ulm university universität
uulm

Statistical Computing 2017
Abstracts der 49. Arbeitstagung

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

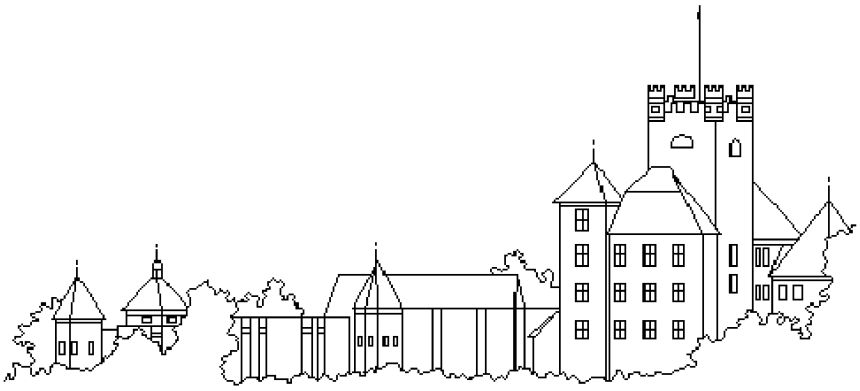
Ulmer Informatik-Berichte

Nr. 2017-01
July 2017

International Graduate School
in Molecular Medicine Ulm

SyStaR

Statistical Computing 2017



49. Arbeitstagung

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

23.07. - 25.07.2017, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, July 23, 2017

16:50 – 17:00		**Opening of the workshop: H. A. Kestler**
		Introduction: H. A. Kestler
17:00 – 18:00	Daniel Braun (Ulm)	Optimal Statistical Decision-Making with Limited Resources
18:00 – 20:00		**Dinner**
		Introduction: M. Schmid
20:00 – 21:00	Helga Wagner (Linz)	Sparse Bayesian Modelling

Monday, July 24, 2017

		Introduction: H. A. Kestler
09:00 – 10:00	Barbara Hammer (Bielefeld)	Transfer Learning and Learning with Concept Drift
10:00 – 12:00		**Chair: A. Mayr**
10:00 – 10:20	Sarah Schnackenberg (Dortmund)	Online Linear Discriminant Analysis for Data Streams with Concept Drift
10:20 – 10:40		**Coffee Break**
10:40 – 11:00	Sebastian Meyer (Erlangen)	Social contact data in endemic-epidemic models and probabilistic forecasting with surveillance
11:00 – 11:20	Dja-Shin Wang (Taiwan)	Monitoring optimization parameter to minimize energy consumption for carbon emission reduction
11:20 – 11:40	Sebastian Krey (Köln)	Structural Health Monitoring for Resource-efficient Usage of Fibre-Reinforced Plastic
11:40 – 12:00	Benjamin Mayer (Ulm)	Sample size estimation for pilot experiments in animal research using a Markov Chain Monte Carlo approach
12:00 – 13:20		**Lunch**

Monday, July 24, 2017

13:20 – 16:40		**Chair: E. Waldmann**
13:20 – 15:00	Sarah Brockhaus and David Rügamer (München)	Boosting Functional Regression Models
15:20 – 15:40	Moritz Berger (Bonn)	Tree-Based Modelling of Varying Coefficient Terms
15:40 – 16:00	Christian Thiele (Osnabrück)	The cutpointr package: Improved and tidy estimation of optimal cutpoints
16:00 – 16:20	Laura Beggel (München)	mlrFDA: an R toolbox for functional data analysis
16:20 – 16:40	Michel Lang (Dortmund)	batchtools: Parallelization on high-performance computing clusters with R
16:40 – 17:00		**Coffee Break**
17:00 – 18:00		**Chair: B. Bischl**
17:00 – 17:20	Leonie Weinhold (Bonn)	The betaboost package – a boosting framework to estimate and select models for bounded outcomes like Health Related Quality of Life data
17:20 – 17:40	Hoang Nguyen (Madrid)	Variational Inference for high dimentional factor copulas
17:40 – 18:00	Elisabeth Waldmann (Erlangen)	Variable Selection and Allocation in Joint Models for Longitudinal and Time-to-Event Data via Boosting
18:00 – 20:00		**Dinner**
20:00 – 21:00	Sarah Brockhaus and David Rügamer (München)	Hands-on Tutorial: Boosting Functional Regression Models

Tuesday, July 25, 2017

		Introduction: E. Waldmann
09:00 – 09:40	Heidi Seibold (Zürich)	Model-Based Recursive Partitioning for Stratified and Personalised Treatment Effect Estimation
09:40 – 10:20		**Chair: U. Mansmann**
09:40 – 10:00	Katharina Hees (Heidelberg)	Incorporation of historical data in clinical trials using robust prior
10:00 – 10:20	Jörn Lötsch (Frankfurt)	Supervised and unsupervised machine-learning methods for pain research
10:20 – 10:40		**Coffee Break**

Tuesday, July 25, 2017

10:40 – 12:00		**Chair: J. M. Kraus**
10:40 – 11:00	Robin Szekely (Ulm)	How feature selection in multi-class classifier systems affects the prediction of diagnostic phenotypes
11:00 – 11:20	Andrea Bommert (Dortmund)	A multi-criteria approach to find predictive and sparse models with stable feature selection for highdimensional data
11:20 – 11:40	Britta Velten (Heidelberg)	Adaptive penalization for high-dimensional regression and classification with non-exchangeable features
11:40 – 12:00	Tobias Hepp (Erlangen)	Variable selection for model-based gradient boosting using random probes
12:00 – 13:20		**Lunch**
		Introduction: A. Mayr
13:20 – 14:00	Christian Staerk (Aachen)	High-Dimensional Variable Selection via Low-Dimensional Adaptive Learning
14:00 – 15:00		**Chair: L. Lausser**
14:00 – 14:20	Corinna Ernst (Köln)	Using generalized additive models for CNV detection on multi gene panels
14:20 – 14:40	Pascal Schlosser (Freiburg)	High-dimensional genome-wide association studies (GWAS): A field for data restructuring?
14:40 – 15:00		Posterteaser
15:00 – 15:40		**Coffee Break + Postersession**
15:40 – 17:00		**Chair: G. Casalicchio**
15:40 – 16:00	Lyn-Rouven Schirra (Ulm)	Accelerating the Detection of Ordinal Structures
16:00 – 16:20	Roman Hornung (München)	Ordinal Forests: A versatile tool for ordinal regression
16:20 – 16:40	Gunther Schauberger (München)	A Common Framework and Software Package for the Inclusion and Selection of Covariates in Bradley-Terry Models
16:40 – 17:00	Thomas Welchowski (Bonn)	Sparse kernel deep stacking networks
17:00 – 18:00		Working group meeting on **Statistical Computing 2018** and other topics (all welcome)
18:00 – 20:00		**Dinner**

Contents

Optimal statistical decision-making with limited resources

Daniel Braun[1]

Expected utility maximization is the foundation of statistical decision-making, but ignores the problem of limited computational resources. Here we discuss information-theoretic approaches for decision-making with limited resources including limited sample size or model uncertainty. We discuss how such limitations can lead to interesting deviations from Bayes-optimal behavior including robustness, the formation of abstractions and the coupling of perception and action.

