Teaching Guide				
Identifying Data				2022/23
Subject (*)	Statistical Modeling of High Dimensional Data		Code	614G02013
Study programme	Grao en Ciencia e Enxeñaría de Datos			
Descriptors				
Cycle	Period	Year	Type	Credits
Graduate	1st four-month period	Second	Obligatory	6
Language	SpanishGalicianEnglish			
Teaching method	Face-to-face			
Prerequisites				
Department	Matemáticas			
Coordinador	Cao Abad, Ricardo	E-mail	ricardo.cao@udc.es	
Lecturers	Cao Abad, Ricardo López Cheda, Ana	E-mail	ricardo.cao@udc.es ana.lopez.cheda@udc.es	
Web	http://dm.udc.es/staff/ricardo_cao/			
General description	This course provides a first contact with the statistical modelization of high dimensional data: multivariate analysis methods, statistical tools and computer programs for the analysis of high dimensional data, identification of the advantages and disadvantages of the different methods, and critical procedures and interpretation of the results related to the proposed problem.			

Study programme competences / results	
Code	Study programme competences / results
A17	CE17 - Capacidade para a construción, validación e aplicación dun modelo estocástico dun sistema real a partir dos datos observados e a análise crítica dos resultados obtidos.
A20	CE20 - Coñecemento das ferramentas informáticas no campo da análise dos datos e modelización estatística, e capacidade para seleccionar as máis adecuadas para a resolución de problemas.
B2	CB2 - Que os estudantes saiban aplicar os seus coñecementos ao seu traballo ou vocación dunha forma profesional e posúan as competencias que adoitan demostrarse por medio da elaboración e defensa de argumentos e a resolución de problemas dentro da súa área de estudo
B3	CB3 - Que os estudantes teñan a capacidade de reunir e interpretar datos relevantes (normalmente dentro da súa área de estudo) para emitir xuízos que inclúan unha reflexión sobre temas relevantes de índole social, científica ou ética
B7	CG2 - Elaborar adecuadamente e con certa orixinalidade composicións escritas ou argumentos motivados, redactar plans, proxectos de traballo, artigos científicos e formular hipóteses razoables.
B8	CG3 - Ser capaz de manter e estender formulacións teóricas fundadas para permitir a introdución e explotación de tecnoloxías novas e avanzadas no campo.
B9	CG4 - Capacidade para abordar con éxito todas as etapas dun proxecto de datos: exploración previa dos datos, preprocesado, análise, visualización e comunicación de resultados.
B10	CG5 - Ser capaz de traballar en equipo, especialmente de carácter multidisciplinar, e ser hábiles na xestión do tempo, persoas e toma de decisións.
C1	CT1 - Utilizar as ferramentas básicas das tecnoloxías da información e as comunicacións (TIC) necesarias para o exercicio da súa profesión e para a aprendizaxe ao longo da súa vida.

Learning outcomes			
Learning outcomes	Study programme competences / results		
Knowledge related to the main techniques of multivariate statistical analysis.	A17	B2 B8 B9 B10	C1

Awareness of the main problems when working with high dimensional data.	A17 A20	B2 B3 B9 B10	C1
Ability to select the main variables and models in real data problems.	A17 A20	B2 B3 B8 B9	C1
Ability to apply the main techniques of multivariate analysis to real or simulated datasets.	A17 A20	B2 B3 B7 B8 B9 B10	C1
Ability to interpret the results and awareness of the limitations of the different models related to multivariate statistical analysis.	A17 A20	B2 B3 B7 B8 B9 B10	C1
Ability to work with advanced software related to statistical analysis.	A20	B2 B10	C1

