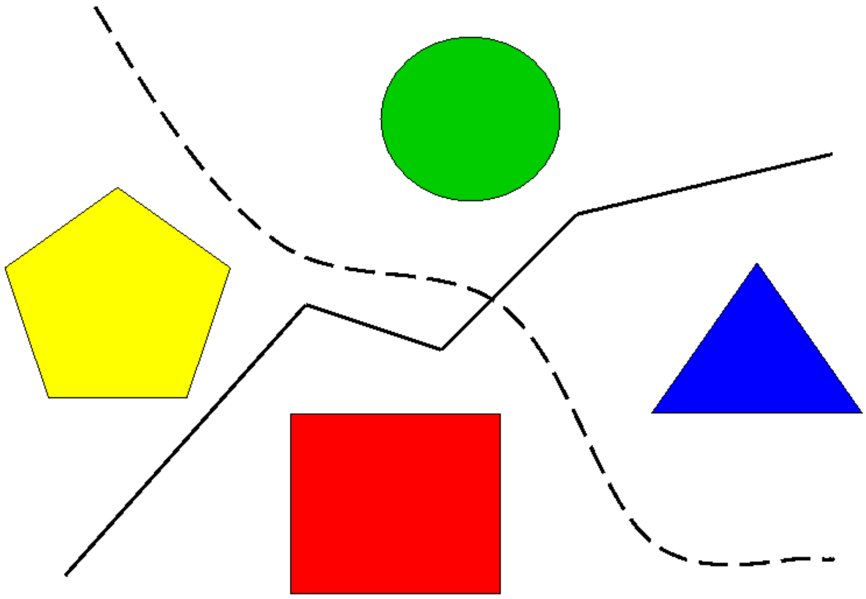
UNIVERSITÄT DORTMUND

# 4. Herbstkolloquium
# des Graduiertenkollegs
# "Statistische Modellbildung"



**Statistische Modellbildung**

**Zu diesem Kolloquium wird eingeladen**

## Freitag / Samstag, 23./24. November 2007

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

Statistische Modellbildung

# 4. Herbstkolloquium des Graduiertenkollegs "Statistische Modellbildung"

***Freitag, 23. November 2007***

**Abfahrt nach Witten:** **ab Dortmund gegen 14.00 Uhr**

## Vortragsprogramm I

| | | |
|---|---|---|
| **15.30 h** | **Begrüßung**<br>**Prof. Dr. Joachim KUNERT** | |
| **15.45 h** | **Hans-Helge MÜLLER**<br>Philipps-Universität Marburg, Institut für Medizinische Biometrie und Epidemiologie, Marburg, Germany | **Design-modifications with control of the type I error rate: Extensions of the CRP-approach considering nuisance parameters or multiple selection** |
| **16.30 h** | **Pierre DRUILHET**<br>Université Blaise Pascal, Laboratoire de Mathématiques, Clermont-Ferrant, France | **A review of recent results in optimal cross-over designs** |
| **17.15 h** | **Augustyn MARKIEWICZ**<br>Agricultural University of Poznán, Department of Mathematical and Statistical Methods, Poznán, Poland | **Optimal designs in multivariate linear models** |

## Diskussionen zu den Projektbereichen

**19.00 h** **Posterausstellung:**
**Präsentation der Dissertations- und Postdoc-Projekte im Kolleg**
**Diskussionen in Arbeitsgruppen**

## *Samstag, 24. November 2007*

### Vortragsprogramm II

| | | |
|---|---|---|
| **9.00 h** | **Karin SAHMER**<br>Institut Universitaire de Technologie de Caen, Département Statistique et Traitement Informatique des Données, Lisieux, France | **Properties of the clustering of variables around latent components. Applications to sensometrics** |
| **9.45 h** | **Daniel MÜLLENSIEFEN**<br>Goldsmith College<br>University of London, United Kingdom | **Dealing with ambiguity in empirical models of music cognition** |