[1] Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany

daniel.braun@uni-ulm.de

Sparse Bayesian Modelling

Helga Wagner [1]

Sparse modelling and variable selection is one of the most important issues in regression type models, as in applications often a large number of covariates on comparably few subjects are available. Estimation of regression effects in such *large p, small n* problems is ill-conditioned: estimated regression effects typically have large standard errors, estimation results are instable and fitted models have no good predictive performance. To identify those regressors which have a non-negligible effect, many methods have been developed, to identify zero regression coefficients.
In a Bayesian approach sparsity can be achieved by specifying appropriate prior distributions on the regression coefficients, which induce shrinkage of small, irrelevant coefficients to zero. This goal can be accomplished either by absolutely continuous priors with a spike at zero or by spike and slab priors, which are a mixture of a spike component with small variance centered at zero and a comparably flat slab. For both types of prior distributions Bayesian inference can be accomplished using MCMC methods.
These types of prior distributions have been employed beyond standard regression models to achieve sparsity in more complex models, e.g. in random effects or state space models. Using spike and slab priors also on variance and covariance parameters allows to start with a rather flexible model, e.g. in state space models with time-varying effects of all potential regressors and classify effects as zero, constant or varying over time during estimation.
Recently priors have been developed that allow to exploit the structure of predictors, e.g. the ordering structure of the levels of ordinal predictors or the grouping of variables. This talk reviews sparse Bayesian modelling in various model classes and discusses how sparsity in modelling the effect of a categorical predicator can be achieved in a Bayesian approach.

[1]Department of Applied Statistics, Johannes Kepler Universitat, Altenbergerstrasse 69, A-4040 LINZ, Austria

Helga.Wagner@jku.at

Transfer learning and learning with concept drift

Barbara Hammer[1]

One of the main assumptions of classical machine learning is that data are generated by a stationary concept. This, however, is violated in practical applications e.g. in the context of life long learning, for the task of system personalization, or whenever sensor degradation or non-stationary environments cause a fundamental change of the observed signals. Within the talk, we will give an overview about recent developments in the field of learning with concept drift, and we will address two particular challenges in more detail: (1) How to cope with a fundamental change of the data representation which is caused e.g. by a misplacement or exchange of sensors? (2) How to deal with drifting concepts which change either rapidly or smoothly over time, e.g. caused by a non-stationary environment? We will present novel intuitive distance-based classification approaches which can tackle such settings by means of suitable metric learning or adaptive memory concepts, respectively, and we will demonstrate their performance in different application domains ranging from computer vision to the control of protheses.

[1] Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany

bhammer@techfak.uni-bielefeld.de

Online Linear Discriminant Analysis for Data Streams with Concept Drift

Sarah Schnackenberg[1], Uwe Ligges[1], and Claus Weihs[1]

There are various methods in the literature that adapt and extend classical classification methods, particularly linear discriminant analysis (LDA), for handling data streams.
Pang et al. (2005) [1] focus on Fisher's discriminant analysis and use standard formulas for updating the means and variances iteratively with new observations. The resulting classification rule is identical to that built on all observations at a time in batch mode: every data point has the same influence and is assigned the same weight, respectively.
Observing data streams, the underlying distribution of features and/or target variable may change over time (*concept drift* [2]). In such a case, exact updates and equal weighting of all observations in the data streams may lead to bad performance of the classifier. A weighting adapted to the currentness of new observations can prove a better performance.
Kuncheva and Plumpton (2008) [3] deal with an adaptive online version of the canonical LDA. They also use standard formulas for updating the relevant values. In addition, they extend the adaptive online version with an option to adjust to concept drift by introducing a fixed or adaptive learning rate. Hence, new observations are assigned different weights indirectly.
Such methods that adapt to concept drift may lead to a good goodness of fit. Nevertheless, the forecasting quality of the classifier to predict new observations can still be bad in case the underlying distribution gradually changes further on.
In the talk existing methods for online LDA are extended. Under some assumptions we estimate a model for the time depending concept drift that is used to predict the forthcoming distribution of the features. The estimated parameters of these distributions are finally used in the LDA to build the classification rule and hence to predict new observations. In a simulation study we consider different kinds of concept drift and compare the new extended methods with the methods these are based on.

[1] Fakultat Statistik, Technische Universitat Dortmund, 44221 Dortmund

schnackenberg@statistik.tu-dortmund.de,ligges@statistik.tu-dortmund.de,
weihs@statistik.tu-dortmund.de

References

1 Pang, S., Ozawa, S., and Kasabov, N. (2005): Incremental Linear Discriminant Analysis for Classification of Data Streams. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 35(5), 905–914.

2 Hoens, T.R., Polikar, R., and Chawla, N.V. (2012): Learning from streaming data with concept drift and imbalance: an overview. Progress in Artificial Intelligence, 1(1), 89–101.

3 Kuncheva, L.I. and Plumpton, C.O. (2008): Adaptive Learning Rate for Online Linear Discriminant Classifiers. Structural, Syntactic, and Statistical Pattern Recognition. Lecture Notes in Computer Science, 5342, 510–519.

Social contact data in endemic-epidemic models and probabilistic forecasting with surveillance

Sebastian Meyer[1], Johannes Bracher[2], and Leonhard Held[2]

Routine surveillance of notifiable infectious diseases gives rise to weekly counts of reported cases stratified by region and age group. A well-established approach to the statistical analysis of such surveillance data are endemic-epidemic time-series models (`hhh4`) as implemented in the `R` package **surveillance** [1]. Autoregressive model components reflect the temporal dependence inherent to communicable diseases. Spatial dynamics are largely driven by human travel and can be captured by movement network data or a parametric power law based on the adjacency matrix of the regions. Furthermore, the social phenomenon of "like seeks like" produces characteristic contact patterns between subgroups of a population, in particular with respect to age. We thus incorporate an age-structured contact matrix in the `hhh4` modelling framework to

1. assess age-specific disease spread while accounting for its spatial pattern [2],
2. improve probabilistic forecasts of infectious disease spread [3].

We analyze weekly surveillance counts on norovirus gastroenteritis from the 12 city districts of Berlin, in six age groups, from week 2011/27 to week 2015/26. The following year (2015/27 to 2016/26) is used to assess the quality of the predictions.

References

1 Sebastian Meyer, Leonhard Held, and Michael Hohle. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. Journal of Statistical Software, 2017. URL http://arxiv.org/abs/1411.0416. In press.

2 Sebastian Meyer and Leonhard Held. Incorporating social contact data in spatio-temporal models for infectious disease spread. Biostatistics, 18(2):338–351, 2017. doi:10.1093/biostatistics/kxw051.

3 Leonhard Held, Sebastian Meyer, and Johannes Bracher. Probabilistic forecasting in infectious disease epidemiology: The thirteenth armitage lecture. bioRxiv, 2017. doi:10.1101/104000.

