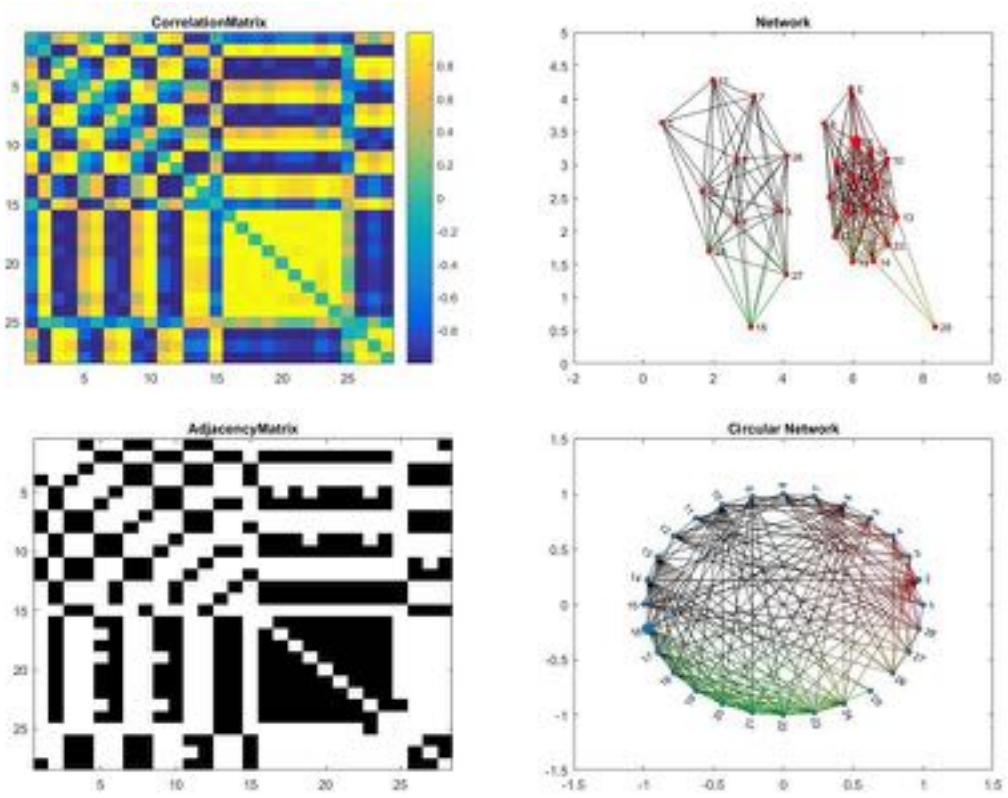
Smart Data Analytics

Wolfgang Karl Härdle

Cathy Yi-Hsuan Chen



Wolfgang Karl Härdle

Wolfgang Karl Härdle completed his Dr. rer. nat. in Mathematics at Heidelberg University and received his habilitation in Economics at Friedrich Wilhelm Universität Bonn.

He was the founder and Director of Collaborative Research Center CRC 373 "Quantification and Simulation of Economic Processes" (1994 - 2003) and also the Director of C.A.S.E. (Center for Applied Statistics and Economics) (2001 - 2014). He founded the CRC 649 "Economic Risk" (2005 - 2016) and is now directing the Sino-German International Research Training Group IRTG1792 "High dimensional non stationary time series analysis" (2013-2022). He has been teaching Master courses at Ladislaus von Bortkiewicz Chair of Statistics at Humboldt-Universität zu Berlin for more than twenty years.

His research focuses on dimension reduction techniques, computational statistics and quantitative finance. He has published many books and more than 300 papers in top statistical, econometrics and finance journals and has a high citation indices across different ranking platforms, see hu.berlin/wkh

He is among the top 1% of economists registered at REPEC and has similar top notch rankings in other scales, such as the Handelsblatt ranking, Scopus and Google scholar.

