ulm university universität ulm

Statistical Computing 2018
Abstracts der 50. Arbeitstagung

HA Kestler, M Schmid, L Lausser,
JM Kraus, A Fürstberger (eds)

Ulmer Informatik-Berichte

Nr. 2018-01
July 2018

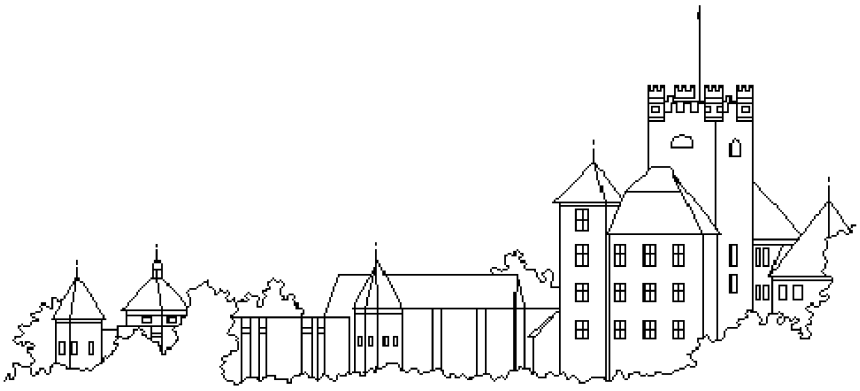
International Graduate School
in Molecular Medicine Ulm

Sponsored by

International Graduate School in Molecular Medicine Ulm (one invited speaker)

RStudio® ($1.000)

Statistical Computing 2018



50. Arbeitstagung

der Arbeitsgruppen Statistical Computing (GMDS/IBS-DR), Klassifikation und Datenanalyse in den Biowissenschaften (GfKl).

08.07. - 11.07.2018, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, July 08, 2018

20:00		**Opening of the workshop: H. A. Kestler**
		Introduction: H. A. Kestler
20:00 – 21:00	Jan Beyersmann (Ulm)	Sampling and Re-Sampling Complex Time-to-Event data: Beyond Standard Cohort and Standard Bootstrap

Monday, July 09, 2018

09:20 – 10:20		**Chair: J. Beyersmann**
09:20 – 09:40	Marius Greiff (Dortmund)	Optimization of stochastic functions by optimizing expectation and variance
09:40 – 10:00	Moritz Berger (Bonn)	Tree-Structured Modeling of Time-Varying Coefficients for Discrete Time-to-Event Data
10:00 – 10:20	Hryhorii Chereda (Göttingen)	Graph-based Convolutional Neural Networks for analyzing pathways in cancer
10:20 – 10:50		**Coffee Break**
		Introduction: M. Schmid
10:50 – 11:50	Lutz Edler (Heidelberg) and Berthold Lausen (Essex)	Statistical Computing in Germany : The early years of "Arbeitstagungen der Arbeitsgruppen" and the flair and impact of Reisensburg
12:00 – 13:30		**Lunch**
		Introduction: A. Fürstberger
13:40 – 14:20	Jennifer Pohle (Bielefeld)	State architectures in hidden Markov models
14:20 – 15:00		**Chair: A. Fürstberger**
14:20 – 14:40	Colin Griesbach (Erlangen)	New insights in how to boost joint models for longitudinal and time-to-event outcomes
14:40 – 15:00	Anja Rappl (Erlangen)	More than one way: Exploring the capabilities of different estimation approaches to joint models
15:00 – 15:30		**Coffee Break**

Monday, July 09, 2018

15:30 – 16:30		**Chair: R. Backofen**
15:30 – 15:50	Andrea Bommert (Dortmund)	Benchmark for Filter Methods for Feature Selection in High-dimensional Data
15:50 – 16:10	Robin Szekely (Ulm)	Utilizing foreign classes for feature selection
16:10 – 16:30	Giuseppe Casalicchio (München)	Visualizing the Feature Importance for Black Box Models
16:30 – 16:50		**Break**
		Introduction: P. Schlosser
16:50 – 17:50	Phil Bowsher (Indianapolis)	An Introduction to Shiny and R Markdown With Applications in Research and Drug Development
18:00 – 20:00		**Dinner**
20:00 – 21:00	Phil Bowsher (Indianapolis)	Hands-on Tutorial: An Introduction to Shiny and R Markdown With Applications in Research and Drug Development

Tuesday, July 10, 2018

09:20 – 10:20		**Chair: G. Casalicchio**
09:20 – 09:40	Roman Hornung (München)	Random forests for multi-omics data
09:40 – 10:00	Jörn Lötsch (Frankfurt am Main)	A machine-learning based approach to the association of next-generation sequencing derived genotypes with (pain-related) phenotypes
10:00 – 10:20	Lisa Schäfer (Ulm)	Repetitive structures pass screening for ordinal relationships
10:20 – 10:50		**Coffee Break**
		Introduction: A. Benner
10:50 – 11:50	Rolf Backofen (Freiburg)	Computational Genomics: Informatics Problems and Applications in Life Sciences
12:00 – 13:30		**Lunch**

Tuesday, July 10, 2018

		Introduction: A. Mayr
13:40 – 14:20	Maike Hohberg (Göttingen)	Mixed ordered-continuous copula GAMLSS with an application to poverty dimensions
14:20 – 15:00		**Chair: A. Mayr**
14:20 – 14:40	Daniel Schalk (München)	`compboost`: A Modular Framework for Component-Wise Boosting in R
14:40 – 15:00	Elisabeth Waldmann (Erlangen)	A Discussion on Adaptive Step-lengths in Gradient Boosting
15:00 – 15:30		**Coffee Break**
15:30 – 16:30		**Chair: T. Beißbarth**
15:30 – 15:50	Malte Jastrow (Dortmund)	Multimodal likelihood functions occurring in mixture modeling
15:50 – 16:10	Marcus Vollmer (Greifswald)	Randomized Stepwise Regression – a Modification of Stepwise Methods in Generalized Linear Models and its Application on Sepsis Data
16:10 – 16:30	Regina Stegherr (Ulm)	Analysing unmeasured baseline covariates in studies with delayed entry using a joint model: fitting the model in presence of left-truncation
16:30 – 16:50		**Break**
16:50 – 17:50		Working group meeting on **Statistical Computing 2019**
18:00 – 20:00		**Dinner**

Wednesday, July 11, 2018

09:20 – 10:20		**Chair: R. Schuler**
09:20 – 09:40	Sebastian Krey (Köln)	Tensile Stress Monitoring of Fibre-Reinforced Plastic
09:40 – 10:00	Tobias Hepp (Bonn)	Dynamic reference intervals from contaminated data sources
10:00 – 10:20	Jakob Richter (Dortmund)	Selection of Optimal Subgroup Weights for Survival Analysis
10:20 – 10:50		**Coffee Break**
10:50 – 11:50		**Chair: L. Lausser**
10:50 – 11:10	Thomas Welchowski (Bonn)	Correlation-Adjusted Regression Survival Scores for High-Dimensional Variable Selection
11:10 – 11:30	Dominic Edelmann (Heidelberg)	The distance correlation coefficient for right-censored survival data
12:00 – 13:30		**Lunch**
13:30		**Departure**

Contents

Sampling and re-sampling complex time-to-event data: beyond standard cohort and standard bootstrap

Jan Beyersmann[1]

Roughly two thirds of all primary outcomes in medical research are survival or time-to-event outcomes. The statistical standard approach is a Kaplan-Meier analysis, accompanied by a Cox model if regression and testing is desired, applied to prospective cohort data. This talk will consider situations where either a) collecting full cohort data is too expensive or logistically even infeasible or b) more informative outcomes than the common "time-to-composite-endpoint" are investigated. Our motivating data examples are a) an ongoing observational study on the occurrence and the impact of infections resistant to Carbapenem, an antibiotic of last resort, and b) a randomized clinical trial on Graft-versus-Host-Disease prophylaxis in leukemia patients undergoing stem cell transplantation. In the first example, occurrence of the infection is both a time-to-event outcome, subject to competing risks, and a time-dependent exposure, and one practical challenge is that all infected patients shall be sampled, but data on only a fraction of the non-infected patients shall be collected. In the second data example, a relevant outcome is the probability to be alive without immunosuppressive therapy, but neither Kaplan-Meier nor competing risks methodology applies to analyzing this non-monotone probability curve, but resampled time-simultaneous confidence bands do.
The link between these two problems is the modern counting process/martingale approach to event history data which we will exploit to a) analyse data where certain patient trajectories are oversampled and b) resample event times that are not necessarily independent. The latter is required by the common Efron bootstrap, but neither holds in the presence of oversampling nor in common pharmaceutical studies. The talk will start with some survival or event history fundamentals, namely that the analysis of time-to-event data is based on hazards and that there is an overemphasis on survival functions and the Kaplan-Meier method.

[1] Institute of Statistics, Ulm University, 89081 Ulm, Germany

jan.beyersmann@uni-ulm.de

Optimization of stochastic functions by optimizing expectation and variance

Marius Greiff [1], *Günter Rudolph* [1]

For deterministic functions, the same input always leads to the same output. For stochastic functions, this is not true. So, when optimizing a stochastic function f, one has to deal with the problem of varying outputs $f(x, C)$ for the same input x due to the effects of a random variable C. One possibility is to consider the expectation and variance of $f(x, C)$ and to choose x such that the expected value of $f(x, C)$ is optimal (e.g. maximal) and the variance of $f(x, C)$ is minimal. Minimal variance of $f(x, C)$ can be interpreted as minimal expected deviation from the expected value.

We investigate the optimization of expensive stochastic functions $f(x, C)$ with $x \in \mathbb{R}$ and C being a one dimensional random variable. Because f is an expensive function, we want to evaluate it seldom. Therefore, we use a surrogate model of f to estimate the expectation $\mathrm{E}(f(x, C))$ and variance $\mathrm{V}(f(x, C))$. For example the expectation can be estimated as

$$\hat{E}(f(x, C)) = \int_{c_l}^{c_u} \hat{f}(x, c) \cdot p_C(c) dc$$

using numerical integration. c_l and c_u denote the lower and upper integration limits, $\hat{f}$ is the surrogate model of f and p_C is the probability density function of C.

Aiming for both an optimal expectation and a minimal variance results in a Pareto front. Our goal is to get a good approximation of the true Pareto front, i.e. accurate estimates for expectation and variance. We perform a simulation study to analyze how well our approach works. We consider different functions f and different numbers of evaluations of f.

[1] Department of Computer Science, TU Dortmund University, 44221 Dortmund, Germany

marius.greiff@tu-dortmund.de, Guenter.Rudolph@tu-dortmund.de

Tree-Structured Modeling of Time-Varying Coefficients for Discrete Time-to-Event Data

Moritz Berger[1], Marie-Therese Puth[1], Gerhard Tutz[2] and Matthias Schmid[1]

Time-to-event models are a popular tool to analyse data where the outcome variable describes the time to the occurrence of a specific event of interest. We focus on the analysis of time-to-event outcomes that are measured on a discrete time scale, which is a likely scenario when events occur between pairs of consecutive points in time (e.g., between two follow-up visits of an epidemiological study).

In the literature there exists a variety of regression models for discrete time-to-event data, see for example [1]. The main principle is to model the *discrete hazard* function $\lambda(t|\boldsymbol{x}) = P(T = t|T \geq t, \boldsymbol{x})$, $t = 1, 2, \ldots$, where T denotes the event time and $\boldsymbol{x}$ is a set of explanatory variables $\boldsymbol{x}^\top = (x_1, \ldots, x_p)$. Commonly it is assumed that the effects of the explanatory variables are constant over time, and the dependency on time is solely modeled by separate intercepts for each t, also referred to *baseline coefficients*. In many applications, however, this assumption may be too restrictive.

A more general approach is to allow the effects to vary over time. Frequently it is natural to assume that the coefficients are represented by smooth functions of t. A popular strategy is to use *P-splines* propagated by [2], where the smooth functions are expanded by B-spline basis functions and an additional *penalty term* is used to obtain stable estimates.