[1] Institute of Medical Informatics, Biometry, and Epidemiology, Friedrich-Alexander-Universitat Erlangen-Nurnberg, Waldstr. 6, 91054 Erlangen, Germany
[2] Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84,8001 Zurich, Switzerland

`seb.meyer@fau.de`

Monitoring optimization parameter to minimize energy consumption for carbon emission reduction

Dja Shin Wang[1], Ting-Wei Yeh[2], Tong Yuan Koo[2]

Machine tools are responsible for environmental impacts owing to their consumption of electrical energy. Improving energy efficiency of manufacturing processes requires knowledge about the energy consumption as a function of the machine tool and cutting process itself. One of the processes widely used in manufacturing is turning. As a result, many companies are adopting the concept of cleaner production. Cleaner production relies on the creation of products using systems that do not pollute and conserve natural resources.

The past research represent that how to section machine tools or scheduling to reduce idle time for energy consumption. This study control chart to monitor turning process which monitor surface roughness and energy to know whether the cutting parameters (depth of cut, feed rate, cutting speed) are maintained on setting or not. It is purpose to find the outlier and return to check cutting parameter if it is wrong or not. According to the analysis, control chart for surface roughness is sensitive to feed rate, but it is not sensitive to depth of cut and cutting velocity. MEWMA control chart is great for monitoring three cutting parameters. Finding outliers fast and exactly to make turning process stable, in order to achieve green production for energy efficiency, reducing energy consumption and reducing pollution.

[1]Department of Business Administration, TransWorld University, 1221, Zheng Nan Rd., Douliu 640, Yunlin, Taiwan, R.O.C.

[2]Department of Industrial Engineering and Management, National Yunlin, University of Science and Technology, Douliu, Taiwan, R.O.C.

shin@twu.edu.tw, m10321308@yuntech.edu.tw, kooty@yuntech.edu.tw

References

1 Camposeco-Negrete, C. (2013). "Optimization of cutting parameters for minimizing energy consumption in turning of AISI 6061 T6 using Taguchi methodology and ANOVA." Journal of Cleaner Production 53: 195-203.

2 Park, J. and C.-H. Jun (2015). "A new multivariate EWMA control chart via multiple testing." Journal of Process Control 26: 51-55.

3 Li, W., et al. (2011). An investigation into fixed energy consumption of machine tools. Glocalized Solutions for Sustainability in Manufacturing, Springer: 268-273.

4 Yan, J. and L. Li (2013). "Multi-objective optimization of milling parameters–the trade-offs between energy, production rate and cutting quality." Journal of Cleaner Production 52: 462-471.

5 Lee, B. and Y. Tarng (2000). "Cutting-parameter selection for maximizing production rate or minimizing production cost in multistage turning operations." Journal of Materials Processing Technology 105(1): 61-66.

6 Roberts, S. (2000). "Control chart tests based on geometric moving averages." Technometrics 42(1): 97-101.

7 Vera do Carmo, C., et al. (2004). "Comparative study of the performance of the CuSum and EWMA control charts." Computers & Industrial Engineering 46(4): 707-724.

Structural Health Monitoring for Resource-efficient Usage of Fibre-Reinforced Plastic

Sebastian Krey[1], Thomas Bartz-Beielstein[1], Yixi Fu[1] and Carolin Gorecki[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.
In this work we present how machine learning methods used on features, which were extracted from the signals detected by piezoelectric sensors, can be used to automatically detect damages of varying degree in glass-fibre reinforced plastic plates.
We also show how the signal quality increases by implementing the piezoelectric sensors with a newly developed method directly between the layers of the glass-fibre reinforced plastic compared to the current state of the art usage of externally applied sensors.

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

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

Sample size estimation for pilot experiments in animal research using a Markov Chain Monte Carlo approach

Benjamin Mayer[1], *Andreas Allgoewer*[1]

Background: Statistical determination of sample size is mandatory when planning animal experiments, but usually difficult to implement appropriately. The main reason is that prior information is hardly available and thus the assumptions made cannot be verified reliably. This is especially true for pilot experiments. Statistical simulation may help in these situations.
Methods: We used a Markov Chain Monte Carlo (MCMC) approach in order to verify the assumptions made on different distribution parameters used for power and sample size calculations in animal experiments. Binomial and normal distributions were simulated for categorical and continuous endpoints, respectively.
Results: The simulations showed that the common application for 5-6 animals per group for continuous endpoints is reasonable. Even in case of small effect sizes the statistical power would be sufficiently large ($\geq 80\%$). For categorical outcomes, group sizes should never remain under 8 animals. Otherwise, a sufficient statistical power cannot be guaranteed, even in case of large effects.
Discussion: The MCMC approach demonstrated to be a useful method for calculating the sample size in animals studies which are lacking prior data. Of course, the simulation results particularly depend on the assumptions made on the distributional properties and effects to be detected, but the same also holds in situations where prior data are available. MCMC is therefore a promising approach for a profound planning of pilot experiments in animal research.

[1] Institute for Epidemiology and Medical Biometry, Ulm University

benjamin.mayer@uni-ulm.de, andreas.allgoewer@uni-ulm.de

Boosting Functional Regression Models

Sarah Brockhaus[1] *and David Rugamer*[1]

Functional data, such as curves, trajectories or images, can be found in many scientific fields. Regression models are a versatile tool for data analysis and various models have been proposed for regression with functional variables. The R add-on package FDboost is a flexible toolbox for the estimation of functional regression models by model-based boosting. It provides the possibility to fit regression models for scalar and functional response with effects of scalar as well as functional covariates, i.e., scalar-on-function, function-on-scalar and function-on-function regression models. In addition to mean regression, quantile regression models as well as generalized additive models for location scale and shape can be fitted with FDboost. Furthermore, boosting can be used in high-dimensional data settings with more covariates than observations. The methods for scalar-on-function regression are illustrated with spectrometric data of fossil fuels and those for functional response regression with a data set including bioelectrical signals for emotional episodes.

[1] Department of Statistics, LMU Munich, Munich, Germany

sarah.brockhaus@stat.uni-muenchen.de, david.ruegamer@stat.uni-muenchen.de

Tree-Based Modelling of Varying Coefficient Terms

Moritz Berger[1] *, Gerhard Tutz*[2] *and Matthias Schmid*[1]

Generalized linear and additive models are very efficient tools that are widely applied to regression problems in various fields of application. A flexible way to appropriatly capture predictor-response relationships are varying coefficient models, proposed by [1]. The inclusion of varying coefficients is particularly useful if there are several covariates on different scales that potentially interact with one another.
The principle of varying coefficient models is that the coefficients of the covariates $x_1, ..., x_p$ are allowed to change with the value of other variables $z_1, ..., z_p$, called *effect modifiers*. It is typically assumed that the effect modifier z_j changes the coefficients of x_j through a smooth function of unspecified form, respectively. Several strategies have been proposed for the estimation of these smooth functions, for example by [2] and [3].
All the traditional approaches have in common that they aim at distinguishing between varying and non-varying coefficients. Given a specific effect modifier z_j one wants to know if the effect of x_j is constant over the whole range of z_j or varies across values of z_j. Thus, the effect modifier has to be specified beforehand. Then one determines the way it modifies coefficients. However, it is often not known which variable is a potential effect modifier and if more than one effect modifier causes varying coefficients.
To address this issue, we propose a method based on *recursive partitioning techniques* that itself identifies the effect modifiers that induce varying coefficients if they are present. We consider only one set of covariates $x_1, ..., x_p$ and assume that the coefficient of each covariate x_j can be modified by all the other variables x_m, $m \in \{1, ..., p\} \setminus j$. As with classical trees the coefficients are successively split into two regions and in each region the coefficient is fitted by a constant. In each step of the algorithm one chooses a coefficient (corresponding to a covariate), an effect modifier and a corresponding split-point. The method yields an individual tree for each covariate that shows varying coefficients. If varying coefficients are present the relevant effect modifiers are selected simultaneously.