Contents	
Topic	**Sub-topic**
0. Multidimensional distributions	0.1 Concept of multidimensional distribution 0.2. Variance-covariance matrix. Linear transformations. 0.3. Multidimensional normal: definition and properties.
1. Dimension reduction methods	1.1 Objectives of the Principal Component Analysis (PCA) 1.2 Transformations to get incorrelation 1.3 Obtaining the principal components 1.4 Principal components and scale changes 1.5 Interpretation of the principal components 1.6 Factor analysis 1.7 Multidimensional scaling
2. Unsupervised classification	2.1 Objectives of unsupervised classification: hierarchical and non-hierarchical methods 2.2 Cluster analysis: approach and objectives 2.3 Hierarchical tree or dendogram 2.4 Similarities and discrepancies between observations 2.5 Criteria for group formation: simple, complete chaining, group average, centroid method, Ward method 2.6 Non-hierarchical distance-based methods: closest neighbors, k means, methods based on density estimation

3. Supervised classification	3.1 Objectives of supervised classification: classification rules and error criteria 3.2 Discriminant factor analysis: approach, objectives and calculation of discriminant factors 3.3 Fisher's linear discriminant analysis and quadratic discriminant analysis 3.4 Maximum likelihood discriminant rule, Bayes rule, nonparametric discriminant rules 3.5 Relationship with regression models with binary response 3.6 Estimation of Probability of Incorrect Classification: Cross Validation and Bootstrap
4. Models for high-dimensional data	4.1 Variable selection in regression: significance tests. 4.2 The problem of multiple contrasts: false discovery rate (FDR) and familywise error rate (FWER) 4.3 Sparse coefficient regression models: ridge regression, lasso and their variants 4.4 Selection of variables and models with sparse coefficients for classification

Planning				
Methodologies / tests	Competencies / Results	Teaching hours (in-person & virtual)	Student?s personal work hours	Total hours
Oral presentation	A1 B2 B3 B4 C4	30	36	66
ICT practicals	A9 A12 A17 A18 A19 A20 A26 A33 A3 A4 A5 A6 A8 B7 B9 B10 C1 C2 C3	14	21	35
Multiple-choice questions	A19 A24 A25 A1 B3 B8	2	6	8
Problem solving	A17 A33 A2 B2 B5 B6 B7 B8 B10	14	21	35
Personalized attention		6	0	6
(*)The information in the planning table is for guidance only and does not take into account the heterogeneity of the students.				

Methodologies	
Methodologies	Description
Oral presentation	Presentation using the computer.
ICT practicals	Statistical data analysis using R.
Multiple-choice questions	Multiple-choice test on concepts.
Problem solving	Deciding statistical tools and strategies for problem solving. Model formulation for multivariate data. Algorithm formulation for high dimensional data analysis.

Personalized attention	
Methodologies	Description
ICT practicals Problem solving	Attendance and participation in lectures. Written multiple choice test. Multivariate data analysis project. Practicals to be performed by the student.

Assessment			
Methodologies	Competencies / Results	Description	Qualification

Oral presentation	A1 B2 B3 B4 C4	Oral presentation of the work (in pairs of students) mentioned in the "Problem solving" item.	10
ICT practicals	A9 A12 A17 A18 A19 A20 A26 A33 A3 A4 A5 A6 A8 B7 B9 B10 C1 C2 C3	Computer labs using the open statistical software R.	40
Problem solving	A17 A33 A2 B2 B5 B6 B7 B8 B10	Work (in pairs of students).	10
Multiple-choice questions	A19 A24 A25 A1 B3 B8	Comprehension tests.	40

Assessment comments

The assessment will be carried out through two tests on R labs, a work in pairs of students, as well as two written concept tests. The first of the tests on R labs and the first of the concept tests will be held approximately in the middle of the semester, and they will be related to topics 0-2. The corresponding second tests of both R labs and concepts will be held in the official exam date, in January 2023. These second tests will include all the topics given in the course, but those students who attended each of the first tests will be allowed to avoid topics 0-2 so that they will be evaluated only in topics 3-4.

Both the qualification of the R labs tests and the qualification of the concept tests will be 40% of the total qualification each, while the remaining 20% will correspond to the pairs student work, that has to be presented orally, in the second half of the semester. The half of the score of this pairs work (10% of the total score) corresponds to its oral presentation.