***10.30 h Pause***

### Vortragsprogramm III

| | | |
|---|---|---|
| **11.00 h** | **Simo PUNTANEN**<br>University of Tampere, Department of Mathematics, Statistics and Philosophy, Tampere, Finland | **Decomposing the Watson inefficiency in a linear model** |
| **11.45 h** | **Jarkko ISOTALO**<br>University of Tampere, Department of Mathematics, Statistics and Philosophy, Tampere, Finland | **Some considerations on estimation of the given parametric function under the Gauss-Markov model** |

***12.30 h Pause***

***13:15 bei trockenem Wetter: Gruppenfoto auf der Terrasse***

### Vortragsprogramm IV

| | | |
|---|---|---|
| **13:30 h** | **El-Mostafa QANNARI**<br>ENTIAA/INRA, Nantes, France | **Discrimination and classification with collinear data; applications in near infrared spectroscopy** |
| **14:15 h** | **Håvard RUE**<br>Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway | **Approximate Bayesian inference for latent Gaussian models** |
| **15:00 h** | **Brian RIPLEY**<br>University of Oxford, Department of Statistics, United Kingdom | **Classifiying and combining classifiers** |

***15.40 h Abschlussbesprechung und Diskussion***

**Pierre Druilhet** – Université Blaise Pascal, Laboratoire de Mathématiques, Clermont-Ferrand, France

## A review of recent results in optimal cross-over designs

For chronic diseases, cross-over designs are the most common experimental devices for comparing treatments and selecting the most efficient one.

During the last decade, the theory of optimal cross-over designs has been considerably developed, especially due to the seminal paper by Kushner (1997).

In this talk we present the state of art of optimal cross-over designs and we show that the optimal design may highly depend on the analysis model and on the parameter of interest.

---

**Jarkko Isotalo** – University of Tampere, Department of Mathematics, Statistics & Philosophy, Tampere, Finland

## Some considerations on estimation of the given parametric function under the general Gauss-Markov model

In this talk, we consider estimation of the given estimable parametric function under the general Gauss-Markov model. In particular, we try to highlight some differences which occur between the cases of estimation of the expected value and estimation of the given parametric function. Special emphasis is given to the concepts of linear sufficiency and linear completeness, and to the problem of when the ordinary least squares estimator equals the best linear unbiased estimator.

---

**Augustyn Markiewicz** – Agricultural University of Poznán, Department of Mathematical and Statistical Methods, Poznán, Poland

## Optimal designs in multivariate linear models

Design optimality usually studied under univariate linear normal models, it is extended on multivariate linear normal models.
This extension is straightforward in the case of known dispersion matrix by the use of univariate formulation of the multivariate linear model.
In the case of partially unknown dispersion matrix optimality is considered with respect to the precision in maximum likelihood estimation (MLE). Derivation of information matrices and precision matrices in MLE is presented. The relation between design optimality in univariate model and in its multivariate extensions is studied.

**Daniel Müllensiefen** – Goldsmith College, University of London, London, U.K.

## Dealing with ambiguity in empirical models of music cognition

Music perception and cognition as areas of research have seen a boom in empirical research in recent time. But modelling the data arising from music-psychological experiments often confronts the researcher with the fact that there are several distinct and yet equally valid responses to a music stimulus. In these cases, a single 'ground truth' data set does not exist for a given experimental task. This divergence between human perceptions are assumed to arise from differences in individual music backgrounds, i.e. regarding music listening and playing habits. In this talk, we will present data from three experiments, a similarity rating task, a boundary indication task, and a chord label evaluation task. Each of these data sets requires a different handling of the human response data. Data analysis procedures include measures of inter-subject agreement, model-based clustering, and base-line analysis.