[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, gerhard.tutz@stat.uni-muenchen.de,
matthias.schmid@imbie.uni-bonn.de

References

1 Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. Journal of the Royal Statistical Society B 55, pp. 757-796.

2 Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. Annals of Statistics 27 (5), pp. 1491-1518.

3 Lu, Y., Zhang, R. and Zhu, L. (2008). Penalized spline estimation for varying-coefficient models. Communications in Statistics - Theory and Methods 37 (14), pp. 2249-2261.

The cutpointr package: Improved and tidy estimation of optimal cutpoints

Christian Thiele[1], *Gerrit Hirschfeld*[1]

Clinicians often use cutpoints or decision-thresholds to decide e.g. whether or not a patient with a depression score of, say, 20 needs treatment for her or his depression. The R package cutpointr [1] allows for estimating such optimal cutpoints for binary decisions by maximizing a specified metric or by using kernel estimation or parametric distribution-based methods [2, 3]. The latter methods have shown to be preferable to maximizing a metric empirically, especially in small samples. A few of the most popular metrics and existing functions [4] are included and user-defined functions for calculating custom metrics can easily be supplied. Similarly, user-defined cutpoint functions can be supplied to make use of the package's plotting and bootstrapping functionality. This parallelizable routine provides estimates of the cutpoints' variability and various in- and out-of-bag performance metrics. cutpointr follows current tidy programming practices to allow for efficient estimation and use in simulation studies as well as interplay with functions from the tidyverse. ROC curves, precision recall plots and further data can be plotted using convenient built-in functions. In terms of speed and scalability, cutpointr compares favorably to existing solutions. We also discuss future plans for cutpointr, specifically a Shiny-interface to make the package more accessible.

References

1 https://github.com/Thie1e/cutpointr

2 Fluss, R., Faraggi, D., Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. Biometrical Journal, 47(4), 458–472.

3 Leeflang, M. M., Moons, K. G., Reitsma, J. B., Zwinderman, A. H. (2008). Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clinical Chemistry, (4), 729–738.

4 López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C., Gude-Sampedro, F. (2014). OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of Statistical Software, 61(i08).

[1] Hochschule Osnabrück, Albrechtstraße 30, 49076 Osnabrück

c.thiele@hs-osnabrueck.de, g.hirschfeld@hs-osnabrueck.de

mlrFDA: an R toolbox for functional data analysis

Laura Beggel[*,1,2]*,Xudong Sun*[*,1]*,Bernd Bischl*[1]

In functional data analysis, measurements are taken over a continuous space and thus features are intercorrelated across domain. Most regular machine learning toolboxes fail to take this property into consideration. We present `mlrFDA`, an `R` toolbox for functional data analysis and benchmarking. The package's functionality comprises regression and classification of functional data and is applicable to one or multidimensional functional data with varying functional measurement lengths. State-of-the-art functional data analysis methods such as Functional Linear Array Model [1] and Functional Generalized Additive Models [2] are included. Since `mlrFDA` is integrated into the open source machine learning toolbox `mlr` [3] it also inherits `mlr`'s full functionality for benchmarking and hyperparameter optimization. The structure of `mlrFDA` and its use cases will be presented, and a first benchmarking result on several open access functional data sets will be demonstrated.

References

1 Brockhaus, S., Scheipl, F., Hothorn, T., and Greven, S. (2015), The Functional Linear Array Model. Statistical Modelling, 15(3), 279-300.

2 McLean, M. W., Hooker, G., Staicu, A. M., Scheipl, F., Ruppert, D. (2014). Functional generalized additive models. Journal of Computational and Graphical Statistics, 23(1), 249-269.

3 Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. and Jones, Z. (2016). "mlr: Machine Learning in R." Journal of Machine Learning Research, 17(170), pp. 1-5. http://jmlr.org/papers/v17/15-066.html

[1] Ludwig-Maximilians-Universität München, Institut für Statistik, Ludwigstraße 33, 80539 München
[2] Robert Bosch GmbH, Zentralbereich Forschung und Vorausentwicklung, Robert-Bosch-Campus 1, 71272 Renningen
[*] The first two authors contributed equally to this work.

`laura.beggel@googlemail.com, xudong.sun@stat.uni-muenchen.de, bernd.bischl@stat.uni-muenchen.de`

batchtools: Parallelization on high-performance computing clusters with R

Michel Lang[1]

The R package `batchtools` is the successor of the `R` package `BatchJobs` [1]. It provides an implementation of a Map-like operation for a variety of parallel backends: (i) Local execution in the current or externally spawned `R` session, (ii) Local parallel execution (using `parallel` or `snow`), (iii) Execution on loosely connected machines via SSH, (iv) Docker Swarm, (v) IBM Spectrum LSF, (vi) OpenLava, (vii) Univa Grid Engine (formerly Oracle Grind Engine and Sun Grid Engine), (viii) Slurm Workload Manager, and (ix) TORQUE/PBS Resource Manager.

The communication between the master and the computational nodes is kept as simple as possible and relies solely on a shared file system. An in-memory database keeps track of the computational status of all jobs. Unique job seeds ensure reproducibility across systems, log files can conveniently be searched using regular expressions and jobs can be annotated with arbitrary tags. Jobs can be chunked (i.e., merged into one technical cluster job) to be executed as one virtual job on a node (executed sequentially or using multiple local CPUs) in order to reduce the overhead induced by job management and starting/stopping `R`. All in all, the provided tools allow users to work with many thousands or even millions of jobs in an organized and efficient manner.

The `batchtools` package also comes with an abstraction mechanism to assist in conducting large-scale computer experiments. The mechanism is similar to the package `BatchExperiments` [1] which `batchtools` now also supersedes. This extension is especially suited for (but not restricted to) benchmarking machine learning algorithms. For this purpose, the `mlr` package [2] implements the function `batchmark()` which utilizes `batchtools` to conveniently conduct large benchmark studies in parallel.

References

1 Bernd Bischl, Michel Lang, Olaf Mersmann, Jörg Rahnenführer, Claus Weihs (2015). BatchJobs and BatchExperiments: Abstraction Mechanisms for Using R in Batch Environments. Journal of Statistical Software, 64(11), 1–25. `http://www.jstatsoft.org/v64/i11/`.

2 Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, Zachary M. Jones (2016). mlr: Machine Learning in R. Journal of Machine Learning Research, 17(170), 1–5. `http://jmlr.org/papers/v17/15-066.html`

[1] TU Dortmund, Vogelpothsweg 87, 44227 Dortmund

`lang@statistik.tu-dortmund.de.de`

The betaboost package – a boosting framework to estimate and select models for bounded outcomes like Health Related Quality of Life data

Andreas Mayr[1,2], Leonie Weinhold[2], Benjamin Hofner[3], Stephanie Titze[4], Olaf Gefeller[1], Matthias Schmid[2]

The analysis of bounded outcomes like the ones resulting from commonly used Health Related Quality of Life scales is a widely discussed issue [1]. Furthermore, modern epidemiological studies often collect vast amounts of data, leading to large numbers of potential explanatory variables [2]. The methodological challenges in this context are two-fold: (i) the bounded outcome makes traditional statistical modelling approaches problematic and (ii) the classical statistical inference methods become unfeasible in the presence of high-dimensional data. With the betaboost package, we present a statistical software tackling both issues by incorporating flexible approaches for beta regression in a model-based boosting framework [3].

