## **Applied Multivariate Statistics**

Fall-Winter-Semester 2021

University of Mannheim Department of Economics Chair of Statistics Dr. Toni Stocker

### **Applied Multivariate Statistics (AMS) - Content**

Introduction to AMS	1
Matrix Algebra	20
Multivariate Samples	57
Principal Component Analysis (PCA)	77
Biplots	131
Factor Analysis	143
Multidimensional Scaling (MDS)	156
Cluster Analysis	174
Linear Discriminant Analysis (LDA)	187
Binary Response Models	198
Statistical Methods for Data Science	216

# **Introduction to AMS**

### **General Course Information**

#### Prerequisites

- Students in Economics from Mannheim: no problem
- All other students: should have attended two or more courses in Statistics (descriptive statistics, estimating and hypothesis testing)
- A course in *Basic Econometrics* or *Linear Algebra* would be helpful but ist not strictly required.
- The statistical software R will <u>intensively</u> be used throughout this course. Students who are not yet familiar with R should work through chapters 1-5 of the *R introduction* (see course folder) on their own by September 16 at the latest.
- Though R is easy to learn, you need to invest some time at the beginning. But you may benefit from it for a long time.

### **General Course Information**

### **Time and Locations**

Format	Day	Time	Method
Lecture	provided till	Friday noon	Video (not live)
Group Exercise Session 1	Thursday	17:15-18:45	Online, Live (Zoom)
Group Exercise Session 2	Friday	08:30-10:00	Online, Live (Zoom)
Homework Support (optional)	Wednesday	15:00-16:30	Online, Live (Zoom)

Choose one of the two compulsory GE-Sessions! They start in the 2nd week. The Homework Support is during my office hour, but no appointment is needed. It is purely optional and starts in the 2nd week.

#### Contact

Office Hour:	Wednesday,	3:00-4:30 p.m.	only online	(Zoom)
--------------	------------	----------------	-------------	--------

Office: L7, 3-5, 1st floor, 143

Phone: 0621-181-3963

Email: stocker@uni-mannheim.de

### **General Course Information**

### **Course Material**

Slides (Lecture), Assignments (Tutorials), 'Introduction to R' (see p. 2)

Material will be updated weekly (Friday) to find in course folder at **Studierendenportal (ILIAS)** 

#### References

- R. Johnson, D. Wichern (2007): Applied Multivariate Statistical Analysis; Pearson Education 6th ed.
- A. C. Rencher, W. F. Christensen (2012): Methods of Multivariate Analysis; Wiley 3rd ed.
- W. Härdle, L. Simar (2003): Applied Multivariate Statistical Analysis; Wiley.
- A. J. Izenman (2008): Modern Multivariate Statistical Techniques; Springer.
- J. H. Stock and M. W. Watson (2020): Introduction to Econometrics; Pearson 4th ed.

Main Reference

### Examination

#### **Exam + Assignments:**

80% written exam (120 minutes) + 20% Exercises (individual and collaborative part) in terms of points to earn in total.

### **Example:**

	Points
Written Exam:	60 (from 80)
Exercises:	<u>18 (from 20)</u> :
Total:	78 (from 100)

```
=> Grading will be based on 78 points (from 100)
```

```
Minimum for passing: \leq 40
```

#### **Assignments:**

Need to submit homework and attend tutorial. To get full points (20) you need to work at least on 10 assignments (out of 13) in a meaningful way. (See *Guidelines for Assignments*)

### **Issues of Applied Multivariate Statistics (AMS)**

Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. See Renchner (2002), p.1

#### **Objectives**

- Dimension reduction and structural simplification
- Visualization of high-dimensional data
- Investigation of the dependence among variables
- Grouping, discrimination and classification
- Close link to many methods and techniques frequently used in areas like Exploratory Data Analysis (EDA), Data Mining, Statistical Learning or – more generally – what is called or described today as "Data Science"

(see also J+W (2007), p.2 or Izenman (2008), chap. 1

#### **Example 1:** Dimension Reduction

Economic Indicators for the 27 European Union Countries in 2011 (see *WIREs Comput Stat 2012, 4:399–406. doi: 10.1002/wics.1200*)



#### **Example 2:** Brushing and Linking

#### ... using R (iplots)



#### Example 2 ...



#### Example 2 ...

### ... or using Ggobi





#### **Example 3:** Factor Analysis

Consumer Preference (J&W, example 9.9, p. 508)



#### **Example 4:** Distances

Voting results for 15 congressmen from New Jersey (example from R package HSAUR)



Extraction from the distance matrix ...

	Hunt(R) Sand	dman(R) Hov	vard(D)
Hunt(R)	0	8	15
Sandman(R)	8	0	17
Howard(D)	15	17	0

### Example 5: Grouping



**Cluster Dendrogram** 

#### **Example 6:** Dimension Reduction in a Time Series Context

Monthly Returns of 10 subindices of the Euro Stoxx Price Index



What's going on here? Any idea how to proceed?

#### **Example 7:** Classification

#### Labor Market Participation of Married Women in Switzerland (1981) (example from R package AER)



15

### **Example 8:** Discrimination

Heights and weights of students



**16** 

### **Course Outline**

Roughly: Chapters 1-4, 8, 9, 11, 12 from J&W + additional stuff

#### **Timetable and Contents**

- Lecture 1: Introduction
- Lecture 2: Matrix Algebra (part 1)
- Lecture 3: Matrix Algebra (part 2)
- Lecture 4: Multivariate Samples
- Lecture 5: Principal Component Analysis (part 1)
- Lecture 6: Principal Component Analysis (part 2)
- Lecture 7: Biplots
- Lecture 8: Factor Analysis

- Lecture 9 (new!): PCA and FA: Advanced Applications
- Lecture 10: Multidimensional Scaling
- Lecture 11: Cluster Analysis
- Lecture 12: Linear Discriminant Analysis
- Lecture 13: Binary Response Models
- Lecture 14: Statistical Methods for Data Science

Note: This is just a plan! Topics may be skipped; order may be changed; lecture topics may overlap

### Main Objectives

... at the end of the semester you

- know and (hopefully) understand most common methods for analyzing multivariate data and their theoretical background
- can proficiently use R when using multivariate techniques: data import, constructing graphics, inference, model diagnosis and assessment
- have experienced the possibilities and limitations of multivariate methods on the basis of real data examples

т.

Generally: This is an introductory and applied course. Modern multivariate techniques based on machine learning algorithms will hardly be covered.