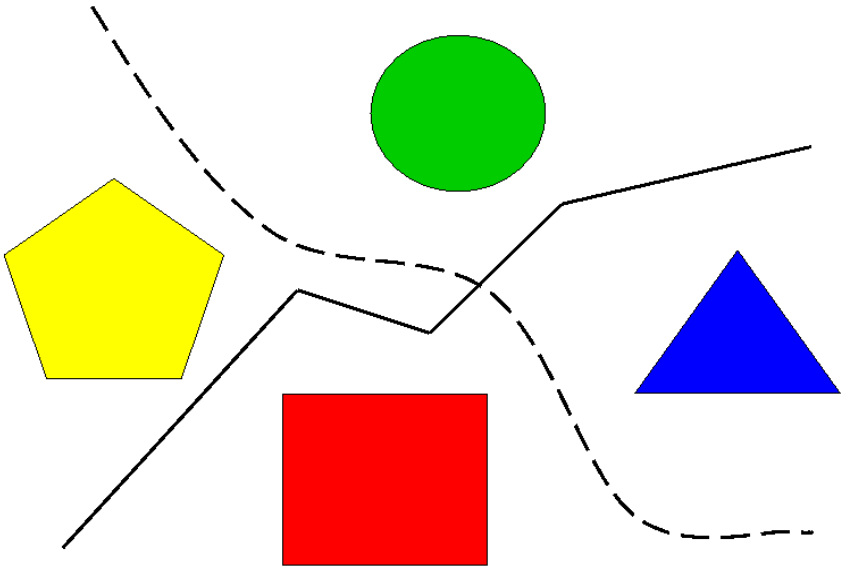
technische universität
dortmund

Fakultät Statistik

7. Herbstkolloquium
des Graduiertenkollegs
"Statistische Modellbildung"

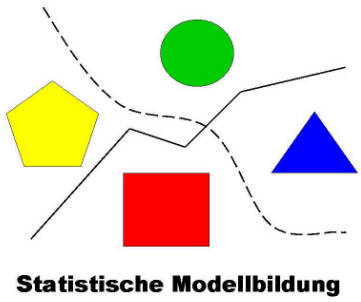


Statistische Modellbildung

Zu diesem Kolloquium wird eingeladen

Freitag / Samstag, 19./20. November 2010

UNIVERSITÄTSKOLLEG BOMMERHOLZ
- Lehr- und Weiterbildungsstätte der TU Dortmund -
Bommerholzer Straße 60, 58456 Witten.
(Tel.: ++49 (0)2302 / 39 60, Fax: ++49 (0)2302 / 39 63 20)



7. Herbstkolloquium des Graduiertenkollegs "Statistische Modellbildung"

Freitag, 19. November 2010

Abfahrt nach Witten: ab Dortmund gegen 14.00 Uhr

Vortragsprogramm I

15:15 **Begrüßung**
Prof. Dr. Joachim Kunert

15:30 **Dr. Tino Ullrich**
Hausdorff Center for Mathematics, Universität Bonn
Sparse signal recovery with random matrices

16:15 **Katrin Ullrich**
Fraunhofer IAIS, B-IT Research School, Universität Bonn
Kernel methods for ligand prediction

17:00 **Dr. Korbinian Strimmer**
Statistics and Computational Biology, Universität Leipzig
High-dimensional variable selection by decorrelation

Diskussion zu den Projektbereichen

19:00 **Posterausstellung:**
Präsentation der Dissertationsprojekte im Kolleg, Diskussion in Arbeitsgruppen

Samstag, 20. November 2010

Vortragsprogramm II

9:00	**Prof. Dr. Ying Chen** *Shanghai University of Finance and Economics, China*	**An analysis of 3-level orthogonal saturated designs**
9:45	**Dr. Alexandra Graf** *Section of Medical Statistics, Medizinische Universität Wien, Austria*	**Maximum inflation of the type 1 error rate when the sample size is adapted in a pre-planned interim look**
10:30	***Pause***	