In our software, two different variants for beta-regression are implemented: while classical beta regression focuses on modelling the expected value via the mean parameter, an extended version additionally models the precision parameter based on covariates in the spirit of distributional regression [4].

Our software incorporates a boosting algorithm [3] which is originally a machine learning approach but was later adapted to estimate statistical models [5]. An inherent advantage of these statistical boosting algorithms [6] is that they (i) can deal with high-dimensional data, (ii) are able to simultaneously select the most influential predictors from a potentially large amount of candidate variables, (iii) still yield statistical models that are in the same way interpretable as if they were estimated via classical approaches and (iv) allow to incorporate different type of predictor effects (e.g., linear, non-linear, spatial).

With the betaboost package, we provide a powerful tool for the analysis of bounded outcomes while incorporating automated data-driven variable selection.

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-University Erlangen-Nuremberg, Germany
[2] Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Germany
[3] Section Biostatistics, Paul-Ehrlich-Institut, Langen, Germany
[4] Department of Nephrology and Hypertension, Friedrich-Alexander-University Erlangen-Nuremberg, Germany

andreas.mayr@fau.de

References

1 Hunger M, Doring A, Holle R. Longitudinal beta regression models for analyzing health-related quality of life scores over time. BMC medical research methodology 2012 Sep 17;12.

2 Kao RR, Haydon DT, Lycett SJ, Murcia PR. Supersize me: how whole-genome sequencing and big data are transforming epidemiology. Trends Microbiol 2014 May;22(5): 282-91.

3 Schmid M, Wickler F, Maloney KO, Mitchell R, Fenske N, Mayr A. Boosted Beta Regression. Plos One 2013 Apr 23;8(4).

4 Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. J R Stat Soc C-Appl 2005;54: 507-44.

5 Mayr A, Binder H, Gefeller O, Schmid M. The Evolution of Boosting Algorithms From Machine Learning to Statistical Modelling. Method Inform Med 2014;53(6): 419-27.

6 Mayr A, Fenske N, Hofner B, Kneib T, Schmid M. Generalized additive models for location, scale and shape for high dimensional dataua flexible approach based on boosting. J R Stat Soc C-Appl 2012;61: 403-27.

Variational Inference for high dimentional factor copulas

Hoang Nguyen[1] *, Concepcion Ausin*[1] *, Pedro Galeano*[1]

Vine copula is a graphical tool used to derive the dependence structure of variables. Based on coupling the bivariate and modelling the relationship, the vine generalizes a tree structure among them [1]. However, the number of parameters as well as the form of tree vine becomes explosive in high dimensional setting. Alternatively, factor copulas ([2], [3])as the truncated C-vines rooted at the latent variables are proposed for tackling the problem. As in any factor model setting, it is assumed that each variables is affected by some common latent factors. We employ the Variational Bayesian approximation [4] to estimate the different specifications of the factor copula models. Conditional on the latent factors, variables become independent which allows the algorithm to run in a parallel setting. Besides, the independence assumption of the latent models also reduces the computational burden for the conditional posterior distribution. We could also choose the best type of copulas using the model selection criteria. The algorithm is fast implemented in C++ and the result is easy extracted to R.

References

1 K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. Insurance: Mathematics and economics, 44(2):182–198, 2009.

2 P. Krupskii and H. Joe. Factor copula models for multivariate data. Journal of Multivariate Analysis, 120:85–101, 2013.

3 P. Krupskii and H. Joe. Structured factor copula models: Theory, inference and computation. Journal of Multivariate Analysis, 138:53–73, 2015.

4 A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. Journal of Machine Learning Research, arXiv preprint arXiv:1603.00788, 2016.

[1] Universidad Carlos III de Madrid - Statistics Department

hoang.nguyen@uc3m.es, concepcion.ausin@uc3m.es, pedro.galeano@uc3m.es

Variable Selection and Allocation in Joint Models for Longitudinal and Time-to-Event Data via Boosting

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

Joint Models for longitudinal and time-to-event data have gained a lot of attention in the last few years as they are a helpful technique to approach a data structure common in clinical studies where longitudinal outcomes are recorded alongside event times [1]. Those two processes are often linked and the two outcomes should thus be modeled jointly in order to prevent the potential bias introduced by independent modelling. Commonly, joint models are estimated in likelihood based expectation maximization or Bayesian approaches using frameworks where variable selection is problematic and which do not immediately work for high-dimensional data. A boosting algorithm rendered possible the selection of covariates even in high-dimensional settings [2]. This contribution extends this algorithm to the automated variable allocation. This approach is necessary when there is no prior knowledge available on the question which dependent variable the individual covariates have influence on.

References

1 Faucett, C. and Thomas, D. (1996). Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: a Gibbs Sampling Approach. Statistics in Medicine, 15, pp 1663 – 1685.

2 Waldmann, E., Taylor-Robinson D., Klein N., Kneib T., Pressler T., Schmid, M. and Mayr, A. (2016): Boosting Joint Models for Longitudinal and Time-to-Event Data. Biometrical Journal. online first

[1] Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg
[2] Institut für Statistik, Ludwig-Maximilians-Universität München

elisabeth.waldmann@fau.de

Model-Based Recursive Partitioning for Stratified and Personalised Treatment Effect Estimation

Heidi Seibold[1], Achim Zeileis[2], Torsten Hothorn[1]

Established statistical methods for the analysis of randomised experiments with two treatment groups estimate a universal (i.e. constant) treatment effect that applies to all subjects observed and – even more importantly – to all future subjects. Common use cases are clinical trials where patients are assigned to standard care vs. a new treatment or A/B testing where website users see different versions of a website.
We propose model-based trees [1] and model-based random forests [2] as a way to relax the assumption that all subjects have the same treatment effect and estimate stratified or personalised treatment effects that depend on characteristics of the subject, e.g. the biomarkers of a patient or the user history of a website user.
The R package model4you [3] provides a simple user interface that allows the user to define the model estimating the overall treatment effect and then partition this model. The model is easy to define, since it is the model that has been used historically and for clinical trials it is already defined in the study protocol. The stratified and personalised models can be visualised and are easy to interpret.

References

1 H. Seibold, A. Zeileis, T. Hothorn. Model-based Recursive Partitioning for Subgroup Analyses. International Journal of Biostatistics, 2016; 12(1), 45-63. doi: 10.1515/ijb-2015-0032.

2 H. Seibold, A. Zeileis, T. Hothorn. Individual Treatment Effect Prediction for Amyotrophic Lateral Sclerosis Patients. Statistical Methods in Medical Research, 2017; online first. doi: 10.1177/0962280217693034.