Alternatively, it is often more appropriate to assume that the effects of the explanatory variables do not vary over the whole range of t, but are constant over several adjacent points in time. That is, one assumes that the time-varying coefficients are represented by *piecewise constant* functions. These functions can be obtained by using recursive partitioning techniques. To address this issue, we propose to utilize the tree-based approach that was recently introduced by [3]. By iterative splitting in one of the explanatory variables (with regard to the effect modifier t) the method yields a tree for each variable that shows time-varying coefficients. Thereby, the algorithm itself identifies the coefficients (corresponding to an explanatory variable) that deviate from a constant and the corresponding partitioning of the observation times.

[1]Department of Medical Biometry, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität Bonn, Sigmund-Freud-Straße 25, 53127 Bonn, Germany
[2]Department of Statistics, Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 München, Germany

moritz.berger@imbie.uni-bonn.de, puth@imbie.meb.uni-bonn.de, gerhard.tutz@stat.uni-muenchen.de, matthias.schmid@imbie.uni-bonn.de

References

1 Tutz, G. and M. Schmid (2016). *Modeling Discrete Time-to-Event Data*. New York: Springer.

2 Eilers, P. and B. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*, 89-102.

3 Berger, M., G. Tutz and M. Schmid (2018). Tree-structured modelling of varying coefficients. *Statistics and Computing*. doi: 10.1007/s11222-018-9804-8.

Graph-based Convolutional Neural Networks for analyzing pathways in cancer

Hryhorii Chereda[1], Frank Kramer[1], Tim Beißbarth[1]

In recent years deep learning was applied to a wide range of problems in various areas. Such deep learning tools as convolutional neural networks (CNNs) have been shown to work well in natural language processing and computer vision, especially at image classification tasks. Furthermore, CNN's have been applied to bioinformatic challenges like patient stratification tasks. Nowadays, deep learning (including CNNs) is extending to Non-Euclidean domains such as graph-structured data and manifolds. We are planning to map gene-expression data to the vertices of biological pathways and feed this graph-structured data into CNNs in order to classify patients.
The usual CNN architecture consists of three types of layers: convolutional layers, pooling layers, and fully connected layers. The first two layers utilize the structure of the data preparing informative features for the fully connected neural network layers. In our work, we consider three popular, but different approaches developed for application of CNN on graph-structured data.
Our research aims to compare these approaches in order to address the question if the use of graph-based CNNs is able to provide valuable classification improvements by utilizing prior pathway knowledge.
Preliminary results show that the utilizing of the WNT signaling pathways as a prior knowledge does not improve the performance of the classifier in the case of breast cancer patients. The set of features covered by the WNT pathways (hundreds of genes) is not informative enough. Hence, the future work will concern more informative biological pathways that cover thousands of genes.

[1] University Medical Center Goettingen, Department of Medical Statistics, Germany

hryhorii.chereda@med.uni-goettingen.de

Statistical Computing in Germany : The early years of "Arbeitstagungen der Arbeitsgruppen" and the flair and impact of Reisensburg

Lutz Edler[1]

The "Statistical Computing 2018" as 50th Meeting of the Working Groups (WG) "Statistical Computing" (GMDS/DR) and Biostatistics (DfKl) at Reisensburg should be a reason for breathing-time and thinking back. This contribution is specific for celebrating the 50^{th} Reisensburg conference and aims to address the origins and the first periods of the famous Reisensburg conferences until the mid 1990ties and the 25^{th} meeting [1], i.e. about the time when the DfKL formally joint the series of meetings around 1994 and when heat-lightening of big (genomic) data started. No much time should be spent to the pre-history of computational statistics beginning in the 16th century when both statistics and computing started at about the same time [2], a fact that sheds some light on the duality of "statistical computing" and "computational statistics" the two labels chosen by GMDS and the German Region (DR) of the International Biometric Society (IBS) for the two WGs. Therefore, this presentation will consider mainly in the last four decades of the last century. We start with a view on the 1960ties when the dinosaur mainframe computers (IBM and UNIVAC) and languages such as FORTRAN and ALGOL became available, when modern programming started and when soon the first statistical programs were developed. The first three Reisensburg "Arbeitstagungen" were two meetings in 1974 and one in 1975, imbedded in a series of seven meetings all framed by the DVM-107 BMFT project initiated, organized and headed by Norbert Victor [3,4]. This project addressed, in particular, portability of programs, standardization and program evaluation and their segmentation (later called modularity!) and developed SYSTAT as program package. During the 1970ties, it must have been the flair of Reisensburg which fixed the "Arbeitstagung" at Reisensburg Castle from the 9th meeting onwards, mainly sponsored by the MEDIS Institute and GSF (Munich) and such connected again with Norbert Victor [3] and soon also with Allmut Hörmann [5]. The 1970ties exhibited also the first statistical analysis systems BMDP, SPSS, P-STAT and GENSTAT in UK. The working group "Statistische Auswertesysteme" of the GMDS was founded on 4^{th} October 1974 in Mainz with its first Reisensburg Meeting from 2^{nd} to 4^{th} April 1975. The next step to settle statistical computing at Reisensburg took place during the 1980ties by the foundation of the WG "Computational Statistics" within the DR in 1984 and a first joint meeting of the two WGs of GMDS and DR $24^{th} - 27^{th}$ June 1985 with Allmut. Hörmann, Reinhold Karl-Heinz Jöckel representing the two WGs [6]. The 1980ties saw the rise of the PC and workstation but also the "collision" of statistical computing with modern statistical methodology such as survival analysis, generalized linear models, robust and exact methods, exploratory data analysis and graphical analysis, bootstrap, and expert

[1] Division of Biostatistics-C060, German Cancer Research Center,,Im Neuenheimer Feld 280, 69120 Heidelberg, retired

edler@dkfz.de / lutzedler@gmx.de

and knowledge based systems. This fruitful inclusion of new (bio)statistical methods has characterized Reisensburg conference continued since then and may be seen as one of its pillars of success. Relations between Reisensburg and COMPSTAT, the International Association for Statistical Computing (IASC) and the Journal of Computational Statistics and Data Analysis (CSDA) will be addressed.

References

1 Dirschedl P and Ostermann R: Computational Statistics. Papers Collected on the Occasion of the 25th Conference on Statistical Computing at Schloß Reisensburg. Springer Berlin/Heidelberg, 1994

2 Sint PP: Remark on the history of computational statistics in Europe. In [1]

3 Victor N: The roots of computational statistics in Germanny. In [1]

4 Edler L: Obituary for Professor Dr. Dr. h.c. Norbert Victor (1940-2011).CSDA 56(5),998-1051, 2012

5 Hörmann A: Twenty-five working conferences on statistical computing-reflections on twenty years of Reisensburg Meetings. In [1]

6 Edler L: Computational Statistics und Biometrie. Wer treibt wen? Computational statistics and biometry. Which discipline drives which? GMDS Med.Inform.Biom.Epidemiol. 2005,1 (2): Doc10

State architectures in hidden Markov models

Jennifer Pohle[1]

Hidden Markov models (HMMs) are flexible, doubly stochastic time series models which find application in many different disciplines, such as speech recognition, economics, biology and medicine (for an overview, see e.g. Zucchini et al., 2016). In its basic form, an HMM assumes the observations to be driven by an underlying unobserved Markov chain with finite state space. Each state is associated with a specific distribution that generates the observation when the state is active. To completely specify such a basic HMM and fit it to data, it is necessary to choose a family of state-dependent distributions and to fix the number of states. After a general introduction to HMMs, this talk focuses on the construction and dimension of the state space of an HMM.
First, in scenarios where the number (and interpretation) of the states is not clear a priori, the selection of a model with an appropriate number of states is of crucial importance. However, traditional information criteria like AIC and BIC tend to select models with a very large number of states, leading to models that are not interpretable anymore (Pohle et al., 2016). The main reason for this is that basic HMMs as outlined above involve a very simplistic dependence structure, and the real data-generating process typically exhibits more complex structures and dependencies than are accounted for in such simple models. However, additional states are often able to absorb these structures, leading AIC and BIC to select models with a larger number of states than the number of genuine states. This is demonstrated using simulation studies and a case study on animal movement data. When feasible, then the model formulation should of course be improved to overcome any substantial lack of fit. However, incorporating all structures in the model is often computational infeasible. Therefore, a pragmatic step-by-step approach is proposed to select the number of states in cases where the focus lies on meaningful inference.
In the last part of the talk, a scenario with a more complex dependence structure will be described in more detail, namely multivariate time series in which the different variables are correlated but do not evolve synchronously. This can be modeled using coupled hidden Markov models (Brand, 1997). They can be seen as HMMs with an extended state space, such that the standard HMM techniques, for example for parameter estimation, state decoding, and residual analyses, can still be applied. We further extend coupled HMMs to let the observed variables additionally depend on covariates, which leads to coupled Markov switching regression models. The potential usefulness of this new class of models is illustrated using a case study on electronic health record data.

[1] Bielefeld University, Universitätsstr. 25, 33615 Bielefeld

jennifer.pohle@uni-bielefeld.de

References

1 Brand, M. (1997): Coupled hidden Markov models for modeling interacting processes. *Technical Report 405*, Massachusetts Institute of Technology Media Laboratory, Cambridge.

2 Pohle, J., Langrock, R., van Beest, F. M., Schmidt, N. M. (2017): Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, **22**, 270-293.

3 Zucchini, W., MacDonald, I. L. & Langrock, R. (2016), Hidden Markov Models for Time Series: An Introduction using R, Second Edition, Chapman & Hall/CRC, Boca Raton.

New insights in how to boost joint models for longitudinal and time-to-event outcomes

Colin Griesbach[1], *Andreas Mayr*[2], *Elisabeth Waldmann*[1]

When analyzing data where event-times are recorded alongside a longitudinal outcome, one commonly used approach in practice is separate modeling of the two outcomes without considering any interaction effects. Joint modeling on the other hand accounts for those interactions by combining a longitudinal and a survival submodel in one single joint likelihood. This approach corrects for event-dependent dropout in longitudinal analysis as well as for endogenous, time-varying covariates measured with error in survival analysis. Waldmann *et al.* [2] initially proposed an algorithm to fit a joint model via gradient boosting techniques discussed in [1]. Aim of the present work is to lay more focus on the survival component of the model and expand the algorithm to also allow for variables which only have an impact on the survival part of the model. Furthermore advantages of flexible step-lengths ν originally discussed in [3] with respect to reduced computational effort are presented.

References

1 Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012): Generalized additive models for location scale and shape for high-dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society Series C - Applied Statistics*. 61(3): 403-427.

2 Waldmann, E., Taylor-Robinson D., Klein N., Kneib T., Pressler T., Schmid, M. and Mayr, A. (2017): Boosting Joint Models for Longitudinal and Time-to-Event Data. *Biometrical Journal*.

3 Waldmann, E., Griesbach, C. and Mayr, A. (2018) Variable Selection and Allocation in Joint Models for Longitudinal and Time-to-Event Data via Boosting *in preparation*

[1] Department of Medical Informatics, Biometrics and Epidemilology, Friedrich-Alexander-Universität Erlangen-Nürnberg
[2] Department of Medical Biometrics, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität Bonn

colin.griesbach@fau.de, amayr@uni-bonn.de, elisabeth.waldmann@fau.de

More than one way: Exploring the capabilities of different estimation approaches to joint models.

Anja Rappl[1], Elisabeth Waldmann[1]

Confronted with data capturing longitudinal and time-to-event outcomes simultaneously the data generating processes are best modelled jointly to avoid bias and account for data drop-out [1,2]. The rising popularity of these joint models in recent years fostered the development of different software routines for facilitated analysis incorporating the current state of research. However, these routines differ in model formulation, estimation technique and analytical purpose. Therefore, a comparison of different routines may identify their capabilities and limitations.
The focus of our comparison will be on `R` routines exclusively. The `R` package `JM` by [4] was one of the first routines available for joint modelling and `joineRML` by [3] expands joint models to a multivariate setting. Both use expectation maximization for estimation. In contrast, `JMboost` by [5] uses boosting and thus is the first package to offer automated variable selection. These three candidates will be examined via a simulation study and a real data example. The results may serve as a guideline for applicants and as a promoter for a more routine application of joint models.

