Analysis of High Dimensional Data (I002004)

Course Specifications
Valid as from the academic year 2017-2018

Course size (nominal values; actual values may depend on programme)
- Credits 3.0
- Study time 75 h
- Contact hrs 30.0 h

Course offerings and teaching methods in academic year 2018-2019
- A (semester 2)
  - English
  - 15.0 h seminar: practical PC room classes
  - 15.0 h lecture

Lecturers in academic year 2018-2019
- Thas, Olivier
- LA26 lecturer-in-charge

Offered in the following programmes in 2018-2019

- Master of Science in Bioinformatics (main subject Bioscience Engineering) 3 A
- Master of Science in Chemical Engineering 3 A
- Master of Science in Chemical Engineering 3 A
- Master of Science in Bioscience Engineering: Land and Water Management 3 A
- Master of Science in Bioscience Engineering: Environmental Technology 3 A

Teaching languages
- English

Keywords
- Statistics, multivariate data analysis, data mining, data science

Position of the course

Multivariate and high-dimensional data analysis is basically a collection of many statistical methods that are applicable to large and/or high dimensional data sets. All methods that are covered in this course, are often applied in industry and research institutions. Good knowledge of basic statistical methods and linear regression models are required, as well as notions of matrix algebra (matrix multiplications, matrix inverse, inner product).

Applications are important in this course, and they are related to several different subject fields (e.g. analytical chemistry, ecology, environmental sciences, biotechnology, ...).

Reporting of the results of a statistical analysis is considered to be very important.

Contents

1. Association versus prediction
2. The singular value decomposition (SVD), multidimensional scaling (MDS), principal component analysis (PCA) and the biplot
3. Sparse SVD, sparse PCA
4. High-dimensional linear prediction models (including classification): principal component regression, ridge, lasso and elastic net
5. Model building / feature selection for linear models in high-dimensional settings
6. (Sparse) Fisher's Discriminant analysis
7. Large scale hypothesis testing, the false discovery rate and empirical Bayes
8. One of the following topics (depending on the interest of the students): functional data analysis, canonical correlation analysis, correspondence analysis, analysis of variance, biclustering, factor analysis, cluster analysis, ...

Initial competences

A basic course in probability theory and statistics (linear models, i.e, regression

(Approved)
analysis and analysis of variance), and a good basic knowledge of matrix algebra.

Final competences
1. The student has knowledge of methods for analysing and exploring high-dimensional data sets.
2. The student can detect 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 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
Lecture, seminar: practical PC room classes.

Learning materials and price
A syllabus is available at ca. 10 €. All learning materials are made available through Minerva.

References

Course content-related study coaching
In the practical sessions in the PC classes 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 Minerva.

Evaluation methods
end-of-term evaluation and continuous assessment

Examination methods in case of periodic evaluation during the first examination period
Written examination with open questions, open book examination.

Examination methods in case of periodic evaluation during the second examination period
Written examination with open questions, open book examination.

Examination methods in case of permanent evaluation
Assignment, peer assessment.

Possibilities of retake in case of permanent evaluation
Examination during the second examination period is possible.

Extra information on the examination methods
The final examination consists of a written examination with open questions (part if it is an open book examination). Most of the questions aim to assess the insight of the student in the statistical analysis and how they result in the conclusions. This happens at the borderline between theory and application.
The assignment relates to a real life problem. Students may work together in groups of 3 to 4 persons. The result of the project work is a written report that should satisfy conventional scientific and professional standards.

Calculation of the examination mark
The total mark is a weighted average of:
- Final exam (12/20)
- Assignment (8/20)

(Approved)