two-way MANOVA.

STUDY LOAD

Type	Hours	Percentage
Hours small group	15,0	12.00
Hours large group	30,0	24.00
Self study	80,0	64.00

Total learning time: 125 h

CONTENTS

Multivariate descriptive statistics

Description:
1. Introduction and basic concepts. A review of linear algebra. The geometry of the sample. The cloud of points in Rp i Rn. Metric. Measures of variability. M-ortogonal projection. Eigenvalue-eigenvector decomposition. Generalized singular value decomposition. Graphical representations, the biplot.
2. Prinicipal component analysis (PCA). Components definition. Properties. PCA based on a covariance matrix and on a correlation matrix. Biplots. Goodness of fit.
3. Multidimensional scaling (MDS). Distances and metrics. Euclidian representation of a distance matrix. Associated spectral decomposition. Goodness of fit.
4. Simple correspondence analysis. Contingency tables. Row and column profiles. Inertia and the chi-square statistic. Biplots.
5. Multiple correspondence analysis (MCA). MCA based on the Burt matrix. MCA based on the indicator matrix. Adjusted inertias. Grafical representations.
6. Factor analysis. The factor analysis model. Common and specific factors. Estimation methods: principal factor analysis and maximum likelihood. Graphical representation.
7. Canonical correlation analysis. Objective function. Canonical correlations, variables and weights. Relationships with other methods. Biplots.

Specific objectives:
Perform a multivariate descriptive analysis, both graphically and numerically, for quantitative and categorical data tables.

Related activities:
Several practicals, problems and the project of the course.

Full-or-part-time: 61h
Theory classes: 15h
Practical classes: 6h
Self study : 40h

Multivariate statistical inference.

Description:
Multivariate normal distribution. Sampling statistics. Likelihood ratio test. Covariance matrix testing. Intersection-union test. Hotelling's T2. Tests on the mean vector. Repeated measures analysis. Profile analysis. Comparison of different means. Wilks' lambda. The MANOVA model with one and two factors.

Specific objectives:
Apply multivariate statistical inference.

Related activities:
Practicals and problems.

Full-or-part-time: 29h
Theory classes: 9h
Self study : 20h

Discriminant analysis and cluster analysis.

Description:
1. Discriminant analysis. Parametric discriminant analysis. Discriminant functions. Linear and quadratic discriminant analysis.
2. Cluster analysis. Distances and similarity. Algorithms. Hierarchic methods and partitioning methods. Dendrogram. Ultrametric property. Ward's criterion.

Specific objectives:
Apply discriminant analysis and cluster analysis and the interpret results of these methods.

Related activities:
Practicals and problems.

Full-or-part-time: 32h
Theory classes: 7h 30m
Practical classes: 4h 30m
Self study : 20h

GRADING SYSTEM

Assessment is based on two exams, one midterm exam halfway the course and the other at the end of the course. Practical, problems and project are also assessed. The final course grade is based on the exam results (70%) and on the problems, practicals and a project (30%). The final grade for the course is a weighted mean of the different parts: exams (final exam 70% of which 35% corresponds to the first partial, and 35% to the second partial), practicals and assignments (15%), project (15%, a written report). Students who pass the first partial will not be evaluated for the corresponding materials at the final exam.

BIBLIOGRAPHY

Basic:
- Aluja, T.; Morineau, A. Aprender de los datos : el análisis de componentes principales. EUB, 1999. ISBN 8483120224.
- Johnson, R. A.; Wichern, D.W. Applied multivariate statistical analysis. 6th ed. Harlow, Essex: Pearson Education Limited, 2014. ISBN 9781292037578.
- Krzanowski, W. J. Principles of multivariate analysis : a user's perspective. Oxford University Press, 2000. ISBN 0198507089.
- Lebart, L.; Morineau, A.; Piron, M. Statistique exploratoire multidimensionnelle. 2e éd. Dunod, 1997. ISBN 2100040014.
- Peña Sánchez de Rivera, Daniel. Análisis de datos multivariantes [on line]. McGraw-Hill, 2002 [Consultation: 05/07/2023]. Available on: https://www-ingebook-com.recursos.biblioteca.upc.edu/ib/NPcd/IB_BooksVis?cod_primaria=1000187&codigo_libro=4203. ISBN 8448136101.

Complementary:
- Cuadras, C. M. Métodos de análisis multivariante. 2ª ed. PPU, 1991. ISBN 8476657714.
- Dillon, W. R.; Goldstein, M. Multivariate analysis methods and applications. John Wiley and Sons, 1984. ISBN 0471083178.
- Mardia, K. V.; Kent, J.T.; Bibby, J.M. Multivariate analysis. Academic Press, 1979. ISBN 0124712525.
- Morrison, D. F. Multivariate statistical methods. 3rd ed. McGraw-Hill, 1990. ISBN 0070431876.
- Volle, Michel. Analyse des données. 3e éd. Economica, 1985. ISBN 2717808159.
- Everitt, Brian. An R and S-PLUS companion to multivariate analysis [on line]. London: Springer, 2005 [Consultation: 05/07/2023]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/b138954. ISBN 1852338822.

RESOURCES

Computer material:
- Lecture slides. Slides.