References

1 Wulfsohn, Michael S. and Anastasios A. Tsiatis: A Joint Model for Survival and Longitudinal Data Measured with Error In: Biometrics, 53.1, (1997).

2 Henderson, Robin, Peter Diggle, and Angela Dobson: Joint modelling of longitudinal measurements and event time data In: Biostatistics (Oxford, England), 1.4, (2000).

3 Hickey, Graeme L., Pete Philipson, Andrea L. Jorgensen, and Ruwanthi Kolamunnage-Dona: joineRML: A joint model and software package for time-to-event and multivariate longitudinal data (2017).

4 Rizopoulos, Dimitris: JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data In: Journal of Statistical Software, 35.9, (2010).

5 Waldmann, Elisabeth, David Taylor-Robinson, Nadja Klein, et al.: Boosting joint models for longitudinal and time-to-event data In: Biometrical journal. Biometrische Zeitschrift, (2017).

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany, http://www.imbe.med.uni-erlangen.de/cms/

`anja.rappl@fau.de, elisabeth.waldmann@fau.de`

Benchmark for Filter Methods for Feature Selection in High-dimensional Data

Andrea Bommert[1], Xudong Sun[2], Bernd Bischl[2], Jörg Rahnenführer[1], and Michel Lang[1]

Feature selection has become increasingly important for data analysis. Especially for high-dimensional datasets, it is necessary to filter out the irrelevant features in order to avoid overfitting. With respect to datasets from the bioinformatics domain, feature selection also often allows identifying the features which are relevant for biological processes. When fitting a model for such high-dimensional datasets, one needs to decide first, which feature selection method to use. As there are very many of them, this is not an easy decision.
The purpose of this talk is to compare and to benchmark filter methods. We investigate a variety of different filter methods based on many high-dimensional datasets from various domains. We aim to find out, which methods select the features of a dataset in a similar order. Additionally, we search for optimal filter methods with respect to the predictive accuracy of the selected features and to runtime.

[1] TU Dortmund University, 44221 Dortmund, Germany
[2] Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 München, Germany

bommert@statistik.tu-dortmund.de, xudong.sun@stat.uni-muenchen.de,
Bernd.bischl@stat.uni-muenchen.de, rahnenfuehrer@statistik.tu-dortmund.de,
lang@statistik.tu-dortmund.de

Utilizing foreign classes for feature selection

Robin Szekely[1,2], Ludwig Lausser[1], Viktor Kessler[1], Friedhelm Schwenker[3] and Hans A. Kestler[1]

Reduction to a small set of interpretable biomarkers is one of the key interests in modeling in the context of precision medicine [1]. These might reveal key aspects of a biological system or even the causes of a disease. In a diagnostic setting, the corresponding marker selection processes are typically focused on the analysis of the specific classes of interest. Other entities that might be related to this classification task are usually neglected. However, foreign classes might be influenced by effects that also affect the classes of interest. In this manner, a new tumor classification task can possibly be characterized by a set of biomarkers obtained from another tumor classification experiment. Considering multiple genomes of interest in such a way is a core aspect of pan-genomic analysis [2].
In a previous work, we analyzed feature selection in the context of multi-class classification [3]. We observed that certain feature selection strategies mimic the feature signature of one of their binary base classifiers while excluding selected features from most of the other class combinations. Nevertheless, the corresponding multi-class architectures led to accuracies comparable to those of more heterogeneous feature selections. Following up this work, we systematically evaluate binary classifiers that operate on feature signatures designed for the discrimination of a pair of foreign classes.
We analyze these classifiers in 10×10 cross-validation experiments on 9 publicly available microarray datasets with multiple diagnostic classes. For our experiments, we utilized linear support vector machines, 3-nearest neighbour classifiers and random forests. For 8 out of 9 datasets, we could observe foreign feature combinations that outperformed at least 90% of those feature sets designed for the original classes.

References

1 Kraus J.M., Lausser L., Kuhn P., Jobst F., Bock M., Halanke C., Hummel M., Heuschmann P., Kestler H. A. (2018) Big data and precision medicine: challenges and strategies with healthcare data. International Journal of Data Science and Analytics, 1-9.

2 Vernikos, G., Medini, D., Riley, D. R., Tettelin, H. (2015). Ten years of pan-genome analyses. Current Opinion in Microbiology, 23, 148-154.

3 Lausser, L.*, Szekely, R.*, Schirra, L. R., Kestler, H. A. (2017). The Influence of Multi-class Feature Selection on the Prediction of Diagnostic Phenotypes. Neural Processing Letters, 1-18.

[1] Institute of Medical Systems Biology, Ulm University, 89081 Ulm, Germany
[2] International Graduate School in Molecular Medicine Ulm, Ulm University, 89081 Ulm, Germany
[3] Institute of Neural Information Processing, Ulm University, 89081 Ulm, Germany

robin.szekely@uni-ulm.de, ludwig.lausser@uni-ulm.de, hans.kestler@uni-ulm.de

Visualizing the Feature Importance for Black Box Models

Giuseppe Casalicchio[1], *Christoph Molnar*[1], *Bernd Bischl*[1]

In recent years, a large amount of model-agnostic methods to improve the transparency, trustability and interpretability of machine learning models have been developed. We introduce local feature importance as a local version of a recent model-agnostic global feature importance method. Based on local feature importance, we propose two visual tools: partial importance (PI) and individual conditional importance (ICI) plots which visualize how changes in a feature affect the model performance on average, as well as for individual observations. Our proposed methods are related to partial dependence (PD) and individual conditional expectation (ICE) plots, but visualize the expected (conditional) feature importance instead of the expected (conditional) prediction. Furthermore, we show that averaging ICI curves across observations yields a PI curve, and integrating the PI curve with respect to the distribution of the considered feature results in the global feature importance. Another contribution of our paper is the Shapley feature importance, which fairly distributes the overall performance of a model among the features according to the marginal contributions and which can be used to compare the feature importance across different models.

[1] Department of Statistics, Ludwig-Maximilians-University Munich

giuseppe.casalicchio@stat.uni-muenchen.de

An Introduction to Shiny and R Markdown with Applications in Research and Drug Development

Phil Bowsher[1]

RStudio will be presenting an overview of Shiny, R Markdown and HTML Widgets for the R user community at Statistical Computing 2018 on Monday, July 9th 2018. This is a great opportunity to learn and get inspired about new capabilities for creating compelling analyses with applications in research and drug development. No prior knowledge of R, RStudio or Shiny is needed. This short course will provide an introduction to flexible and powerful tools for statistical analysis, reproducible research and interactive visualizations. The hands-on course will include an overview of how to build Shiny apps and R Markdown documents. Immunogenicity examples will be reviewed and generated for each topic.

Shiny is an open source R package that provides an elegant and powerful web framework for building web applications using R. Shiny combines the computational power of R with the interactivity of the modern web. Shiny allows users the flexibility of pulling in whatever package in R needed to solve a problem. There are no limits to the types of applications one can build, and no constraint on the visualizations that can be used. Developers get the benefit of an open source ecosystem for R, along with the open source ecosystem for Javascript visualization libraries, thereby allowing one to create highly custom applications. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge. This powerful concept allows you to easily deliver results as interactive data explorations instead of static reports to your stakeholders and non R users. Immunogenicity assessments via Shiny will be covered. An introduction to databases via R will be reviewed along with how to connect Shiny apps to databases. An introduction to creating web APIs with your existing R code will also be discussed.

R Markdown is an authoring format that enables easy creation of dynamic documents, presentations, and reports from R. It combines the core syntax of markdown with embedded R code chunks that are run so their output can be included in the final document. R Markdown documents help to support reproducible research and can be automatically regenerated whenever underlying R code or data changes. Various types of R Markdown output will be covered, including blogdown and bookdown. An R Notebook is an R Markdown document with chunks that can be executed independently and interactively, with output visible immediately beneath the input. R Notebooks can be thought of as a special execution mode for R Markdown documents. The OpenFDA package and immunogenicity assessments will be used for the course examples regarding R Markdown reports and R Notebooks.

RStudio will be showcasing several compelling examples as well as learning resources. As part of the short course, some available drug development-related R Shiny apps and R Markdown reports will be illustrated.

[1] Director of Life Sciences & Healthcare at RStudio, Inc. , Indianapolis

phil@rstudio.com

Random forests for multi-omics data

Roman Hornung[1] *and Marvin N. Wright*[2]

In the last years, the availability of gene expression data, methylation data, copy-number data and other types of omics data is increasing rapidly. As a consequence, there is also more multi-omics data becoming available, that is, data for which measurements of different omics data types are present for the same patient. Multi-omics data are well suited for obtaining prediction rules for phenotypic outcomes, such as response to therapy, because each of the involved omics data types can be expected to contribute some unique information valuable for prediction.
However, the structure of multi-omics data is complex. First, the information contained in the different omics data types, denoted as blocks in the following, is strongly overlapping. Second, the levels of predictive information contained in the different blocks differ and depend on the outcome considered. Third, there are strong interactions between variables from different blocks. The complexity of these data makes it challenging to exploit the available predictive information.
The prediction method random forest is known to be able to capture complex dependency structures between the outcome and the covariates. Therefore, in this project we develop a variant of the random forest method that is able to adequately take the block structure of multi-omics data into account. In this process, we discuss several sensible, potentially useful approaches that are applicable to binary, metric and survival outcomes. These approaches are implemented and compared using real multi-omics data with a survival outcome.

[1] Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, Munich 81377, Germany
[2] Leibniz Institute for Prevention Research and Epidemiology, BIPS, Achterstr. 30, Bremen 28359, Germany

hornung@ibe.med.uni-muenchen.de, wright@leibniz-bips.de

A machine-learning based approach to the association of next-generation sequencing derived genotypes with (pain-related) phenotypes

Jörn Lötsch[1,2] and Dario Kringel[1]

Pain, as many other medical conditions, has a complex pathophysiology [1] requiring complex research approaches to reduce its still high medical medial and socio-economic impact [2]. One of the most promising approach to pain and its development toward persistence is the study of the genetic architecture of pain, which has been pursued for more than 50 years [3]. While recent research approaches focused on functional single genetic variants, this limited selection is increasingly perceived as an insufficient basis for genetic diagnostics [4]. Indeed, with the broader availability of next generation sequencing (NGS) the restriction of the analysis to known functional variants has fallen in favor of the full accessibility of the subject's genetic information. However, the resulting large amount of data poses challenges upon data analysis. Furthermore, to challenge functional phenotype association with genotype related data by increasing the size of the study sample as with classical genetic approaches becomes increasingly difficult and cost intensive and the success remains doubtful.

We propose a novel approach to genotype phenotype association that is based on machine learning. The main idea is to train an artificial intelligence, implemented as different types of supervised machine learning, to learn the association of the genetic information with the pain-related phenotype, and to utilize subsequently the trained intelligence to predict a phenotype in new data from genetic information. If this application performs better than guessing the phenotype a genotype-phenotype association can be concluded as supported by the data.