3 `https://r-forge.r-project.org/R/?group_id=261`

[1] Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland
[2] Faculty of Economics and Statistics, University of Innsbruck, Austria

`heidi.seibold@uzh.ch`

Incorporation of historical data in clinical trials using robust prior

Katharina Hees[1] *and Meinhard Kieser*[1]

Recruiting sufficient patients within an acceptable time horizon is an issue for most clinical trials and is especially challenging in the field of rare diseases. It is therefore an attractive option to include historical data from previous (pilot) trials in the analysis of the current study thus reducing the recruitment burden. Various Bayesian methods for the incorporation of historical information in present trials have been proposed in the literature. In case that the current data match sufficiently well with the historical data, these approaches lead to increased power. However, if this assumption is not met, the gain in power may be much smaller than expected while at the same time a type I error inflation occurs. Therefore, so-called robust prior distributions are well-suited since in case of a prior-data conflict they down-weight the extent to which the historical data is incorporated. When planning the sample size for trials incorporating historical data, not only the type I error rate, the power, and the treatment group difference but additionally the variance and the weight of the historical data have to be specified. However, there is usually some uncertainty in the planning phase about the value of these nuisance parameters. We present methods for blinded and unblinded sample size recalculation in the setting of two-arm superiority trials with historical control data where the variance – and in the unblinded setting additionally the extent to which the historical information is incorporated – is estimated mid-course and the sample size is recalculated accordingly. The operating characteristics of these methods are investigated in terms of actual type I error rate, power, and expected sample size. Application is illustrated with a clinical trial example in patients with systemic sclerosis, a rare connective tissue disorder.

[1] Universitätsklinikum Heidelberg, Im Neuenheimer Feld 130.3, 69120 Heidelberg

hees@imbi.uni-heidelberg.de

Supervised and unsupervised machine-learning methods for pain research

Jörn Lötsch[1,2] and Alfred Ultsch[3]

Along with the increasing molecular and clinical knowledge pathomechanisms of disease, the data acquired during biomedical research become increasingly complex. This poses challenges on the bioinformatical analytics that are increasingly accommodated by current developments in machine learning. In the present study, a complex data set acquired during phenotype assessment of human sensitivity to pain was analyzed. Local UV-B irradiation and capsaicin application as experimental models of heat hyperalgesia of the skin were assessed in 82 healthy subjects using a variety of noxious stimuli. The hypersensitization effects were assessed with a clinically established quantitative sensory testing (QST) battery [1]. Supervised machine-learned analysis implemented as random forests [2] followed by computed ABC analysis [3] pointed at heat pain thresholds as the most relevantly affected of the nine QST parameters. Decision tree analysis of the CART style [4] indicated that in particular UV-B additionally modulated the sensitivity to cold. The emergent self-organizing feature map (ESOM) [5] was used as unsupervised machine-learning technique to identify pain response subgroups. It was possible to present a whole biomedical study on a Kohonen map [5] enhanced by a U-matrix [6]. The results were biologically plausible with respect to both, underlying molecular mechanisms and sex differences in pain sensitivity. The analysis of a complete biomedical study was possible by combining supervised and unsupervised machine-learned methods.

[1] Goethe - University, Institute of Clinical Pharmacology, Theodor – Stern - Kai 7, 60590 Frankfurt am Main, Germany
[2] Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Project Group Translational Medicine and Pharmacology TMP, Theodor – Stern - Kai 7, 60596 Frankfurt am Main
[3] DataBionics Research Group, University of Marburg, Hans – Meerwein – Straße 6, 35032 Marburg, Germany

j.loetsch@em.uni-frankfurt.de, ultsch@Mathematik.Uni-Marburg.de

References

1 Rolke, R., Baron, R., Maier, C., Tölle, T.R., Treede, R.D., Beyer, A., Binder, A. et al.: Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain 123, 231-243 (2006)

2 Breiman, L.: Random Forests. Mach. Learn. 45, 5-32 (2001)

3 Ultsch, A., Lötsch, J.: Computed ABC Analysis for Rational Selection of Most Informative Variables in Multivariate Data. PLoS One 10, e0129767 (2015)

4 Breimann, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman and Hall, Boca Raton (1993)

5 Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol Cybernet 43, 59-69 (1982)

6 Ultsch, A., Lötsch, J.: Machine-learned cluster identification in high-dimensional data. Journal of biomedical informatics 66, 95-104 (2017)

How feature selection in multi-class classifier systems affects the prediction of diagnostic phenotypes

Robin Szekely[1,2], Ludwig Lausser[1], Lyn-Rouven Schirra[1,3] and Hans A. Kestler[1]

Machine learning algorithms for high-dimensional molecular profiles are often modified by internal feature selection processes. The corresponding feature signature can stratify the training process of a classifier and often provides insight in the background of a specific diagnostic task. By identifying various molecular causes of a disease, feature selection can be regarded as a basis for modern analysis methods such as being used in precision medicine or personalized medicine [1].
We compare two schemes for incorporating feature selection processes in multi-class classifier systems. Both schemes operate on the level of the individual base classifier of a classifier system. They are analyzed in combination with the well known One-against-One and One-against-All strategies [3]. The stability of feature selections [2] and the overall generalization ability are evaluated in 10×10 cross-validation experiments with a linear support vector machine on eight publicly available microarray multi-class datasets.

References

1 Jameson J. L., Longo D. L. (2015). Precision medicine - personalized, problematic, and promising. The New England Journal of Medicine 372(23), 2229-2234

2 Lausser L., Müssel C., Kestler H. A. (2013). Measuring and visualizing the stability of biomarker selection techniques. Computational Statistics 28(1), 51-65

3 Schirra L. R., Schmid F., Kestler H. A., Lausser L. (2016). Interpretable classifiers in precision medicine: Feature selection and multi-class categorization. In: Schwenker F, Abbas HM, El Gayar N, Trentin E (eds) Artificial Neural Networks in Pattern Recognition, ANNPR 2016, Springer, Lecture Notes in Artificial Intelligence, vol 9896, pp 105-116

[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 Number Theory and Probability Theory, Ulm University, 89081 Ulm, Germany

robin.szekely@uni-ulm.de

A multi-criteria approach to find predictive and sparse models with stable feature selection for highdimensional data

Andrea Bommert[1], Jörg Rahnenführer[1], and Michel Lang[1]

Finding a good predictive model for a high-dimensional data set is often challenging. For genetics' data, it is not only important to find a model with high predictive accuracy, but it is also important that this model uses only few features and that the selection of these features is stable. This is because in bioinformatics' applications, the models are used not only for prediction but also for drawing biological conclusions, which makes the interpretability and reliability of the models crucial. We suggest using three criteria when fitting a predictive model to a high-dimensional data set: the classification accuracy, the stability of the feature selection, and the number of chosen features.
As it is unclear which measure is best for evaluating the stability of a model, we first compare a variety of stability measures. We find that for the stability assessment behaviour of these measures it is most important if a measure contains a correction term for large numbers of chosen features. While the uncorrected measures show a very similar stability assessment behaviour, the results for the corrected measures differ noticeably.
Then, we analyse Pareto fronts to find models that perform well considering all three target criteria. We conclude that it is possible to find models with a stable selection of few features without losing much predictive accuracy compared to models fitted only considering the classification accuracy.