His professional experience includes financial engineering, analysis of unstructured data and dynamic decision analytics. He currently focuses his research on sentiment distillation, crypto currencies and DEDA - Digital Economy & Decision Analytics. He has supervised more than 50 PhD students and is holding up long-term research relations to partners in the USA, Singapore, Prague, Warsaw, Paris, Cambridge, Beijing, Xiamen and Taipei among others.

Cathy Yi-Hsuan Chen

Cathy Yi-Hsuan Chen received her PhD diploma in Finance at National Cheng-Chi University. She is since 2015.8 Associate Professor at the School of Business & Economics in Humboldt-Universität zu Berlin, and a principal investigator of the International Research Training Group 1792 – High Dimensional Non Stationary Time Series. She is the member of Princeton-Humboldt Cooperation and Collective Cognition Network (CoCCoN), and visiting fellow of Sim Kee Boon Institute for Financial Economics, Singapore Management University.

She has been teaching Master courses at Ladislaus von Bortkiewicz Chair of Statistics at Humboldt-Universität zu Berlin since 2015. The courses she has taught include *Statistics of Financials markets*, *Quantitative Finance*, *Economic Risk Seminar* and *Q-Kolleg* (a joint course collaborated with the National University of Singapore).

She had a five-years of consulting experience in the insurance industry. Through internal rating-based models, developed and customised programming training (Matlab), a company she consulted can benefit from superior risk management, employee training and fulfill all the obligations and regulations required by the authorities. She is currently heading a "transfer project" between Humboldt-Universität and Deutsche Bank, and focussing on credit risk modeling and stress testing.

Her research focuses on quantitative finance, risk modeling and management, behavior finance and text mining in finance. She has published the papers in reputable journals in the past 10 years. She currently concentrates her research on text extraction, text analytics, textual sentiment distillation, textual sentiment modeling, lexicon construction, investor decision and sentiment.

Course Contents

SDA Learning Objectives

The SDA course presents tools and concepts for unstructured data with a strong focus on applications and implementations. It presents the decision analytics in a way that is understandable for non-mathematicians and practitioners who are confronted with day to day number crunching statistical data analysis. All practical examples may be recalculated and modified: software and Quantlets are in www.quantlet.de. The SDA course endows the practitioner with ready to use practical tools for smart data analytics.

SDA Structure

Data are everywhere and the ubiquitous availability of huge amounts of data makes it necessary to develop smart data analytics. Out of the plethora of tools that are available for many scientific disciplines this course offers for the common data analyst an easy access to all levels of analysis without deep computer programming knowledge. SDA provides a wide variety of exercises. In addition a full set of slides is provided making it easier for the participants to reanalyze the presented material. The R and Python programming language are becoming the lingua franca of computational data analysis. They are the common smart data analysis software platforms used inside corporations and in academia. Both are OS independent free open-source programs which are popularized and improved by hundreds of volunteers all over the world.

SDA Literature

Franke J, Härdle WK, Hafner C (2015) Statistics of Financial Markets: an Introduction. 4th ed., Springer Verlag, Heidelberg. ISBN: 978-3-642-54538-2

Chen C YH, Härdle WK, Overbeck L (2017) Applied Quantitative Finance. 3rd extended ed., Springer Verlag, Heidelberg.

Härdle WK, Simar L (2015) Applied Multivariate Statistical Analysis. 4th ed., Springer Verlag, Heidelberg. ISBN 978-3-662-45170-0

Härdle WK, Okhrin O, Okhrin Y (2017) Basics of Computational Statistics, Springer Verlag, Heidelberg.