Vortragsprogramm III

11:00	**Dr. Björn Schelter** *Freiburg Center for Data Analysis and Modeling, Universität Freiburg*	**Multivariate analysis of dynamical processes with applications in the neurosciences**
11:45	***Dr. Manfred Mudelsee*** *Climate Risk Analysis, Hannover*	**Climate time series analysis, with examples from regression and risk analysis**
12:30	***Pause***	
13:15	***Gruppenfoto auf der Terrasse (wetterabhängig)***	

Vortragsprogramm IV

13:30	**Prof. Dr. Friedrich Leisch** *Institut für Statistik, Universität München*	**(Title to be announced)**
14:15	**Roland Winkler** *Deutsches Zentrum für Luft- und Raumfahrt, Köln*	**Cluster analysis on thigh dimensional data**
15:00	.**Prof. Dr. Peter Müller** *Department of Biostatistics, The University of Texas, Houston, U.S.A.*	**Bayesian clustering with regression**
15:30	***Kaffeetrinken, Abschlussbesprechung und Diskussion***	

An analysis of 3-level orthogonal saturated designs

Ying Chen
Shanghai University of Finance and Economics, Shanghai, China
Chi Kin Chan and Bartholomew P. K. Leung
The Hong Kong Polytechnic University, Kowloon, Hong Kong

Although three-level factorial designs with quantitative factors are not the most efficient way to fit a second-order polynomial model, they often find some application, when the factors are qualitative. The three-level orthogonal designs with qualitative factors are frequently used, e.g., in agriculture, in clinical trials and in parameter designs. It is practically unavoidable that, because of the limitation of experimental materials or time-related constraint, we often have to keep the number of experiments as small as possible and to consider the effects of many factors and interactions simultaneously so that most of such designs are saturated or nearly saturated. An experimental design is said to be saturated, if all degrees of freedom are consumed by the estimation of parameters in modelling mean response. The difficulty of analyzing orthogonal saturated designs is that there are no degrees of freedom left to estimate the error variance so that the ordinary ANOVA is no longer available. In this paper, we present a new formal test, which is based on mean squares, for analyzing three-level orthogonal saturated designs. This proposed method is compared via simulation with several mean squares based methods published in the literature. The results show that the new method is more powerful in terms of empirical power of the test. Critical values used in the proposed procedure for some three-level saturated designs are tabulated. Industrial examples are also included for illustrations.

Maximum inflation of the type 1 error rate when the sample size is adapted in a pre-planned interim look

Alexandra Graf
Section of Medical Statistics, Medizinische Universität Wien, Austria

Sample size reassessment in an adaptive interim analysis based on an interim estimate of the effect size can considerably increase the type 1 error rate if the planned fixed sample size test is applied for the final analysis. Adaptive combination tests (see e.g. Bauer (1989), Bauer and Koehne (1994), Brannath et al. (2002)) and tests based on the conditional error function principle (see Proschan and Hunsberger (1995), Mueller and Schaefer (2001), Mueller and Schaefer (2004)) have been proposed which allow for such interim design modifications without compromising on the type 1 error rate. The maximum inflation of the type 1 error rate for such type of design can be calculated by searching for the “worst case” scenarios, i.e. sample size adaption rules in the interim analysis that lead to the largest type 1 error rate inflation. This has been investigated by Proschan and Hunsberger (1995) for the one-sided comparison of two normal means (variance known) with sample sized balanced between groups. They showed that the type 1 error rate may be inflated from 0.05 to 0.1146 if in an interim analysis the second stage sample size is chosen to maximize the conditional error function. However, when sample size and allocation rate to the treatment arms can be modified in an interim analysis the maximum inflation of the type 1 error rate is substantially larger than derived by Proschan and Hunsberger (1995).

Bauer, P (1989). Multistage testing with adaptive designs.
Biometrie und Informatik in Medizin und Biologie 20, 130-148.
Bauer, P. and Koehne, K. (1994). Evaluations of experiments with adaptive interim analysis.
Biometrics 50, 1029-1041.
Brannath, W., Posch, M. and Bauer, P. (2002). Recursive combination tests.
JASA 97, 236-244.
Mueller, H.H. and Schaefer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches.
Biometrics 95, 886-891.