**Hans-Helge Müller** – Philipps-Universität Marburg, Institut für Medizinische Biometrie und Epidemiologie, Marburg, Germany

## Design-modifications with control of the type I error rate: Extensions of the CRP-Approach considering nuisance parameters or multiple selection

During the course of an empirical study it may be advantageous to adapt design elements although the statistical monitoring and analysis has been planned carefully. Redesigning the study by use of the Conditional Rejection Probability (CRP) principle (Müller and Schäfer, 2001, 2004) allows the researcher most flexibility while preserving the overall type I error rate.
In many studies there are two groups to be compared, an experimental group and a control. And for monitoring and analysis of the primary endpoint statistically modelling by means of a Brownian motion with drift and a fixed sample design or a group sequential design is appropriate, at least approximately by asymptotic theory. In such studies one can deal with the unexpected straightforward since the

calculation of a conditional rejection probability is simple by use of standard statistical software packages and since the redesign is nothing else than carefully designing of a new study with the calculated conditional rejection probability taken as the type I error level. Methods for the construction of confidence sets reflecting early stopping for both, significance and futility, e.g. the confidence intervals by Tsiatis et al. (1984), can be developed further to cope with design modifications. However, these calculations will need some more efforts with standard software or will - up to now - need the use of special packages. And connected intervals are no longer guaranteed when using the most common method of Tsiatis et al..
After the introduction of the basic methods for design modifications based on the CRP-principle. the CRP approach will be developed further. Refinements will be presented: first for design extensions in the case of studies with small to moderate sample sizes where statistically modelling with nuisance parameters is more adequate (t-test as an example) and second for the development of multiple testing procedures incorporating selection strategies (flexible marker selection in genome wide association studies as an example).

References

Tsiatis AA, Rosner GL, Metha CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984;40:797-803.
Müller H-H, Schäfer H. Adaptive Group Sequential Designy for Clinical Trials Combining the Advantages of Adaptive and of Classical Group Sequential Approaches. Biometrics 2001;57(3): 886-891.
Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004;23:2497-2508.
Timmesfeld N, Schäfer H, Müller H-H. Increasing the sample size during clinical trials with *t*-distributed test statistics without inflating the type I error rate. *Statistics in Medicine* 2007;26(12):2449-2464.

**Simo Puntanen** – University of Tampere, Department of Mathematics, Statistics & Philosophy, Tampere, Finland

## Decomposing the Watson efficiency in a linear model

This talk is a review of recent results obtained by Chu, Isotalo, Styan, and Puntanen concerning the decomposition of the Watson efficiency of the ordinary least squares estimator of a subset of the parameters. The conditions under which the Watson efficiency in the full model reduces into a function of some other Watson efficiencies is given special attention. In particular, a new decomposition of the Watson efficiency into a product of three particular factors appears to be very useful.

**El-Mostafa Qannari** – ENTIAA/INRA, Nantes, France

## Discrimination and classification with collinear data; application in near infrared spectroscopy

Many rapid analytical techniques such as near infrared spectroscopy involve the calibration of the spectrometer or other instruments in order to predict one or several dependent variables. When the variable to be predicted is nominal (categorical), discrimination analysis and classification methods are generally used in order to establish a prediction model. However, very often, it occurs that the predictor variables are highly collinear. Moreover, the number of predictors may be larger than the number of the available observations. Therefore, it is difficult to establish a stable model using usual techniques. In order to tackle this issue, various methods such as Ridge discrimination, Principal Component discrimination, and Partial Least Squares discrimination analysis (*PLS-DA*) are discussed in the literature. We give herein a general strategy of analysis that encompasses all these techniques. We start by giving a new presentation (formalization) of Discriminant Analysis. This consists in setting up patterns or prototypes associated to the various groups and deriving latent variables in such a way that scores in each group are as highly clustered about their patterns as possible. When the conformity between observations and group patterns is investigated by means of the coefficient of correlation, Fisher's Canonical Discriminant Analysis (*FCDA*) is retrieved. If the covariance is used instead of the coefficient of correlation, a new and simple formalization of PLS Discriminant Analysis (*PLS-DA*) is achieved. It is worth noting that these two methods are respectively based on the eigenstructure of $T^{-1}B$ and $B$ (where $T$ and $B$ are respectively the total and between-groups covariance matrices). Thus, it turns out that *PLS-DA* analysis corresponds to a shrinkage of matrix $T^{-1}$ towards the identity matrix. From this standpoint, it emerges that a hole range of possibilities is offered by considering the eigenstructure of matrix $(\alpha I + (1 - \alpha)T)^{-1}B$, where α is a scalar between 0 and 1. *FCDA* corresponds to the case $\alpha = 0$ whereas *PLS-DA* corresponds to the case $\alpha = 1$. The potential of the general approach is discussed and the methods of analysis are illustrated on the basis of real data sets in near infrared spectroscopy.