The approach was applied to a data set obtained in a human experimental pain setting where n = 82 subjects had been investigated with respect to the pain perception thresholds of noxious heat stimuli before and after sensitization with capsaicin, an activator of TRPV1 ion channels [5]. Firstly, a phenotype class structure was obtained by means of Gaussian mixture modeling that indicated two phenotype groups with high or low capsaicin-induced hypersensitization to heat. Secondly, a role of the genic information in this two-classes problem was assessed. NGS genotype information was available from n = 75 subjects [6] and consisted of d = 278 loci of the TRPV1 gene and, in addition, of the TRPA1 gene that has also been implicated previously in human heat pain perception. Following data preprocessing aimed at eliminating non-informative genetic variants by applying Shannon information, chi square statistics and computed ABC analysis, d = 31 genetic variants were further analyzed for association with pain heat related phenotypes. Several methods of supervised machine learning implemented as (i) random

[1] Institute of Clinical Pharmacology, Goethe - University, Theodor Stern Kai 7, 60590 Frankfurt am Main, Germany

[2] Fraunhofer Institute of Molecular Biology and Applied Ecology - Project Group Translational Medicine and Pharmacology (IME-TMP), Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany

j.loetsch@em.uni-frankfurt.de, Kringel@med.uni-frankfurt.de

forests, (ii) adaptive boosting, (iii) k-nearest neighbors, (iv) naïve Bayes, (v) support vector machines, and for comparison, (vi) binary logistic regression predicted, in a cross-validated scenario using 1,000 Monte-Carlo random sample splits into 2/3 training and 1/3 test data subsets, the phenotype group association consistently better when based on the observed genotypic information, than when using a random permutation of the exomic sequences. The balanced accuracy of the prediction, however, never exceeded 65%, which was nevertheless the expected result considering the complexity of pain unlikely to be solely controlled by two single genes. To exclude that this performance was, however, due to poor implementation of the machine learning methods, the classification was repeated using a positive control data set created by sorting the original genotype information at each locus in descending order data of the number of variant alleles along the sorted phenotype classes. In these data, all machine-learned methods classified with 100% accuracy or close to that. By contrast, classical chi square based statistics failed to produce significant results when applying alpha correction.
The analysis shows that the complexity of the genotype is a relevant factor and machine-learned methods provide biologically plausible results, outperforming classical statistical genotype versus phenotype association analyses.

Acknowledgement Landesoffensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz (LOEWE), LOEWE-Zentrum für Translationale Medizin und Pharmakologie (JL) and European Union Seventh Framework Programme (FP7/2013, grant agreement no. 602919, JL, GLORIA).

References

1 Julius D, Basbaum AI. Molecular mechanisms of nociception. Nature. 2001;413(6852):203-10.

2 Verhaak PF, Kerssens JJ, Dekker J, Sorbi MJ, Bensing JM. Prevalence of chronic benign pain disorder among adults: a review of the literature. Pain. 1998;77(3):231-9.

3 Godinova AM. [Genetic analysis of migraine]. Zh Nevropatol Psikhiatr Im S S Korsakova. 1965;65(8):1132-8.

4 Lötsch J, Geisslinger G. Relevance of frequent mu-opioid receptor polymorphisms for opioid activity in healthy volunteers. Pharmacogenomics J. 2006;6(3):200-10.

5 Lötsch J, Geisslinger G, Heinemann S, Lerch F, Oertel BG, Ultsch A. QST response patterns to capsaicin- and UV-B-induced local skin hypersensitization in healthy subjects: a machine-learned analysis. Pain. 2017.

6 Kringel D, Geisslinger G, Resch E, Oertel BG, Thrun MC, Heinemann S, et al. Machine-learned analysis of the association of next-generation sequencing based human TRPV1 and TRPA1 genotypes with the sensitivity to heat stimuli and topically applied capsaicin. Pain. 2018(in revision).

Repetitive structures pass screening for ordinal relationships

Lisa M. Schäfer[1,2], Ludwig Lausser[1] and Hans A. Kestler[1]

Ordinal classifiers are constraint by an assumed ordinal relationship of class labels. The order in the label space should be reflected by the embedding of the data in the feature space. Depending on additional constraints about the layout of the decision regions various ordinal pattern can occur. As incorrect assumptions lead to a low performance these classifiers can be used to screen for potential class orders in the data [1]. This explorative analysis is guided by the hypothesis about the order. Incorrect or not-reflected class orders lead to low class-wise sensitivities. Orders with high class-wise sensitivities are a set of possible candidates.
However, it is not guaranteed that those candidates show an ordinal pattern representation in the feature space. By conducting cross-validation experiments with artificial datasets we could show that depending on the type of base classifiers the ordinal classifier proposes class orders that represent different patterns in the feature space and that some repetitive (non-ordinal) structures are not rejected by the ordinal classifier. We observe this effect for base classifiers with disconnected decision regions.

References

1 Lattke R, Lausser L, Müssel C, Kestler HA. Detecting ordinal class structures. In Schwenker F, Roli F, Kittler J, (eds), Multiple Classifier Systems (MCS 2015), volume LNCS 9132, pp. 100-111. Springer, 2015.

[1] Institute of Medical Systems Biology, Ulm University, 89081 Ulm, Germany
[2] International Graduate School in Molecular Medicine Ulm, Ulm University, 89081 Ulm, Germany

lisa-1.schaefer@uni-ulm.de

Computational Genomics: Informatics Problems and Applications in Life Sciences

Rolf Backofen[1]

Computational Genomics refers to the use of computational and statistical methods to decipher the information in genomic data. Especially the broad usage of high-throughput sequencing (HTS) produces large amounts of genomic data to be analyzed. I will talk about specific problems in computational genomics that were addressed in my groups, and will highlight both specific related computational tasks and important applications in life science. Topics will include sequence-structure alignment of RNA, the genome-wide detection of RNA-protein interactions, the analysis of single-cell sequencing data, and a workflow-management system based on Galaxy for the general analysis of HTS data. Our Galaxy server is one of the largest server in Europe with more than 600 users.

[1] Lehrstuhl für Bioinformatik, Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Georges-Köhler-Allee 106, D-79110 Freiburg, Germany

backofen@informatik.uni-freiburg.de

Mixed ordered-continuous copula GAMLSS with an application to poverty dimensions

Maike Hohberg[1], Francesco Donat[2,], Thomas Kneib[1] and Giampiero Marra[3]*

Although widely acknowledged as a multidimensional phenomenon, poverty is ubiquitously assessed using single number indices, such as the Multidimensional Poverty Index [1], inhibiting a thorough analysis of the potentially varying dependence between poverty dimensions. Multivariate regression is one tool to better acknowledge multidimensionality. To model the relations between poverty or welfare dimensions, which are subsumed into one response vector, copulas have been proven to be useful (e.g.[2]).
A second issue in poverty analysis concerns distributional aspects. Especially for program targeting and risk factor analysis, poverty analysis must move beyond mean effects. Examples include concepts such as inequality or vulnerability that take the scale parameter into account and can be analyzed using Generalized additive models for location, scale, and shape (GAMLSS, [3]) which estimate the effect of covariates on the whole conditional distribution of a poverty dimension.
We propose multivariate copula GAMLSS [4] to tackle both challenges. These models are implemented in the `R` package `gjrm` [5] and comprise a wide range of potential marginal distributions (continuous, binary, discrete) and copulas whose dependence parameter can also vary with covariates.
When looking at poverty, it often occurs that one dimension is measured in ordered categories whereas the other one is continuous. In this talk, we focus on two important poverty dimensions: expenditures as a continuous variable and health which is often assessed in categories such as "poor", "healthy", "very healthy". This combination of a continuous and an ordered categorical variable has not been treated within the copula GAMLSS context although this is a very relevant case in a lot of applications, especially in economics where many outcomes (health, education, subjective well-being, etc.) are measured in ordered categories.
We theoretically extend bivariate copula GAMLSS to the case where one response variable is measured in ordinal categories making use of a latent variable representation and relating the categories to an underlying continuous variable. In this way, we can follow the approach developed in [4], i.e. estimating the copula dependence and marginal distribution parameters simultaneously within a penalized likelihood framework using a trust region algorithm. The new model is integrated into the `gjrm` package and applied to Indonesian household data.

[1] Chair of Statistics, University of Goettingen, Humboldtallee 3, 37073 Goettingen
[2] European Central Bank, Sonnemannstrasse 20, 60314 Frankfurt am Main
[3] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT
[*] This paper should not be reported as representing the views of the European Central Bank. The views expressed are those of the authors and do not necessarily reflect those of the Bank.

`mhohber@uni-goettingen.de`

References

1 Alkire, S., Foster, J. E., Seth, S., Santos, M. E., Roche, J. M., and Ballón, P. (2015), *Multidimensional poverty measurement and analysis*. Oxford: Oxford University Press.

2 Decancq, K.(2014).Copula-based measurement of dependence between dimensions of well-being. *Oxford Economic Papers* **66**(3), 681 – 701.

3 Rigby, R.A. and Stasinopoulos, D.M. (2005), Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**, 507 – 554.

4 Marra, G., Radice, R. (2017), Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis* **112**, 99 – 113.

5 Marra, G., Radice, R. (2018), GJRM: Generalised Joint Regression Modelling. R package version 0.1-4.

`compboost`: A Modular Framework for Component-Wise Boosting in R

Daniel Schalk[1], *Janek Thomas*[1], *Bernd Bischl*[1]

Component-wise boosting applies the boosting framework to statistical models, e. g., general additive models using component-wise smoothing splines[1]. Boosting these kinds of models maintains interpretability and enables unbiased model selection in high dimensional feature spaces. A well-known implementation of this principle is the R package `mboost`[2].
The `R` package `compboost` is another implementation of component-wise boosting written in `C++` to obtain high runtime performance and full memory control. The main idea is to provide a modular class system which can be extended without editing the source code. Therefore, it is possible to use `R` functions as well as `C++` functions for custom base-learners, losses, logging mechanisms or stopping criteria.
Using those extensions, `compboost` can be configured, for instance, to take a performance measure of the `mlr` package, or any other package, as logger and for early stopping. These customizations apply not just for logging, but also for creating new base-learners or losses allowing `compboost` to be highly flexible and extendible.

References

1 Schmid, M., & Hothorn, T. (2008). *Boosting additive models using component-wise P-splines*. Computational Statistics & Data Analysis, 53(2), 298-311.

2 Hofner, B., Hothorn, T., Kneib, T., & Schmid, M. (2011). *A framework for unbiased model selection based on boosting*. Journal of Computational and Graphical Statistics, 20(4), 956-971.

[1] LMU Munich, Ludwigstr. 33, 80539 Munich

`d.schalk@mail.de, janek.thomas@stat.uni-muenchen.de, bernd_bischl@stat.uni-muenchen.de`

A Discussion on Adaptive Step-lengths in Gradient Boosting

Elisabeth Waldmann[1], *Andreas Mayr*[2]

Gradient boosting methods are iterative updating schemes, in which the gradient of the loss-function in the current step is fitted to the data. The best fitting variable is then selected and updated by a fracture of the parameter suggested. There has been a lot of evidence for gradient boosting algorithms working best, if the main tuning parameter is the number of iterations rather than the step-length [1]. The latter is hence traditionally fixed to $\nu = 0.1$. Those findings however are based on models with one outcome and problems arise when trying to fit models with more dimensions, such as in generalized additive models for location, shape and scale [2] or joint models for longitudinal and time-to-event data [3]. In the first case a whole set of parameters is modelled rather than only the mean. In the second case, a longitudinal outcome shares part of its predictor with a survival outcome. The resulting joint likelihood, which is used as the loss-function, is a product of two very different likelihoods: a Gaussian distribution and a Cox-type likelihood. The resulting gradients are very different for the two different outcomes. Simulations have shown that adaptive step-lengths, which mitigate the difference in scale, lead to more stable results than simply using a fixed step-length of $\nu = 0.1$.
A further interesting point of the idea of adaptive step-lengths is using them as convergence criterion, where a model is regarded as converged if the optimal step-length is $\nu = 0$.

References

1 Matthias Schmid and Torsten Hothorn. (2008) Boosting additive models using component-wise P-splines as base-learners. *Computational Statistics & Data Analysis*, 53(2):298–311.

2 Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012): Generalized additive models for location scale and shape for high-dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society Series C - Applied Statistics*. 61(3): 403-427.