[1] TU Dortmund University, 44221 Dortmund, Germany

bommert@statistik.tu-dortmund.de, rahnenfuehrer@statistik.tu-dortmund.de, lang@statistik.tu-dortmund.de

Adaptive penalization for high-dimensional regression and classification with non-exchangeable features

Britta Velten[1] *, Wolfgang Huber*[1]

Penalization schemes like LASSO [1] or Ridge regression are routinely used to regress a response of interest on a high-dimensional set of features. Commonly used approaches assume that the features are exchangeable: the same penalty factor is used for each model coefficient. In many applications, however, additional information is available about the features. Such information can include structural knowledge (e.g., feature sets comprising multiple data types and data qualities, such as in biology: transcriptome, genome, epigenome) and/or different prior probabilities for different feature classes (e.g., based on gene or pathway annotation or prior studies). We present a hierarchical Bayesian model that enables differentially penalizing groups of features based on external covariates and adapts the penalty to the information content of each group in a data-driven way. In an application to drug response prediction for cancer patients from multiple 'omic data types [2], the method identifies meaningful differences between 'omic data types. Using available covariates extends the range of applications of penalized regression, improves model interpretability and can improve prediction performance.

References

1 Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, 1996

2 Dietrich, Oles, Sellner et al, Drug Perturbation Based Stratification of Blood Cancer, submitted

[1] European Molecular Biology Laboratory (EMBL), Meyerhofstr. 1, 69117 Heidelberg, Germany

britta.velten@embl.de, whuber@embl.de

Variable selection for model-based gradient boosting using random probes

Tobias Hepp[1], Janek Thomas[2], Andreas Mayr[1,2], Bernd Bischl[2]

Variable selection in regularized regression models like the *lasso* or *gradient boosting* algorithms is usually controlled by method-specific tuning-parameters that define the degree of penalization. While these parameters are commonly determined using resampling strategies like cross-validation, bootstrapping and similar methods, their focus on minimizing the prediction error often results in the selection of many variables without true effect on the outcome [1].
Therefore, we propose a new method to determine the optimal number of iterations in model-based boosting for variable selection [2] inspired by *probing*, a method used in related areas of machine learning research [3]. The general notion of probing involves the artificial inflation of the data with random noise variables, so-called *probes* or *shadow variables*. Using the first selection of a shadow variable as stopping criterion, the algorithm is applied only once without the need to optimize any hyperparameters in order to extract a set of informative variables from the data, thereby making its application very fast and simple in practice. Furthermore, simulation studies show that the resulting models tend to be more strictly regularized compared to the ones resulting from cross-validation, thereby substantially reducing the high number of false discoveries.

References

1 Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics.* 34(3), 1436-1462.

2 Thomas J., Hepp T., Mayr A. and Bischl B. (2017). Probing for sparse and fast variable selection with model-based boosting. *Computational and Mathematical Methods in Medicine,* accepted

3 Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research.* 3(Mar), 1157-1182.

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg
[2] Department of Statistics, Ludwig-Maximilians-Universität München

tobias.hepp@uk-erlangen.de

High-Dimensional Variable Selection via Low-Dimensional Adaptive Learning

Christian Staerk[1] *, Maria Kateri*[1] *and Ioannis Ntzoufras*[2]

Nowadays one often faces the challenging scenario in which high-dimensional data is observed. In a high-dimensional setting, where the number of explanatory variables p is possibly much larger than the sample size n, one is particularly interested in variable selection in order to identify a sparse and interpretable model.
In this talk we give a short and selective overview of variable selection methods which have been proposed in the recent past. In particular, we contrast ℓ_1 and ℓ_0-type regularization methods. While ℓ_1-type methods like the Lasso [1] are computational efficient due to their convexity, they typically require quite strong conditions on the design matrix for variable selection consistency. On the other hand, ℓ_0-type methods like the recently proposed Extended Bayesian Information Criterion [2] can be shown to be variable selection consistent under mild conditions, but they lead to combinatorial and in general NP-hard optimization problems.
We therefore propose an Adaptive Subspace method (AdaSub) which aims at identifying the best model with respect to an ℓ_0-type selection criterion [3], [4]. AdaSub is based on the idea of adaptively solving low-dimensional sub-problems in order to provide a solution to the original high-dimensional problem. We show that the particular form of adaptive learning can be motivated in a Bayesian way. Furthermore, we analyse the limiting properties of AdaSub. In particular, we show that AdaSub converges to the best model according to the criterion used, provided that the so-called ordered importance property (OIP) is satisfied.
In a simulation study we compare the performance of AdaSub with other well-known variable selection methods. Finally we illustrate the application of AdaSub on high-dimensional real data examples with ten thousands of explanatory variables.

[1] Institute of Statistics, RWTH Aachen University, Germany
[2] Department of Statistics, Athens University of Economics and Business, Greece

christian.staerk@rwth-aachen.de, maria.kateri@rwth-aachen.de, ntzoufras@aueb.gr

References

1 Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", J. R. Statist. Soc. B, 58(1), 267–288.

2 Chen, J. and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces", Biometrika, 95(3), 759–771.

3 Staerk, C., Kateri, M. and Ntzoufras, I. (2016), "An Adaptive Subspace Method for High-Dimensional Variable Selection", Proceedings of the 31st International Workshop on Statistical Mod- elling, Rennes, France, ed. by J. Dupuy and J. Josse, 295–300.

4 Staerk, C., Kateri, M. and Ntzoufras, I. (2017), "High-Dimensional Variable Selection via Low-Dimensional Adaptive Learning", submitted.

Using generalized additive models for CNV detection on multi gene panels

Corinna Ernst[1] , Eric Hahnen[1] , Andreas Beyer[2] and Rita K. Schmutzler[1]

Targeted sequencing, which is restricted to the exons of genes known or assumed to be implicated in a special phenotype, decreases costs, storage requirements, and computation times significantly in comparison to whole genome and whole exome approaches. Hence, so-called multi gene panel approaches have become a widely-used tool in clinical diagnostics. Targeted sequencing data is typically characterized by strong biases based on local mappability, GC-content, and further factors affecting capture efficiency. Recent studies revealed that existing tools for copy number variation (CNV) detection on targeted sequencing data – which are mainly designed for the purposes of whole exome approaches exclusively – are not fully able to face these difficulties as they show a notable lack of accuracy and robustness.

We present an approach for CNV detection which is tailored to the challenges of multi gene panel analysis. Our method relies on a generalized additive model (GAM), which models obsrved read count frequencies as a product of two smooth functions. Input data is assumed to consist of mapped reads originating from m, $m \geq 30$ samples which are captured on the same gene panel, and to be re-aligned around indels and filtered for duplicates. Inter-sample normalization occurs position-wise as proposed by Anders and Huber [1] for the aim of RNA-seq data normalization.

Genome-wide generalized additive models (GAMs) have recently been shown to comprise a powerful tool for the identification of ChIP-Seq peaks and genomic regions of aberrant methylation [2]. We present a GAM which models the mean of observed read counts as a product of two smooth functions, namely, a generic background function that contributes to all m samples and a sample-specific smooth function. The latter function is used for final CNV calling. It is assumed to deviate significantly from zero in case a CNV exists.