Mueller, H.H. and Schaefer, H. (2004). A general statistical principle for changing a design any time during the course of a trial.
Statistics in Medicine 23, 2497-2508.
Proschan, M.A. and Hunsberger, S.A. (1995). Designed extension of studies based on conditional power.
Biometrics 51, 1315-1324.

New features for archetypal analysis in R

Friedrich Leisch
Institut für Statistik, Ludwig-Maximilians-Universität München

Archetypal analysis (Cutler & Breiman, 1994) has the aim to represent observations in a multivariate data set as convex combinations of a few, not necessarily observed, extremal points (archetypes). The archetypes themselves are restricted to being convex combinations of the individuals in the data set and lie on the boundary of the data set, which makes the analysis very sensitive to outliers. Although archetypal analysis is about the data set boundary, practice has shown that in many cases one primarily is interested in the archetypes of the large majority than of the totality.

R package archetypes (Eugster & Leisch, 2009) is a new freely available and very flexible implementation of the algorithm, that can easily be modified. We have adapted the original archetypes estimator to be a robust M-estimator and present an iteratively reweighted least squares fitting algorithm (Eugster & Leisch, 2010). Our robust archetypal analysis algorithm is based on weighting the residuals and observations, respectively. As a byproduct we hence obtain weighted archetypal analysis which enables us to represent additional information available from the data set, like the importance of observations or the correlation between observations.

Climate Time Series Analysis, with Examples from Regression and Risk Analysis

Manfred Mudelsee
Climate Risk Analysis, Hannover

Climate is a paradigm of a complex system. Analysing climate data is an exciting challenge, which is increased by non-normal distributional shape, serial dependence, uneven spacing and timescale uncertainties.

This talk follows a recent book by the author (sample at www.manfredmudelsee.com/book), which presents bootstrap resampling as a computing-intensive method able to meet the above mentioned challenge. We look on the bootstrap in following estimation techniques: regression and extreme values. We learn about climate and the need to do quantitative climatology by means of statistical methods.

Bayesian Clustering with Regression

Peter Müller
Department of Biostatistics, The University of Texas, Houston, USA

We discuss models for covariate-dependent clustering, i.e., probability models for random partitions that are indexed by covariates. The motivating application is inference for a clinical trial. As part of the desired inference we wish to define clusters of patients. Defining a prior probability model for cluster memberships should include a regression on patient baseline covariates. We review some current approaches, and propose a new model based on an extension of product partition models (PPM). We define an extension of the PPM to include the desired regression. This is achieved by including in the cohesion function a new factor that increases the probability of experimental units with similar covariates to be included in the same cluster.

Multivariate Analysis of Dynamical Processes with Applications in the Neurosciences

Björn Schelter
Freiburg Center for Data Analysis and Modeling, Universität Freiburg

Networks are ubiquitous in the neurosciences. To understand the characteristic behavior of dynamical networks, its topology and specific types of interaction between constituents of the network have to be inferred. To this end, recent methodological developments in the field of linear and non-linear approaches to analyze multivariate time series and point processes are addressed, with particular focus on various Granger causality measures to infer the direct directed interaction structure of the processes. Kalman filter approaches are discussed for processes, which are nonstationary and contaminated by observation noise. The abilities and limitations of the techniques are presented.

High-dimensional variable selection by decorrelation

Korbinian Strimmer (joint work with Verena Zuber)
Statistics and Computational Biology, Universität Leipzig

Variable selection problems are ubiquitous in genomic high-throughput data analysis, e.g., to identify genetic signatures for tumor classification or medical biomarkers relevant for disease prediction. Standard approaches based on t-scores and marginal correlation do not take account of the correlation structure among genes.
In my talk I discuss a model selection approach based on decorrelation, which leads to the CAT [1] and CAR [2] score variable ranking criteria.

We show that CAT and CAR scores are highly efficient for prediction and classification, show favorable theoretical properties, and are competitive with the best modern regression and classification approaches available.