In summary, the assessment scoring will be the following:

Work in pairs: 20% of the total score (10% solving the practical exercise in R and 10% oral presentation).Concept tests: two concept tests will be carried out (each one of them corresponds to the 20% of the total score). The first of the exams, related to Topics 0-2, will take place in the middle of the quadrimester. The second exam, related to Topics 3-4, will be carried out the day of the official exam. Note that students will avoid Topics 0-2 in the second exam if they pass the first exam, unless they pass it but want to get a higher score. Moreover, if students do not take the first exam, they will be evaluated (the day of the official exam) with two parts: one related to Topics 0-2 and another related to Topics 3-4. The sum of these two parts is the 40% of the total score.Tests on R labs: following the same idea as for the concept tests, there will be two tests on R labs (each one of them corresponds to the 20% of the total score), using statistical software, R. The first exam, related to the R labs of Topics 0-2, will take place in the middle of the quadrimester. The second exam, related to the R labs of Topics 3-4, will be carried out the day of the official exam. Note that students will avoid Topics 0-2 in the second test on R labs if they pass the first exam, unless they pass it but want to get a higher score. Moreover, if students do not take the first exam, they will be evaluated (the day of the official exam) with two parts: one related to Topics 0-2 and another related to Topics 3-4. The sum of these two parts is the 40% of the total score.To pass the subject is necessary to obtain a score of at least 5 out of 10 overall.

On July opportunity, students could avoid those tests with scores of at least 4 out of 10 in January tests. In this respect, at least 4 out of 10 is required in partial tests (T0-T2 or T3-T4) or in the full test

(T0-T4), either in the concepts part or in the R labs part, to be kept for computing the final grade. Only students that didn't take any test will be qualified as NON ATTENDANT in the first opportunity (January-February). In July (2nd opportunity) only students that didn't take the final exam will be qualified as NON ATTENDANT.

If a student wants to take a test in a specific official language (Spanish or Galician), he/she must inform the professor at least 1 week in advance.

Sources of information

Basic	- Anderson, T.W. (2003). An Introduction to Multivariate Statistical Analysis. Wiley - Chatfield, C., Collins, A. J. (1980). Introduction to multivariate analysis. Chapman & Hall - Giraud, C. (2014). Introduction to High-Dimensional Statistics. Chapman & Hall/CRC - Goldstein, M., Dillon,W. R. (1984). Multivariate Analysis: Methods and Applications. Wiley - Jambu, M. (1991). Exploratory and Multivariate Data Analysis. Boston, Academic Press - Jobson, J.D. (1994). Applied Multivariate Data Analysis. Springer-Verlag - Johnson, R. A., Wichern, D. W. (2007). Applied multivariate statistical analysis. Prentice Hall - Koch, I. (2014). Analysis of Multivariate and High-Dimensional Data. Cambridge University Press - Mardia, K.V., Kent, J.T., Bibby, J.M. (1994). Multivariate Analysis. Academic Press. Academic Press - Muirhead, R.J. (1982). Aspects of multivariate statistical theory. John Wiley & Sons - Rencher, A.C. (1998). Multivariate Statistical Inference and Applications. Wiley - Wainwright, M.J. (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press
Complementary	

Recommendations
Subjects that it is recommended to have taken before
Introduction to Databases/614G02008
Linear Algebra/614G02001
Multivariable Calculus /614G02006
Discrete Mathematics/614G02002
Fundamentals of Programming II/614G02009
Fundamentals of Programming I/614G02004
Statistical Inference/614G02007
Probability and Basic Statistics/614G02003
Subjects that are recommended to be taken simultaneously
Regression Models/614G02012
Subjects that continue the syllabus
Simulation and Resampling Techniques/614G02036
Statistical Analysis of Complex Data/614G02031
Machine Learning III/614G02026
Information Retrieval/614G02027
Machine Learning I/614G02019
Machine Learning II/614G02021
Statistical Analysis of Dependent Data/614G02022
Other comments

(*)The teaching guide is the document in which the URV publishes the information about all its courses. It is a public document and cannot be modified. Only in exceptional cases can it be revised by the competent agent or duly revised so that it is in line with current legislation.