3 Waldmann, E., Griesbach, C. and Mayr, A. (2018) Variable Selection and Allocation in Joint Models for Longitudinal and Time-to-Event Data via Boosting *in preparation*

[1] Department of Medical Informatics, Biometrics and Epidemilology, Friedrich-Alexander-Universität Erlangen-Nürnberg
[2] Department of Medical Biometrics, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität Bonn

elisabeth.waldmann@fau.de, amayr@uni-bonn.de

Multimodal likelihood functions occurring in mixture modeling

Malte Jastrow[1], *Claus Weihs*[1]

A popular framework for black-box optimization benchmarking (BBOB) is given by the so-called BBOB functions[1]. These functions are created to confront optimization algorithms with a broad range of difficulties, especially multimodality[2]. Due to the artificial nature of these test functions a practical relevance for data analysis tasks is not directly given. The initial goal that lead to this talk was to derive alternative test functions directly from statistical applications.
A common optimization problem in statistics is parameter estimation by maximizing the likelihood function. In some cases analytical solutions exist (e.g. linear models), but there are other applications, like mixture distribution models, where the use of non-linear optimization techniques is inevitable. It is even known, that likelihood functions of mixture models can be multimodal[3], which makes them ideal candidates for alternative test functions.
In this talk likelihood functions of normal mixtures are presented and investigated. In the simple case of two equally-weighted univariate normals with known variances, the likelihood is observed to be (highly) multimodal, if a considerable difference between the two distributions' variances exists. It is analyzed how the attraction area size of the global optimum and the number of local optima can be controlled during function generation.
The highly favored optimization method for mixture models, expectation-maximization[4], does not guarantee to discover the global optimum on those highly multimodal functions. In a simulation study the optimization performance of the EM-algorithm is compared to other common nonlinear optimization techniques.
Further aspects, beyond this talk, are the expansion to mixing more than two distributions, multivariate distributions and non-normal distributions. Particularly the question, whether the presented multimodal behavior can only be detected in mixtures of normals, is of high interest.

[1] Computational Statistics, TU Dortmund University, 44221 Dortmund, Germany

jastrow@statistik.tu-dortmund.de, weihs@statistik.uni-dortmund.de

References

1 Hansen, N., et al. *Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions.* Diss. INRIA, 2009.

2 Mersmann, O., et al. *Exploratory landscape analysis.* Proceedings of the 13th annual conference on Genetic and evolutionary computation. ACM, 2011.

3 Jin, C., et al. *Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences.* Advances in Neural Information Processing Systems. 2016.

4 McLachlan, Geoffrey, and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Randomized Stepwise Regression – a Modification of Stepwise Methods in Generalized Linear Models and its Application on Sepsis Data

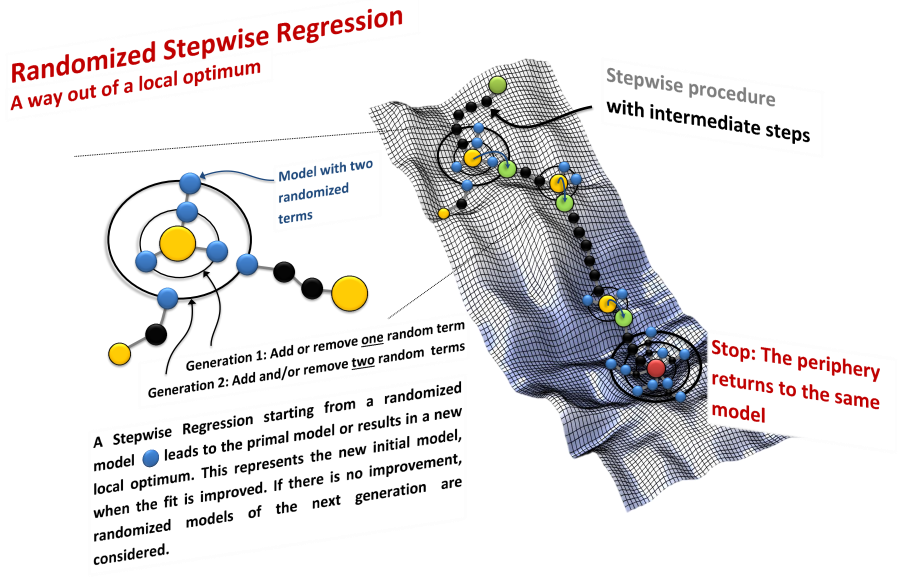
Marcus Vollmer[1]

Background:
Generalized linear models are widely used in medical statistics for predicting the outcome of a response variable, e.g. survival, progression, number of tumors or length of hospital stay using predictor variables [1,2]. To decide which predictor set provides the best model the all-subset method is appropriate [3]. However, all-subset method can easily fail due to combinatorial explosion if many covariates and interaction terms are involved. Often the multivariable search for relevant predictors is performed using statistically significant univariate variables or greedy hierarchical procedures, like the stepwise logistic regression with backward-elimination and forward-selection. It has been shown that stepwise procedures have serious drawbacks, like inflation of Type I error rates, instability of the selection, and biased estimation of coefficients [2,4]. Beside this, the methods result only in local optima when optimizing the model selection criterion, such as the Bayesian Information Criterion (BIC).
Aim:
This motivates us to address the problem of finding the global optimized model by adding randomized steps within the stepwise methodology.
Methods & Results:
Our randomized strategy starts with a random initial model. The standard stepwise approach will be conducted by adding or removing one predictor (1^{st} generation) to improve the model gradually. Once the usual procedure stops at a local optimum, we will try to continue by alter the model formula randomly. We add or remove two predictors at the same time (2^{nd} generation) and start the standard approach again. The best of the local optima forms the new starting point. The procedure ends, if even a k-th generation mutation will lead to the same model. To rate the performance of the new selection strategy, we investigate predictors of 90-day survival of 878 surgical patients with septic shock or severe sepsis between 2006 and midyear 2013 treated at the University Hospital Greifswald [5]. Stepwise logistic regression with different initial models were compared to the randomized approach and LASSO [6]. All procedures were optimized according to BIC.

[1] Institute of Bioinformatics, Universtity Medicine Greifswald

marcus.vollmer@uni-greifswald.de

Figure 1: Illustration of the stepwise regression with randomized steps. The highland structure is representing the BIC values from the model space. Color codes: yellow circles represents local optima, black circles as intermediate steps, blue circles as randomized models, green circles as new starting models, red circle as the final model.



References

1 F. Harrell, *Regression Modeling Strategies*. Springer, corrected ed., Jan. 2001.

2 E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health)*. Springer, 1st ed., Nov. 2008.

3 J. Fox, *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, 2008.

4 R. Mundry and C. L. Nunn, "Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution.," *The American Naturalist*, vol. 173, no. 1, pp. 119–123, 2009. PMID: 19049440.

5 C. S. Scheer, C. Fuchs, S.-O. Kuhn, M. Vollmer, S. Rehberg, S. Friesecke, P. Abel, V. Balau, C. Bandt, K. Meissner, et al., "Quality improvement initiative for severe sepsis and septic shock reduces 90-day mortality: a 7.5-year observational study," *Critical care medicine*, vol. 45, no. 2, pp. 241–252, 2017.

6 R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological), pp. 267–288, 1996.

Analysing unmeasured baseline covariates in studies with delayed entry using a joint model: fitting the model in presence of left-truncation

Regina Stegherr[1],Jan Beyersmann[1],Peter Bramlage[2] and Tobias Bluhmki[1]

The natural choice for 'time zero' (baseline) in a randomized clinical trial is study entry, in particular, covariate information is available at study entry. In other observational studies study entry may happen after the time origin leading to left-truncated data. One specific example is the analysis of diabetes register-based data, where a relevant timescale is 'time-since-first-antidiabetic-medication'. Since such data is collected in calendar time, some patients enter the study upon their first medication, but others have a known date of therapy initiation before start of data collection. A relevant baseline covariate would be glycated haemoglobin (HbA1c) levels in diabetes patients. The challenge is that such data is typically measured upon study entry and, hence, not at baseline for the left-truncated patients, but, e.g., HbA1c will have changed in the random and patient-specific time interval between start of medication and study entry. The problem has been summarized in a letter by Keiding and Knuiman [1]. We propose a joint model to investigate the impact of a baseline covariate, possibly unmeasured due to delayed entry, on the time to the event of interest. This contrasts with the standard use of joint models where the aim is to analyze the effect of the current value of the covariate on the hazard of an event. Thereby, parameter estimation bears several challenges as for maximization of the likelihood integrals over random effects need to be approximated using Gauss-Hermite quadrature rules. The implementation of our approach was done in the statistic software R by adding a second set of pseudo-adaptive Gauss-Hermite quadrature rules to the original jointModel function [2] to account for left-truncation. Our approach shows proper performance in a simulation study and was applied to data from a German diabetes register to evaluate the effect of HbA1c at therapy initiation on the risk of treatment failure.

References

1 Keiding N and Knuiman M (1990), Letter to the editor in survival analysis in natural history studies of disease. Statistics in Medicine, 9(10):1221-1222.

2 Rizopoulos D (2010). JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. Journal of Statistical Software, 35, 1–33.

[1] Insitute of Statistics, Ulm University, Helmholtzstraße 20, 89081 Ulm, Germany
[2] Intitute of Pharmacology and Preventive Medicine, Menzelstraße 21, 15831 Mahlow, Germany

regina.stegherr@uni-ulm.de, jan.beyersmann@uni-ulm.de, peter.bramlage@ippmed.de, tobias.bluhmki@uni-ulm.de

Tensile Stress Monitoring of Fibre-Reinforced Plastic

Sebastian Krey[1], Thomas Bartz-Beielstein[1] and Patrick Franck[1]

Fibre-reinforced plastic (FRP) and other composite materials gain more and more usage in important industries like automotive, aerospace, or renewable energy. Today the wings of wind generators, large parts of newly constructed aircrafts and weight critical components of cars are built from different types of fibre-reinforced plastic. While the structure of these materials permits extremely lightweight design of constructional elements, there are currently no methods for an easy health monitoring of these complex materials or for the prediction of the remaining service life. Regular downtimes for resource intensive manual inspections are the consequences.
Structural Health Monitoring methods for FRP components allow a cost efficient and secure usage of these materials. The availability of always up to date information about the condition of the different components permits planning the downtimes for maintenance work as needed instead of fixed service intervals. The usage of ultrasonic waves induced and measured with piezoelectric elements is one option for the inspection of these materials. The propagation of ultrasonic waves through the FRP and the distinct signal changes caused by changing material characteristics can be used for the detection of damages in the structure.
The consequences of tensile stress are especially difficult to detect manually. Even small elongations can result in severe damages within the composite, which are often not visible.
In this work we present how tensile stress changes the signals detected by the piezoelectric sensors and how this can be used to detect damages within the composite long before the structure completely fails.
We also compare the results obtained by externally applied piezoelectric elements with sensors placed between the layers of glass-fibre reinforced plastic using our self developed method. Especially the major concern of industrial users regarding the sturdiness of composites with implemented sensors will be discussed.

[1] Technische Hochschule Köln, Fakultät für Informatik und Ingenieurwissenschaften, Steinmüllerallee 1, 51643 Gummersbach

sebastian.krey@th-koeln.de, thomas.bartz-beielstein@th-koeln.de

Dynamic reference intervals from contaminated data sources

Tobias Hepp[1], Jacob Zierk[2], Andreas Mayr[3]

In medical practice, reference limits (RLs) are required to compare the values of specific biomarkers measured on patients with their distribution in the healthy population. While these intervals could be directly estimated based on voluntary study participants, many laboratories will not be able to do so due to the expenses involved. Since simply adopting RLs from external sources rises the question of transferability, so-called indirect approaches have been suggested to determine reference limits from already available intra-laboratory data bases [1]. As a consequence of these data being comprised of all kinds of patients with no specific focus on the biomarker of interest, they are therefore contaminated with pathological values. While the current strategies based on truncated power-normal-distributions work well for static distributions, they are rather cumbersome to use if the reference distribution dynamically depends on other patient characteristics such as age [2]. Therefore, we suggest an alternative approach based on a mixture of generalized additive models for location scale and shape (GAMLSS) [3]. Despite the high flexibility of the resulting procedure, our preliminary simulation results already provide meaningful RLs under proper specification.

References

1 Arzideh, et al. (2009). An improved indirect approach for determining reference limits from intra-laboratory data bases exemplified by concentrations of electrolytes. *Laboratoriumsmedizin.* 33(2), 52-66.