[1] V. Zuber and K. Strimmer. 2009. Gene ranking and biomarker discovery under correlation. Bioinformatics 25: 2700-2707.
[2] Zuber, V., and K. Strimmer. 2010. Variable importance and model selection by decorrelation. http://arxiv.org/abs/1007.5516

Kernel Methods for Ligand Prediction

Katrin Ullrich
Fraunhofer IAIS, B-IT Research School, University of Bonn

With the increasing amount of available data, machine learning and so-called kernel methods have become more and more important in recent years. In many practical issues, one is interested in the hidden rule which generates complex data pairs. Therefore, usually one is looking for a function $f : X \to Y$ out of a hypothesis space H mapping from an input or instance space X into an output or label space Y. In many practical problem settings we have available a set of training data of known data pairs. In order to find a predictor function for unseen instances, we assume the “best” function f being the one which minimizes the expected loss EL(y,f(x)) for an appropriate loss-function L. As we do not know the distribution P on X x Y, but have the training data, we estimate the expected loss by the empirical loss. Finally, this leads to the minimization of the regularized risk functional . To find a solution f* one has to make assumptions on the hypothesis space H. For kernel methods, H is assumed to be a reproducing kernel Hilbert space of functions generated by a symmetric and non-negative definite kernel function $k : X \times X \to R$.

Ligand prediction is the search for small molecules binding to much larger proteins. As many biochemical reactions are triggered by cell proteins it is very interesting to find new ligands for proteins. I will present approaches for ligand prediction via kernel methods.
Ligand prediction is the search for small molecules binding to much larger proteins. As many biochemical reactions are triggered by cell proteins it is very interesting to find new ligands for proteins. One distinguishes between the case of either already known ligands for a target protein or proteins with no ligands known in advance (orphan targets). Well-performing algorithms for virtual screening of proteins can be beneficial for pharmaceutic industry as screening of proteins in a laboratory is very expensive and time-consuming. In my talk I will present approaches for ligand prediction via kernel methods.

Sparse Signal Recovery with Random Matrices

Tino Ullrich
Hausdorff Center for Mathematics, University of Bonn

In recent years sparsity has become an important concept in applied mathematics, especially in mathematical signal and image processing, in the numerical treatment of PDEs as well as inverse problems, and statistics. The key idea is that many types of functions and signals arising naturally in these contexts can be described by only a small number of significant degrees of freedom. The novel theory of Compressive Sensing heavily uses this model and predicts, quite surprisingly, that sparse high-dimensional signals can be recovered efficiently from what was previously considered highly incomplete measurements. This discovery has led to a fundamentally new approach towards certain signal and image recovery problems.
Interpreting the signal as a vector x of length N with small support set (non-zero entries) of size s we are able to recover this signal by asking a number m of questions which scales linearly in the sparsity s. To be more precise, these questions are realized by a matrix vector multiplication with a proper matrix A and y = Ax. Note, that the system y = Ax is highly under-determined. However, the signal x is recovered from y via a convex L1-norm-minimization program. The exact recovery is guaranteed if the measurement matrix A satisfies the Restricted Isometry Property (RIP). Remarkably, mainly random constructions for measurement matrices A reduce the effort (number of rows m) significantly. Using independent mean-zero sub-Gaussian entries one can prove that m = slog(N/s) rows suffice. The main tool are concentration inequalities for the size of the entries of A. Further constructions involve partial random circulant matrices or Fourier matrices with random samples and provide a similar behavior.
This talk is intended to give an introduction to compressive sensing mainly focused on the RIP of random matrices. If time permits we will also comment on lower bounds for optimal recovery.

Cluster analysis on high dimensional data

Roland Winkler
Deutsches Zentrum für Luft- und Raumfahrt, Köln

A data set may contain of one or more 'clouds' of data objects. The task for cluster analysis is, to find the location of these clouds and to provide a partitioning of the data objects. Many algorithms are available and often very successful as long as the number of attributes (dimensions) is reasonably low. In high dimensions however, many clustering algorithms do not provide meaningful results any more.
In this talk, I will give an overview on the challenges and principle problems of high dimensional data in clustering. The Statistics Department, TU Dortmund provided an artificial example data set that is connected to the BaBar experiment of the Stanford Linear Accelerator. This data set is used as an example to show the problems of high dimensional data sets.