GHENT UNIVERSITY

Course Specifications

Valid as from the academic year 2023-2024

Analysis of High Dimensional Data (C003549)

Course size *(nominal values; actual values may depend on programme)*

Credits 5.0 **Study time 150 h**

Course offerings and teaching methods in academic year 2023-2024

A (semester 1)	English	Gent	group work
			lecture
			seminar
			independent work

Lecturers in academic year 2023-2024

Clement, Lieven	WE02	lecturer-in-charge

Offered in the following programmes in 2023-2024	**crdts**	**offering**
Master of Science in Statistical Data Analysis	5	A

Teaching languages

English

Keywords

Statistics, multivariate data analysis, high dimensional data

Position of the course

Modern high throughput technologies easily generate data on thousands of variables; e.g. genomics, chemometrics, environmental monitoring, ... Conventional statistical methods are no longer suited for effectively analysing such high-dimensional data. Multivariate statistical methods may be used, but for often the dimensionality of the data set is much larger than the number of (biological) samples. Modern advances in statistical data analyses allow for the appropriate analysis of such data.
Methods for the analysis of high dimensional data rely heavilly on multivariate statistical methods. Therefore a large part of the course content is devoted to multivariate methods, but with a focus on high dimensional settings and issues.
Multivariate statistical analysis covers many methods. In this course, only a few are discussed. A selection of techniques is made based on our experience that they are frequently used in industry and research institutes (e.g. principal component analysis, cluster analysis, classification methods). Central in the course are applications from different fields (analytical chemistry, ecology, biotechnology, genomics, ...).

Contents

1. Dimension reduction: Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Multidimensional Scaling (MDS) and biplots for dimension-reduced data visualisation
2. Sparse SVD and sparse PCA
3. Prediction with high dimensional predictors: principal component regression; ridge, lasso and elastic net penalised regression methods
4. Classification (prediction of class membership): (penalised) logistic regression, linear and quadratic discriminant analysis, (sparse) Fisher discriminant analysis
5. Evaluation of prediction models: sensitivity, specificity, ROC curves, mean squared error, cross validation
6. Feature selection
7. assocation versus prediction
8. Large scale hypoptheses testing: FDR, FDR control methods, empirical Bayes (local) FDR control
9. One of the following topics (depending on the interest of the students): functional data analysis, cannonical correlation analysis, correspondence analysis, biclustering, factor

analysis, model based clustering, ...

Initial competences

Having successfully completed the course “analysis of continuous data”, or having acquired otherwise the corresponding competences (knowledge of the theory and practice of linear statistical models). A good knowledge in matrix algebra is also required.

Final competences

1 The student has knowledge of methods for analysing and exploring high-dimensional data sets.
2 The student can see and quantify structures in large high dimensional/multivariate datasets, using the software R.
3 The student can value and interpret the statistical data analyses of high-dimensional data correctly.
4 The student can correctly report the results of the data analyses according to scientific standards.
5 The student can comprehensively read scientific papers related to the course content.
6 The student can take responsibility and initiative in a group effort.

Conditions for credit contract

Access to this course unit via a credit contract is determined after successful competences assessment

Conditions for exam contract

This course unit cannot be taken via an exam contract

Teaching methods

Group work, Seminar, Lecture, Independent work

Extra information on the teaching methods

Theory: 15 hours lectures and 7,5 hours guided self study (discussion of papers)
Exercises: 15 hours PC room classes and 25 hours group or individual work (assignment)

Learning materials and price

Slides and all material for the PC labs are available from Ufora. A syllabus is available for approx. 10 EUR. The pdf file is also available from Ufora.

References

Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction. IMS Monographs.
Efron, B., and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.
Hastie T., Tibshirani R. and Friedman J. (2009). The Elements of Statistical Learning. Springer.
Johnson R. and Wichern D. (2008). Applied Multivariate Statistical Analysis (6th edition). Prentice Hall.
Ramsay J. and Silverman B. (2005). Functional Data-analysis (second edition). Springer-Verlag.
Ramsay J. and Silverman B. (2002). Applied Functional Data-analysis: Methods and Case Studies. Springer-Verlag.

Course content-related study coaching

In the practical sessions in the PC labs the students are coached by an assistant.
Students can make an appointment to ask questions to the lecturer. Questions and answers can be exchanged in Ufora.

Assessment moments

end-of-term and continuous assessment

Examination methods in case of periodic assessment during the first examination period

Written assessment with open-ended questions

Examination methods in case of periodic assessment during the second examination period

Written assessment with open-ended questions

Examination methods in case of permanent assessment

Peer and/or self assessment, Assignment

Possibilities of retake in case of permanent assessment

examination during the second examination period is possible in modified form

Extra information on the examination methods

Theory and exercises: periodical evaluation (open book, written exam with open-ended questions) and non-periodical evaluation (2 homework assignments and 1 project assignment).

Calculation of the examination mark

Theory and exercises: periodical evaluation (50%) and non-periodical evaluation (1 homework assignment1 (12.5%) and 1 project assignment (37.5%)).
To pass for this course, the student must pass for both the periodical and the non-periodical evaluation.
If the student fails for this course in the first examination period and if he/she wants a retake in the second examination period, the non-periodical evaluation will be presented in a revised form in the second examination period.
Peer assessment is used to correct the marks for the project assignment so as to provide a more representative score for the student's individual contribution. A student's individual score for the project work will at most deviate 20% of average score for the group. 20% of the score of the project will be based on a report that is used to evaluate the process that student went through to arrive at the result (focus on the skills to participate in a group effort).