2 Zierk et al. (2013). Indirect determination of pediatric blood count reference intervals. *Clinical Chemistry and Laboratory Medicine* 51(4), 863-872

3 Rigby and Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics).* 54(3), 507-554.

[1] IMBIE, Universitätsklinikum Bonn
[2] Kinder- und Jugendklinik, Universitätsklinikum Erlangen
[3] Institut für Statistik, Ludwig-Maximilians-Universität München

tobias.hepp@uk-erlangen.de

Selection of Optimal Subgroup Weights for Survival Analysis

Jakob Richter[1], *Katrin Madjar*[1], *Jörg Rahnenführer*[1]

Survival analysis is a central aspect in cancer research with the aim of predicting the survival time of a patient on the basis of his features as precisely as possible. Often it can be assumed that the relation between covariables and the survival time is not equal across different subgroups of patients (e.g. cohorts from different hospitals). The aim is to predict the survival function for a specific subgroup based on data of all subgroups to benefit from the larger sample size. In standard subgroup analysis only the patients of the subgroup of interest are included in the model. This might lead to instable results, especially for smaller subgroups. As an alternative we propose a model that uses all patients but assigns them individual weights based on their subgroup affiliation.
Weyer and Binder [2] aim at improving the stability and prediction quality for a specific subgroup by including an additional weighted subgroup. They study the effects for a set of fixed weights for the additional subgroup in a stratified Cox model on the stability and performance. In our approach we use multiple additional subgroups to improve the prediction quality of the Cox model. The optimal subgroup weights are determined by optimizing the cross-validated Concordance index through Bayesian optimization [1]. In an adapted version of the cross-validation only the patients of the subgroup of interest are included in the test while all subgroups are used for training. This ensures that we include subgroups that lead to a high predictive accuracy in our subgroup by assigning a high weight to them. Subgroups that deteriorate the predictive performance (e.g. due to different covariable survival time relationship) will be assigned lower weights. We will show that with our subgroup weights optimization approach, we can improve the predictive performance over the naive approaches, of including all related subgroups or excluding them. As an application example we use ten non-small cell lung cancer studies as subgroups and optimize the prediction quality for each subgroup using all other subgroups with optimized weights.

References

1 D. R. Jones. A Taxonomy of Global Optimization Methods Based on Response Surfaces. 21(4): 345–383. doi: 10.1023/A:1012771025575.

2 V. Weyer and H. Binder. A weighting approach for judging the effect of patient strata on high-dimensional risk prediction signatures. 16:294. doi: 10.1186/s12859-015-0716-8.

[1] TU Dortmund University, Department of Statistics

richter@statistik.tu-dortmund.de, madjar@statistik.tu-dortmund.de,
rahnenfuehrer@statistik.tu-dortmund.de

Correlation-Adjusted Regression Survival Scores for High-Dimensional Variable Selection

Thomas Welchowski[1], Verena Zuber[2] and Matthias Schmid[1]

Abstract

Background
The development of classification methods for personalized medicine is highly dependent on the identification of predictive genetic markers. In survival analysis it is often necessary to discriminate between influential and non-influential markers. It is common to perform univariate screening using Cox scores, which quantify the associations between survival and each of the markers to provide a ranking. Since Cox scores do not account for dependencies between the markers, their use is suboptimal in the presence highly correlated markers.
Methods
As an alternative to the Cox score, we propose the correlation-adjusted regression survival (CARS) score for right-censored survival outcomes. By removing the correlations between the markers, the CARS score quantifies the associations between the outcome and the set of "de-correlated" marker values. Estimation of the scores is based on inverse probability weighting, which is applied to log-transformed event times. For high-dimensional data, estimation is based on shrinkage techniques.
Results
The consistency of the CARS score is proven under mild regularity conditions. In simulations, survival models based on CARS score rankings achieved higher areas under the precision-recall curve than competing methods. Two example applications on prostate and breast cancer confirmed these results. CARS scores are implemented in the *R* package *carSurv*.
Conclusions
In research applications involving high-dimensional genetic data, the use of CARS scores for marker selection is a favorable alternative to Cox scores even when correlations between covariates are low. Having a straightforward interpretation and low computational requirements, CARS scores are an easy-to-use screening tool in personalized medicine research.

Keywords: Biomarker discovery; Breast cancer; Multi-gene signature; Personalized medicine; Prostate cancer; Survival modeling

[1] Department of Medical Biometry, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität Bonn, Sigmund-Freud-Str. 25, 53127 Bonn,Germany
[2] MRC Biostatistics Unit, Cambridge University, Cambridgeshire, United Kingdom

welchow@imbie.meb.uni-bonn.de, verena.zuber@mrc-bsu.cam.ac.uk, matthias.schmid@imbie.uni-bonn.de

The distance correlation coefficient for right-censored survival data

Dominic Edelmann[1] *and Axel Benner*[1]

Székely, Rizzo, and Bakirov introduced the powerful concept of distance correlation as a measure of dependence between random variables. In contrast to Pearson correlation, which only measures linear dependence, distance correlation can detect any kind of dependence including nonlinear or even nonmonotone associations. In biomedical applications, one is interested in finding associations between patient characteristics (e.g. the expression of a certain gene) and survival time. This kind of data usually involves right-censoring, implying that the exact survival time cannot be observed for individuals who leave the study before they decease. Under this missingness of information, it is not straightforward to derive consistent estimates for the distance correlation. In this talk, we present a suitable distance correlation coefficient for right-censored survival data accounting for censored observations via inverse probability of censoring weights (IPCW). In particular, we show that the population distance correlation between covariate data and survival times can be approximated by an IPC-weighted U-statistic. Using a simulation study, we demonstrate that this coefficient is helpful to reveal associations which can not be found with standard methods. Finally, we apply this coefficient on variable screening for ultrahigh dimensional DNA methylation data.

[1] German Cancer Research Center, Heidelberg Germany

dominic.edelmann@dkfz.de, benner@dkfz.de

List of technical reports published by the University of Ulm

Some of them are available by FTP from ftp.informatik.uni-ulm.de
*Reports marked with * are out of print*

91-01 Ker-I Ko, P. Orponen, U. Schöning, O. Watanabe
Instance Complexity

*91-02** K. Gladitz, H. Fassbender, H. Vogler
Compiler-Based Implementation of Syntax-Directed Functional Programming

*91-03** Alfons Geser
Relative Termination

*91-04** J. Köbler, U. Schöning, J. Toran
Graph Isomorphism is low for PP

91-05 Johannes Köbler, Thomas Thierauf
Complexity Restricted Advice Functions

*91-06** Uwe Schöning
Recent Highlights in Structural Complexity Theory

*91-07** F. Green, J. Köbler, J. Toran
The Power of Middle Bit

*91-08** V.Arvind, Y. Han, L. Hamachandra, J. Köbler, A. Lozano, M. Mundhenk, A. Ogiwara, U. Schöning, R. Silvestri, T. Thierauf
Reductions for Sets of Low Information Content

*92-01** Vikraman Arvind, Johannes Köbler, Martin Mundhenk
On Bounded Truth-Table and Conjunctive Reductions to Sparse and Tally Sets

*92-02** Thomas Noll, Heiko Vogler
Top-down Parsing with Simulataneous Evaluation of Noncircular Attribute Grammars

92-03 Fakultät für Informatik
17. Workshop über Komplexitätstheorie, effiziente Algorithmen und Datenstrukturen

*92-04** V. Arvind, J. Köbler, M. Mundhenk
Lowness and the Complexity of Sparse and Tally Descriptions

*92-05** Johannes Köbler
Locating P/poly Optimally in the Extended Low Hierarchy

*92-06** Armin Kühnemann, Heiko Vogler
Synthesized and inherited functions -a new computational model for syntax-directed semantics

*92-07** Heinz Fassbender, Heiko Vogler
A Universal Unification Algorithm Based on Unification-Driven Leftmost Outermost Narrowing

*92-08** Uwe Schöning
On Random Reductions from Sparse Sets to Tally Sets

*92-09** Hermann von Hasseln, Laura Martignon
Consistency in Stochastic Network

92-10 Michael Schmitt
A Slightly Improved Upper Bound on the Size of Weights Sufficient to Represent Any Linearly Separable Boolean Function

92-11 Johannes Köbler, Seinosuke Toda
On the Power of Generalized MOD-Classes

92-12 V. Arvind, J. Köbler, M. Mundhenk
Reliable Reductions, High Sets and Low Sets

92-13 Alfons Geser
On a monotonic semantic path ordering

*92-14** Joost Engelfriet, Heiko Vogler
The Translation Power of Top-Down Tree-To-Graph Transducers

93-01 Alfred Lupper, Konrad Froitzheim
AppleTalk Link Access Protocol basierend auf dem Abstract Personal Communications Manager

93-02 M.H. Scholl, C. Laasch, C. Rich, H.-J. Schek, M. Tresch
The COCOON Object Model

93-03 Thomas Thierauf, Seinosuke Toda, Osamu Watanabe
On Sets Bounded Truth-Table Reducible to P-selective Sets

93-04 Jin-Yi Cai, Frederic Green, Thomas Thierauf
On the Correlation of Symmetric Functions

93-05 K.Kuhn, M.Reichert, M. Nathe, T. Beuter, C. Heinlein, P. Dadam
A Conceptual Approach to an Open Hospital Information System

93-06 Klaus Gaßner
Rechnerunterstützung für die konzeptuelle Modellierung

93-07 Ullrich Keßler, Peter Dadam
Towards Customizable, Flexible Storage Structures for Complex Objects

94-01 Michael Schmitt
On the Complexity of Consistency Problems for Neurons with Binary Weights

94-02 Armin Kühnemann, Heiko Vogler
A Pumping Lemma for Output Languages of Attributed Tree Transducers

94-03 Harry Buhrman, Jim Kadin, Thomas Thierauf
On Functions Computable with Nonadaptive Queries to NP

94-04 Heinz Faßbender, Heiko Vogler, Andrea Wedel
Implementation of a Deterministic Partial E-Unification Algorithm for Macro Tree Transducers

94-05 V. Arvind, J. Köbler, R. Schuler
On Helping and Interactive Proof Systems

94-06 Christian Kalus, Peter Dadam
Incorporating record subtyping into a relational data model

94-07 Markus Tresch, Marc H. Scholl
A Classification of Multi-Database Languages

94-08 Friedrich von Henke, Harald Rueß
Arbeitstreffen Typtheorie: Zusammenfassung der Beiträge

94-09 F.W. von Henke, A. Dold, H. Rueß, D. Schwier, M. Strecker
Construction and Deduction Methods for the Formal Development of Software

94-10 Axel Dold
Formalisierung schematischer Algorithmen

94-11 Johannes Köbler, Osamu Watanabe
New Collapse Consequences of NP Having Small Circuits

94-12 Rainer Schuler
On Average Polynomial Time

94-13 Rainer Schuler, Osamu Watanabe
Towards Average-Case Complexity Analysis of NP Optimization Problems

94-14 Wolfram Schulte, Ton Vullinghs
Linking Reactive Software to the X-Window System

94-15 Alfred Lupper
Namensverwaltung und Adressierung in Distributed Shared Memory-Systemen

94-16 Robert Regn
Verteilte Unix-Betriebssysteme

94-17 Helmuth Partsch
Again on Recognition and Parsing of Context-Free Grammars: Two Exercises in Transformational Programming

94-18 Helmuth Partsch
Transformational Development of Data-Parallel Algorithms: an Example

95-01 Oleg Verbitsky
On the Largest Common Subgraph Problem

95-02 Uwe Schöning
Complexity of Presburger Arithmetic with Fixed Quantifier Dimension

95-03 Harry Buhrman,Thomas Thierauf
The Complexity of Generating and Checking Proofs of Membership

95-04 Rainer Schuler, Tomoyuki Yamakami
Structural Average Case Complexity

95-05 Klaus Achatz, Wolfram Schulte
Architecture Indepentent Massive Parallelization of Divide-And-Conquer Algorithms

95-06 Christoph Karg, Rainer Schuler
Structure in Average Case Complexity

95-07 P. Dadam, K. Kuhn, M. Reichert, T. Beuter, M. Nathe
ADEPT: Ein integrierender Ansatz zur Entwicklung flexibler, zuverlässiger kooperierender Assistenzsysteme in klinischen Anwendungsumgebungen

95-08 Jürgen Kehrer, Peter Schulthess
Aufbereitung von gescannten Röntgenbildern zur filmlosen Diagnostik

95-09 Hans-Jörg Burtschick, Wolfgang Lindner
On Sets Turing Reducible to P-Selective Sets