**Brian Ripley** – University of Oxford, Department of Statistics, Oxford, U.K.

## Classifying and Combining Classifiers

Classifiers are algorithms to allocate future observations to one of a set of pre-specified classes: the algorithms are almost always developed from a set of 'training' examples.
Classifiers have been developed in several fields: engineering, computer science, machine learning and (of course) statistics. In the last twenty years there have been a number of ideas to combine classifiers into better ones, with names such as 'ensemble methods', 'bagging', 'boosting' and 'stacked generalization'.

The talk will discuss:
- the differences between 'methods' and 'algorithms'
- ways to classify classification methods
- the principles behind combining methods
- ideas on how to choose methods to combine

---

**Håvard Rue** – Norwegian University of Science and Technology, Department of Mathematical Sciences, Trondheim, Norway

## Approximate Bayesian inference for latent Gaussian models

Latent Gaussian models are used within a number of different applications: smoothing-spline models, state-space models, stochastic volatility models, spatial and spatio-temporal models, and log-Gaussian Cox processes among others. The challenging case is when the data is Gaussian so the posterior is analytically intractable. MCMC has "traditionally" been used for inference, but in this talk I will present a deterministic approach: Integrated Nested Laplace Approximations (INLA). The INLA-approach provides approximations the posterior marginals in just a small fraction of the time used for MCMC, and, provide practically exact results in a wide range of examples. The INLA-approach can also be extended to address issues like like detecting ``surprising'' observations, and choosing among different competing models.

---

**Karin Sahmer** – Institut Universitaire de Technologie de Caen, Départmenent Statistique et Traitement Informatique des Données, Lisieux, France

## Properties of the clustering of variables around latent components. Application to sensometrics

In this talk, the properties of the method of clustering of variables around latent components (CLV) are investigated. A statistical model is postulated. This model is especially appropriate for sensory profiling data. It sheds more light on the method CLV. The clustering criterion can be expressed in terms of the parameters of the model.

It is shown that, under weak conditions, the hierarchical algorithm of CLV finds the correct partition while the partitioning algorithm depends on the partition used as a starting point.

Furthermore, the performance of CLV on the basis of a sample is investigated by means of a simulation study. It is shown that this performance is comparable to the performance of known methods such as the procedure Varclus of the software SAS.

*Gruppenfoto 3. Herbstkolloquium, 24./25. November 2006*

*Participants of the 3rd Autumn Symposium, November 24/25, 2006*

Teilnehmer von links nach rechts (*Participants from left to right*)
1. Reihe: Joachim Hartung, Wibke Stansen, Ana Moya, Corinna Auer, Gero Szepannek, Joachim Kunert, Wolfgang Urfer, Katja Ickstadt, Silke Straatmann, Eva Brune
2. Reihe: Christian Hartmann, Götz Trenkler, Katrin Kuhr, Anne Krampe, Sibylle Sturtz, Katrin Sommer, Shalabh, Sonja Kuhnt
3.Reihe: Jonas Kaiser, Gernot Fink, Ingileif Halgrímsdóttir, Werner Brannath, Julia Schiffner, Vladislav Ponyatovskyy, Claus Weihs
4. Reihe: Martin Gebel, Henning Henke, Roland Fried, Andrea Preußer, Nina Kirschbaum, Oliver Sailer, Walter Krämer
5. Reihe: Björn Bornkamp, Michael Hauptmann, Nils Raabe , Uwe Ligges