All examples are presented in R or Python. The Quantlets are available here:

www.quantlet.de

Schedule

Unit 1	
What do we see?	• Basic concepts • Data Management • Structuring Data elements
Unit 2	
Data Analysis	• Sentiment extraction • Stemming, lemmatizing • DTM Dynamic Topic Modeling
Unit 3	
Modern Data Analytics	• Cluster Analysis and Classification • TEDAS Tail Event Driven Asset Allocation • CRIX a CRypto currency IndeX
Unit 4	
Smart Data Analytics	• R tools for text mining • text mining in Quantitative Finance • Applications & Empirics
Unit 5	
Smart Data Analytics	• Network Geometry • Tail Event driven Network AutoRegression • DYTEC DYnamic Tail Event Curves
Unit 6	
Smart Data Analytics	• Financial Risk Meter • DDI Networks Topology • Q3 D3 LSA

Contact

Wolfgang Karl Härdle
Ladislaus von Bortkiewicz Chair of Statistics
School of Business and Economics
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin, Germany

Telefone +49 30 2093 5631

FAX +49 30 2093 5649

E-Mail stat@wiwi.hu-berlin.de

Links

hu.berlin/93629

hu.berlin/irtg1792

http://crix.hu-berlin.de

hu.berlin/**rdc**

hu.berlin/frm

http://quantlet.de

Appendix

Details on course flow and homework (HW)

Unit 1 (3h)

Class room presentations by Professor:

FMRI Data analysis and Neuro Economic RPID tasks risk perception
BCS Introduction into Basics of Computational Statistics, slides 1-1-1-20

HW Unit 1:

1. Calculate the increase of memory of PCs over the last 30 years and check whether the FMRI analysis could have been done 20 years ago
2. prepare 2-5 slides explaining logistic regression
3. install R and run simple programs from Quantlet.de, make sure you have a Github (GH) account.

Unit 2 (3h)

HW presentations
Working with quantlet.de
BCS Introduction into Basics of Computational Statistics, slides 1-21-1-52, slides 2-31, 2-51 on BCS Qs from GH BCS_integrand, BCS_newton
(100 min)
Binomial distributions, Plotting, varying n,p hypergeometric distribution, Poisson distribution, how many errors does a typist do?
How to scrape data from NASDAQ web page
Blockchain Introduction
(80min)

HW Unit 2

1. make an R quantlet to solve HW #1 from unit 1 with R and show it on Github (GH) hint: use the CMB Qs for this work
2. use R with B-spline code to solve HW#1, any comments?
3. Suppose you observe that in n=1000 mails (in 1 week) you have about 2 scams. Use the LvB /Poisson cdf to calculate that you have 6 scam emails in 2 weeks. In Scammyland you have 5 scams on average, what is the probability to have no scam mail.

Unit 3 (3h)

HW presentations (30min)
Blockchain Introduction hash codes (40min)
Presentation of SCRY.INFO, a Chengdu based company that offers blockchain services (40 min)
CRIX a CRyptocurrency IndeX (70min)
Reading CRIX data from JSON file on CRIX web page

HW Unit 3

1. make an R quantlet on GH to produce hash code for the 2 sentences: „I learn a lot from this class when I am proper listening to the professor", „I do not learn a lot from this class when I am absent and playing on my Iphone". Compare the 2 hash sequences
2. Make 3-5 slides (in PPTX) on the DSA (Digital Signature Algorithms)
3. Make slides with R code where you create a JSON data set that you save and read again.
4. Download the CRIX data and make a plot of the time series, analyse its properties, i.e. fit ARMA, ARIMA etc. Is there a GARCH effect?

Unit 4 (3h)

HW presentations (30min)
The econometrics of CRIX (90min)
Web page extraction (60 min)
Sorry to mention that the Teacher computer could in the beginning of class not connect to internet. So the students had to bring their own computer and show it from their GH account

HW Unit 4

(all this to be done on perfect PPTX slides)

1. improve the R quantlets on GH (from CRIX directory on quantlet.de) and make excellent graphics that follow Fig 3,4,5,6 of the „Econometrics of CRIX" paper.
2. make your R code perfect as in the R examples on quantlet.de i.e. make sure that the code is „time independent" by using actual dimensions of the data that you are collecting from crix.hu-berlin.de RecreateFig 7 from „Econometrics of CRIX".
3. redo as many figures as you can.

Unit 5 (3h)

HW presentations (30min)

Final EXAM

Collect all HW s in one word file and leave it on GH for evaluation.