95-10 Boris Hartmann
Berücksichtigung lokaler Randbedingung bei globaler Zieloptimierung mit neuronalen Netzen am Beispiel Truck Backer-Upper

95-11 Thomas Beuter, Peter Dadam
Prinzipien der Replikationskontrolle in verteilten Systemen

95-12 Klaus Achatz, Wolfram Schulte
Massive Parallelization of Divide-and-Conquer Algorithms over Powerlists

95-13 Andrea Mößle, Heiko Vogler
Efficient Call-by-value Evaluation Strategy of Primitive Recursive Program Schemes

95-14 Axel Dold, Friedrich W. von Henke, Holger Pfeifer, Harald Rueß
A Generic Specification for Verifying Peephole Optimizations

96-01 Ercüment Canver, Jan-Tecker Gayen, Adam Moik
Formale Entwicklung der Steuerungssoftware für eine elektrisch ortsbediente Weiche mit VSE

96-02 Bernhard Nebel
Solving Hard Qualitative Temporal Reasoning Problems: Evaluating the Efficiency of Using the ORD-Horn Class

96-03 Ton Vullinghs, Wolfram Schulte, Thilo Schwinn
An Introduction to TkGofer

96-04 Thomas Beuter, Peter Dadam
Anwendungsspezifische Anforderungen an Workflow-Mangement-Systeme am Beispiel der Domäne Concurrent-Engineering

96-05 Gerhard Schellhorn, Wolfgang Ahrendt
Verification of a Prolog Compiler - First Steps with KIV

96-06 Manindra Agrawal, Thomas Thierauf
Satisfiability Problems

96-07 Vikraman Arvind, Jacobo Torán
A nonadaptive NC Checker for Permutation Group Intersection

96-08 David Cyrluk, Oliver Möller, Harald Rueß
An Efficient Decision Procedure for a Theory of Fix-Sized Bitvectors with Composition and Extraction

96-09 Bernd Biechele, Dietmar Ernst, Frank Houdek, Joachim Schmid, Wolfram Schulte
Erfahrungen bei der Modellierung eingebetteter Systeme mit verschiedenen SA/RT–Ansätzen

96-10 Falk Bartels, Axel Dold, Friedrich W. von Henke, Holger Pfeifer, Harald Rueß
Formalizing Fixed-Point Theory in PVS

96-11 Axel Dold, Friedrich W. von Henke, Holger Pfeifer, Harald Rueß
Mechanized Semantics of Simple Imperative Programming Constructs

96-12 Axel Dold, Friedrich W. von Henke, Holger Pfeifer, Harald Rueß
Generic Compilation Schemes for Simple Programming Constructs

96-13 Klaus Achatz, Helmuth Partsch
From Descriptive Specifications to Operational ones: A Powerful Transformation Rule, its Applications and Variants

97-01 Jochen Messner
Pattern Matching in Trace Monoids

97-02 Wolfgang Lindner, Rainer Schuler
A Small Span Theorem within P

97-03 Thomas Bauer, Peter Dadam
A Distributed Execution Environment for Large-Scale Workflow Management Systems with Subnets and Server Migration

97-04 Christian Heinlein, Peter Dadam
Interaction Expressions - A Powerful Formalism for Describing Inter-Workflow Dependencies

97-05 Vikraman Arvind, Johannes Köbler
On Pseudorandomness and Resource-Bounded Measure

97-06 Gerhard Partsch
Punkt-zu-Punkt- und Mehrpunkt-basierende LAN-Integrationsstrategien für den digitalen Mobilfunkstandard DECT

97-07 Manfred Reichert, Peter Dadam
$ADEPT_{flex}$ - Supporting Dynamic Changes of Workflows Without Loosing Control

97-08 Hans Braxmeier, Dietmar Ernst, Andrea Mößle, Heiko Vogler
The Project NoName - A functional programming language with its development environment

97-09 Christian Heinlein
Grundlagen von Interaktionsausdrücken

97-10 Christian Heinlein
Graphische Repräsentation von Interaktionsausdrücken

97-11 Christian Heinlein
Sprachtheoretische Semantik von Interaktionsausdrücken

97-12 Gerhard Schellhorn, Wolfgang Reif
Proving Properties of Finite Enumerations: A Problem Set for Automated Theorem Provers

97-13 Dietmar Ernst, Frank Houdek, Wolfram Schulte, Thilo Schwinn
Experimenteller Vergleich statischer und dynamischer Softwareprüfung für eingebettete Systeme

97-14 Wolfgang Reif, Gerhard Schellhorn
Theorem Proving in Large Theories

97-15 Thomas Wennekers
Asymptotik rekurrenter neuronaler Netze mit zufälligen Kopplungen

97-16 Peter Dadam, Klaus Kuhn, Manfred Reichert
Clinical Workflows - The Killer Application for Process-oriented Information Systems?

97-17 Mohammad Ali Livani, Jörg Kaiser
EDF Consensus on CAN Bus Access in Dynamic Real-Time Applications

97-18 Johannes Köbler,Rainer Schuler
Using Efficient Average-Case Algorithms to Collapse Worst-Case Complexity Classes

98-01 Daniela Damm, Lutz Claes, Friedrich W. von Henke, Alexander Seitz, Adelinde Uhrmacher, Steffen Wolf
Ein fallbasiertes System für die Interpretation von Literatur zur Knochenheilung

98-02 Thomas Bauer, Peter Dadam
Architekturen für skalierbare Workflow-Management-Systeme - Klassifikation und Analyse

98-03 Marko Luther, Martin Strecker
A guided tour through Typelab

98-04 Heiko Neumann, Luiz Pessoa
Visual Filling-in and Surface Property Reconstruction

98-05 Ercüment Canver
Formal Verification of a Coordinated Atomic Action Based Design

98-06 Andreas Küchler
On the Correspondence between Neural Folding Architectures and Tree Automata

98-07 Heiko Neumann, Thorsten Hansen, Luiz Pessoa
Interaction of ON and OFF Pathways for Visual Contrast Measurement

98-08 Thomas Wennekers
Synfire Graphs: From Spike Patterns to Automata of Spiking Neurons

98-09 Thomas Bauer, Peter Dadam
Variable Migration von Workflows in ADEPT

98-10 Heiko Neumann, Wolfgang Sepp
Recurrent V1 – V2 Interaction in Early Visual Boundary Processing

98-11 Frank Houdek, Dietmar Ernst, Thilo Schwinn
Prüfen von C–Code und Statmate/Matlab–Spezifikationen: Ein Experiment

98-12 Gerhard Schellhorn
Proving Properties of Directed Graphs: A Problem Set for Automated Theorem Provers

98-13 Gerhard Schellhorn, Wolfgang Reif
Theorems from Compiler Verification: A Problem Set for Automated Theorem Provers

98-14 Mohammad Ali Livani
SHARE: A Transparent Mechanism for Reliable Broadcast Delivery in CAN

98-15 Mohammad Ali Livani, Jörg Kaiser
Predictable Atomic Multicast in the Controller Area Network (CAN)

99-01 Susanne Boll, Wolfgang Klas, Utz Westermann
A Comparison of Multimedia Document Models Concerning Advanced Requirements

99-02 Thomas Bauer, Peter Dadam
Verteilungsmodelle für Workflow-Management-Systeme - Klassifikation und Simulation

99-03 Uwe Schöning
On the Complexity of Constraint Satisfaction

99-04 Ercument Canver
Model-Checking zur Analyse von Message Sequence Charts über Statecharts

99-05 Johannes Köbler, Wolfgang Lindner, Rainer Schuler
Derandomizing RP if Boolean Circuits are not Learnable

99-06 Utz Westermann, Wolfgang Klas
Architecture of a DataBlade Module for the Integrated Management of Multimedia Assets

99-07 Peter Dadam, Manfred Reichert
Enterprise-wide and Cross-enterprise Workflow Management: Concepts, Systems, Applications. Paderborn, Germany, October 6, 1999, GI–Workshop Proceedings, Informatik '99

99-08 Vikraman Arvind, Johannes Köbler
Graph Isomorphism is Low for ZPP^{NP} and other Lowness results

99-09 Thomas Bauer, Peter Dadam
Efficient Distributed Workflow Management Based on Variable Server Assignments

2000-02 Thomas Bauer, Peter Dadam
Variable Serverzuordnungen und komplexe Bearbeiterzuordnungen im Workflow-Management-System ADEPT

2000-03 Gregory Baratoff, Christian Toepfer, Heiko Neumann
Combined space-variant maps for optical flow based navigation

2000-04 Wolfgang Gehring
Ein Rahmenwerk zur Einführung von Leistungspunktsystemen

2000-05 Susanne Boll, Christian Heinlein, Wolfgang Klas, Jochen Wandel
Intelligent Prefetching and Buffering for Interactive Streaming of MPEG Videos

2000-06 Wolfgang Reif, Gerhard Schellhorn, Andreas Thums
Fehlersuche in Formalen Spezifikationen

2000-07 Gerhard Schellhorn, Wolfgang Reif (eds.)
FM-Tools 2000: The 4th Workshop on Tools for System Design and Verification

2000-08 Thomas Bauer, Manfred Reichert, Peter Dadam
Effiziente Durchführung von Prozessmigrationen in verteilten Workflow-Management-Systemen

2000-09 Thomas Bauer, Peter Dadam
Vermeidung von Überlastsituationen durch Replikation von Workflow-Servern in ADEPT

2000-10 Thomas Bauer, Manfred Reichert, Peter Dadam
Adaptives und verteiltes Workflow-Management

2000-11 Christian Heinlein
Workflow and Process Synchronization with Interaction Expressions and Graphs

2001-01 Hubert Hug, Rainer Schuler
DNA-based parallel computation of simple arithmetic

2001-02 Friedhelm Schwenker, Hans A. Kestler, Günther Palm
3-D Visual Object Classification with Hierarchical Radial Basis Function Networks

2001-03 Hans A. Kestler, Friedhelm Schwenker, Günther Palm
RBF network classification of ECGs as a potential marker for sudden cardiac death

2001-04 Christian Dietrich, Friedhelm Schwenker, Klaus Riede, Günther Palm
Classification of Bioacoustic Time Series Utilizing Pulse Detection, Time and Frequency Features and Data Fusion

2002-01 Stefanie Rinderle, Manfred Reichert, Peter Dadam
Effiziente Verträglichkeitsprüfung und automatische Migration von Workflow-Instanzen bei der Evolution von Workflow-Schemata

2002-02 Walter Guttmann
Deriving an Applicative Heapsort Algorithm

2002-03 Axel Dold, Friedrich W. von Henke, Vincent Vialard, Wolfgang Goerigk
A Mechanically Verified Compiling Specification for a Realistic Compiler

2003-01 Manfred Reichert, Stefanie Rinderle, Peter Dadam
A Formal Framework for Workflow Type and Instance Changes Under Correctness Checks

2003-02 Stefanie Rinderle, Manfred Reichert, Peter Dadam
Supporting Workflow Schema Evolution By Efficient Compliance Checks

2003-03 Christian Heinlein
Safely Extending Procedure Types to Allow Nested Procedures as Values

2003-04 Stefanie Rinderle, Manfred Reichert, Peter Dadam
On Dealing With Semantically Conflicting Business Process Changes.

2003-05 Christian Heinlein
Dynamic Class Methods in Java

2003-06 Christian Heinlein
Vertical, Horizontal, and Behavioural Extensibility of Software Systems

2003-07 Christian Heinlein
Safely Extending Procedure Types to Allow Nested Procedures as Values (Corrected Version)

2003-08 Changling Liu, Jörg Kaiser
Survey of Mobile Ad Hoc Network Routing Protocols)

2004-01 Thom Frühwirth, Marc Meister (eds.)
First Workshop on Constraint Handling Rules

2004-02 Christian Heinlein
Concept and Implementation of C+++, an Extension of C++ to Support User-Defined Operator Symbols and Control Structures

2004-03 Susanne Biundo, Thom Frühwirth, Günther Palm(eds.)
Poster Proceedings of the 27th Annual German Conference on Artificial Intelligence

2005-01 Armin Wolf, Thom Frühwirth, Marc Meister (eds.)
19th Workshop on (Constraint) Logic Programming

2005-02 Wolfgang Lindner (Hg.), Universität Ulm , Christopher Wolf (Hg.) KU Leuven
2. Krypto-Tag – Workshop über Kryptographie, Universität Ulm

2005-03 Walter Guttmann, Markus Maucher
Constrained Ordering

2006-01 Stefan Sarstedt
Model-Driven Development with ACTIVECHARTS, Tutorial

2006-02 Alexander Raschke, Ramin Tavakoli Kolagari
Ein experimenteller Vergleich zwischen einer plan-getriebenen und einer leichtgewichtigen Entwicklungsmethode zur Spezifikation von eingebetteten Systemen

2006-03 Jens Kohlmeyer, Alexander Raschke, Ramin Tavakoli Kolagari
Eine qualitative Untersuchung zur Produktlinien-Integration über Organisationsgrenzen hinweg

2006-04 Thorsten Liebig
Reasoning with OWL - System Support and Insights –

2008-01 H.A. Kestler, J. Messner, A. Müller, R. Schuler
On the complexity of intersecting multiple circles for graphical display

2008-02 Manfred Reichert, Peter Dadam, Martin Jurisch,l Ulrich Kreher, Kevin Göser, Markus Lauer
Architectural Design of Flexible Process Management Technology

2008-03 Frank Raiser
Semi-Automatic Generation of CHR Solvers from Global Constraint Automata

2008-04 Ramin Tavakoli Kolagari, Alexander Raschke, Matthias Schneiderhan, Ian Alexander
Entscheidungsdokumentation bei der Entwicklung innovativer Systeme für produktlinien-basierte Entwicklungsprozesse

2008-05 Markus Kalb, Claudia Dittrich, Peter Dadam
Support of Relationships Among Moving Objects on Networks

2008-06 Matthias Frank, Frank Kargl, Burkhard Stiller (Hg.)
WMAN 2008 – KuVS Fachgespräch über Mobile Ad-hoc Netzwerke

2008-07 M. Maucher, U. Schöning, H.A. Kestler
An empirical assessment of local and population based search methods with different degrees of pseudorandomness

2008-08 Henning Wunderlich
Covers have structure

2008-09 Karl-Heinz Niggl, Henning Wunderlich
Implicit characterization of FPTIME and NC revisited

2008-10 Henning Wunderlich
On span-P$^c c$ and related classes in structural communication complexity

2008-11 M. Maucher, U. Schöning, H.A. Kestler
On the different notions of pseudorandomness

2008-12 Henning Wunderlich
On Toda's Theorem in structural communication complexity

2008-13 Manfred Reichert, Peter Dadam
Realizing Adaptive Process-aware Information Systems with ADEPT2

2009-01 Peter Dadam, Manfred Reichert
The ADEPT Project: A Decade of Research and Development for Robust and Fexible Process Support Challenges and Achievements

2009-02 Peter Dadam, Manfred Reichert, Stefanie Rinderle-Ma, Kevin Göser, Ulrich Kreher, Martin Jurisch
Von ADEPT zur AristaFlow® BPM Suite – Eine Vision wird Realität "Correctness by Construction" und flexible, robuste Ausführung von Unternehmensprozessen

2009-03 Alena Hallerbach, Thomas Bauer, Manfred Reichert
Correct Configuration of Process Variants in Provop

2009-04 Martin Bader
On Reversal and Transposition Medians

2009-05 Barbara Weber, Andreas Lanz, Manfred Reichert
Time Patterns for Process-aware Information Systems: A Pattern-based Analysis

2009-06 Stefanie Rinderle-Ma, Manfred Reichert
Adjustment Strategies for Non-Compliant Process Instances

2009-07 H.A. Kestler, B. Lausen, H. Binder H.-P. Klenk. F. Leisch, M. Schmid
Statistical Computing 2009 – Abstracts der 41. Arbeitstagung

2009-08 Ulrich Kreher, Manfred Reichert, Stefanie Rinderle-Ma, Peter Dadam
Effiziente Repräsentation von Vorlagen- und Instanzdaten in Prozess-Management-Systemen

2009-09 Dammertz, Holger, Alexander Keller, Hendrik P.A. Lensch
Progressive Point-Light-Based Global Illumination

2009-10 Dao Zhou, Christoph Müssel, Ludwig Lausser, Martin Hopfensitz, Michael Kühl, Hans A. Kestler
Boolean networks for modeling and analysis of gene regulation

2009-11 J. Hanika, H.P.A. Lensch, A. Keller
Two-Level Ray Tracing with Recordering for Highly Complex Scenes

2009-12 Stephan Buchwald, Thomas Bauer, Manfred Reichert
Durchgängige Modellierung von Geschäftsprozessen durch Einführung eines Abbildungsmodells: Ansätze, Konzepte, Notationen

2010-01 Hariolf Betz, Frank Raiser, Thom Frühwirth
A Complete and Terminating Execution Model for Constraint Handling Rules

2010-02 Ulrich Kreher, Manfred Reichert
Speichereffiziente Repräsentation instanzspezifischer Änderungen in Prozess-Management-Systemen

2010-03 Patrick Frey
Case Study: Engine Control Application

2010-04 Matthias Lohrmann und Manfred Reichert
Basic Considerations on Business Process Quality

2010-05 HA Kestler, H Binder, B Lausen, H-P Klenk, M Schmid, F Leisch (eds):
Statistical Computing 2010 - Abstracts der 42. Arbeitstagung

2010-06 Vera Künzle, Barbara Weber, Manfred Reichert
Object-aware Business Processes: Properties, Requirements, Existing Approaches

2011-01 Stephan Buchwald, Thomas Bauer, Manfred Reichert
Flexibilisierung Service-orientierter Architekturen

2011-02 Johannes Hanika, Holger Dammertz, Hendrik Lensch
Edge-Optimized À-Trous Wavelets for Local Contrast Enhancement with Robust Denoising

2011-03 Stefanie Kaiser, Manfred Reichert
Datenflussvarianten in Prozessmodellen: Szenarien, Herausforderungen, Ansätze

2011-04 Hans A. Kestler, Harald Binder, Matthias Schmid, Friedrich Leisch, Johann M. Kraus (eds):
Statistical Computing 2011 - Abstracts der 43. Arbeitstagung

2011-05 Vera Künzle, Manfred Reichert
PHILharmonicFlows: Research and Design Methodology

2011-06 David Knuplesch, Manfred Reichert
Ensuring Business Process Compliance Along the Process Life Cycle

2011-07 Marcel Dausend
Towards a UML Profile on Formal Semantics for Modeling Multimodal Interactive Systems

2011-08 Dominik Gessenharter
Model-Driven Software Development with ACTIVECHARTS - A Case Study

2012-01 Andreas Steigmiller, Thorsten Liebig, Birte Glimm
Extended Caching, Backjumping and Merging for Expressive Description Logics

2012-02 Hans A. Kestler, Harald Binder, Matthias Schmid, Johann M. Kraus (eds):
Statistical Computing 2012 - Abstracts der 44. Arbeitstagung

2012-03 Felix Schüssel, Frank Honold, Michael Weber
Influencing Factors on Multimodal Interaction at Selection Tasks

2012-04 Jens Kolb, Paul Hübner, Manfred Reichert
Model-Driven User Interface Generation and Adaption in Process-Aware Information Systems

2012-05 Matthias Lohrmann, Manfred Reichert
Formalizing Concepts for Efficacy-aware Business Process Modeling

2012-06 David Knuplesch, Rüdiger Pryss, Manfred Reichert
A Formal Framework for Data-Aware Process Interaction Models

2012-07 Clara Ayora, Victoria Torres, Barbara Weber, Manfred Reichert, Vicente Pelechano
Dealing with Variability in Process-Aware Information Systems: Language Requirements, Features, and Existing Proposals

2013-01 Frank Kargl
Abstract Proceedings of the 7th Workshop on Wireless and Mobile Ad-Hoc Networks (WMAN 2013)

2013-02 Andreas Lanz, Manfred Reichert, Barbara Weber
A Formal Semantics of Time Patterns for Process-aware Information Systems

2013-03 Matthias Lohrmann, Manfred Reichert
Demonstrating the Effectiveness of Process Improvement Patterns with Mining Results

2013-04 Semra Catalkaya, David Knuplesch, Manfred Reichert
Bringing More Semantics to XOR-Split Gateways in Business Process Models Based on Decision Rules

2013-05 David Knuplesch, Manfred Reichert, Linh Thao Ly, Akhil Kumar, Stefanie Rinderle-Ma
On the Formal Semantics of the Extended Compliance Rule Graph

2013-06 Andreas Steigmiller, Birte Glimm, Thorsten Liebig
Nominal Schema Absorption

2013-07 Hans A. Kestler, Matthias Schmid, Florian Schmid, Dr. Markus Maucher, Johann M. Kraus (eds)
Statistical Computing 2013 - Abstracts der 45. Arbeitstagung

2013-08 Daniel Ott, Dr. Alexander Raschke
Evaluating Benefits of Requirement Categorization in Natural Language Specifications for Review Improvements

2013-09 Philip Geiger, Rüdiger Pryss, Marc Schickler, Manfred Reichert
Engineering an Advanced Location-Based Augmented Reality Engine for Smart Mobile Devices

2014-01 Andreas Lanz, Manfred Reichert
Analyzing the Impact of Process Change Operations on Time-Aware Processes

2014-02 Andreas Steigmiller, Birte Glimm, and Thorsten Liebig
Coupling Tableau Algorithms for the DL SROIQ with Completion-based Saturation Procedures

2014-03 Thomas Geier, Felix Richter, Susanne Biundo
Conditioned Belief Propagation Revisited: Extended Version

2014-04 Hans A. Kestler, Matthias Schmid, Ludwig Lausser, Johann M. Kraus (eds)
Statistical Computing 2014 - Abstracts der 46. Arbeitstagung

2014-05 Andreas Lanz, Roberto Posenato, Carlo Combi, Manfred Reichert
Simple Temporal Networks with Partially Shrinkable Uncertainty (Extended Version)

2014-06 David Knuplesch, Manfred Reichert
An Operational Semantics for the Extended Compliance Rule Graph Language

2015-01 Andreas Lanz, Roberto Posenato, Carlo Combi, Manfred Reichert
Controlling Time-Awareness in Modularized Processes (Extended Version)

2015-03 Raphael Frank, Christoph Sommer, Frank Kargl, Stefan Dietzel, Rens W. van der Heijden
Proceedings of the 3rd GI/ITG KuVS Fachgespräch Inter-Vehicle Communication (FG-IVC 2015)

2015-04 Axel Fürstberger, Ludwig Lausser, Johann M. Kraus, Matthias Schmid, Hans A. Kestler (eds)
Statistical Computing 2015 - Abstracts der 47. Arbeitstagung

2016-03 Ping Gong, David Knuplesch, Manfred Reichert
Rule-based Monitoring Framework for Business Process Compliance

2016-04 Axel Fürstberger, Ludwig Lausser, Johann M. Kraus, Matthias Schmid, Hans A. Kestler (eds)
Statistical Computing 2016 - Abstracts der 48. Arbeitstagung

2016-05 Axel Fürstberger, Johann M. Kraus, Hans A. Kestler (eds)
Classification 2016 - Abstracts of the 5th German-Japanese Symposium

2016-06 Vera Künzle, Sebastian Steinau, Kevin Andrews, and Manfred Reichert
An Approach for Modeling and Coordinating Process Interactions

2017-01 Hans A. Kestler, Matthias Schmid, Ludwig Lausser, Johann M. Kraus, Axel Fürstberger (eds)
Statistical Computing 2017 - Abstracts der 49. Arbeitstagung

2018-01 Hans A. Kestler, Matthias Schmid, Ludwig Lausser, Johann M. Kraus, Axel Fürstberger (eds)
Statistical Computing 2018 - Abstracts der 50. Arbeitstagung

Ulmer Informatik-Berichte
ISSN 0939-5091

Herausgeber:
Universität Ulm
Fakultät für Ingenieurwissenschaften und Informatik
89069 Ulm