We validated our approach on 96 samples that were analyzed on the diagnostic TruRisk^{TM} gene panel comprising 48 genes known or assumed to be implicated in hereditary breast and/or ovarian cancer. We compared the performance of our method to the performance of two other tools adapted to CNV analysis on targeted sequencing data, namely, CnvHunter [3] and ExomeDepth [4]. Evaluation revealed that our approach achieves sensitivities and specificities higher or close to the values achieved by existing tools.

[1] Center for Familial Breast and Ovarian Cancer, Center for Integated Oncology (CIO), Medical Faculty, University Hospital Cologne, Cologne, Germany
[2] Cellular Networks and Systems Biology (CECAD), University of Cologne, Cologne, Germany

corinna.ernst@uk-koeln.de

References

1 S Anders and W Huber. Differential expression analysis for sequence count data. Genome Biology, 11(10), 2010.

2 G Stricker, A Engelhardt, D Schulz, M Schmid, A Tresch, and J Gagneur. Genome-wide generalized additive models. Technical Report biorxiv;047464v1, 2016.

3 https://github.com/imgag/ngs-bits.

4 V Plagnol, J Curtis, M Epstein, KY Mok, E Stebbings, S Grigoriadou, NW Wood, S Hambleton, SO Burns, AJ Thrasher, D Kumararatne, R Doffinger, and S Nejentsev. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics, 28:2747–2754, 2012.

High-dimensional genome-wide association studies (GWAS): A field for data restructuring?

Pascal Schlosser[1] *and Anna Kottgen*[1]

With transcriptomics, proteomics, metabolomics and other Omics reaching large population based cohorts, the field of GWAS reached a new dimension. Instead of one GWAS of a clinical outcome, a study might perform hundreds of thousands of independent GWAS with Omics-like outcomes, which further have unknown and complex covariance structures.
In the context of metabolomics, we propose to restructure outcomes by an unsupervised clustering approach. Thereby, the multiple testing burden can be strongly reduced due to existing correlation structures. Moreover, technical noise can be reduced under the assumption that it is independent of the metabolites given their asigned cluster.
In the German Chronic Kidney Disease (GCKD) study, 1221 European ancestry individuals were genotyped with the Illumina Omni2.5exome chip followed by HRC imputation, and had 1382 metabolites quantified in urine using the HD4 Metabolon platform.
Many potentially genetically regulated biological processes can be described by ratios and other elementary terms of these metabolites. Using the set of metabolites and all their probable combinations, we first identify modules by several clustering techniques. These include hierarchical clustering, Netboost, a boosting-based filtered version of hierarchical clustering, and k-means clustering. Subsequently GWAS are performed with summary statistics of modules as outcomes. Principal component measures as well as basic means and medians are considered.
We evaluate the presented approaches by their statistical performance using positive controls and cross validation. Further benchmarking criteria are the identification of known biological pathways and the potential to de-orphanize unnamed metabolites.

[1] Division of Genetic Epidemiology, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany

schlosser@imbi.uni-freiburg.de

Accelerating the Detection of Ordinal Structures

Lyn-Rouven Schirra[1,2], Ludwig Lausser[1], Florian Schmid[1] and Hans A. Kestler[1]

Supervised classification algorithms are mainly designed for the identification of patterns that allow an accurate discrimination of semantically meaningful classes. They typically operate on an abstract level in which classes (verbal concepts or categories) are replaced by nominal variables without taking into account their interpretation. Knowledge about the semantic relationships between the analyzed concepts are typically neglected.

One exception to this scheme are ordinal classifiers, which take into account an assumed ordering of the classes (e.g. *small* $\prec$ *medium* $\prec$ *large*). The ordering is treated as external domain knowledge and will guide the training process of the classifier system. We have previously shown that ordinal classifier cascades can be used for screening for unknown ordinal relationships [1].

In this study we introduce a novel enhancement of the aforementioned method. The new Cascades algorithm can significantly reduce the computational effort to detect ordinal structures in the data. By utilizing upper bounds of classwise sensitivities we are able to filter unsuitable cascades at an early stage. Thus, in comparison to the old method, fewer class cascades have to be tested in order to find ordinal cascades.
In this study we examined the ability of the new algorithm to detect ordinal relations in the class structure on artificial generated data. Subsequently we evaluated the method on real publicly available data from different domains [2,3].

References

1 Lattke, R., Lausser, L., Müssel, C., and Kestler, H. A. Detecting Ordinal Class Structures. In *Multiple Classifier Systems. MCS 2015*, F. Schwenker, F. Roli, and J. Kittler, Eds., no. 9132 in Lecture Notes in Computer Science. Springer, Cham, 2015, pp. 100–111.

2 Lichman, M. UCI Machine Learning Repository `http:/archive.ics.uci.edu/ml`. University of California, School of Information and Computer Sciences, Irvine, 2013.

3 Edgar, R., Domrachev, M., and Lash, A. E. Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. *Nucleic Acids Research* 30, 1 (2002), 207–210.

[1] Institute of Medical Systemsbiology, Ulm University, 89069 Ulm, Germany
[2] Institute of Number Theory and Probability Theory, Ulm University, 89069 Ulm, Germany

`ludwig.lausser@uni-ulm.de, lyn-rouven.schirra@uni-ulm.de, florian.schmid@uni-ulm.de, hans.kestler@uni-ulm.de`

Ordinal Forests: A versatile tool for ordinal regression

Roman Hornung[1]

In many fields of application, including life sciences, it is sometimes of interest to predict the value of an ordinal response variable. However, there exists few ordinal regression methods, that is, regression methods that take the ordinal nature of such variables adequately into account. Ordinal response variables are often treated as nominal variables, applying prediction techniques for binary target variables to all possible pairs of classes of the ordinal response variable.
With Ordinal Forests (OF) we present a method for ordinal regression with highdimensional and low-dimensional data that is able to predict the values of the ordinal target variable for new observations and at the same time estimate the relative widths of the classes of the ordinal target variable. Using a (permutation-based) variable importance measure it is moreover possible to rank the importances of the covariates. The concept of OF is based on the following assumption that also justifies the well-known ordered probit model: There exists a (possibly latent) refined continuous variable y^* underlying the observed ordinal variable y ($y \in 1, ..., J$), where y^* determines the values of y. The functional relationship between y^* and y takes the form of a monotonically increasing step function.
In the context of conditional inference trees, for situations in which the intervals in y^* that define the classes of y are available, Hothorn et al. (2006) suggest to use — as a metric target variable — the midpoints of these intervals. The idea of considering the midpoints of these intervals is used in OF as well. However, importantly, we consider the common situation in which the intervals are not available. The main concept of OF is the following: choose the intervals in y^* corresponding to the different classes of y by maximizing the prediction performance of a standard regression forest that uses the interval midpoints. Subsequently, construct a regression forest that uses the midpoints of the optimized intervals. The performance of OF is assessed using both real and simulated data. OF is implemented in our R package `ordinalForest` available on CRAN.

References

1 Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: a conditional inference framework. Journal of Computational and Graphical Statistics, 15:651-674.

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, Munich 81377, Germany

`hornung@ibe.med.uni-muenchen.de`

A Common Framework and Software Package for the Inclusion and Selection of Covariates in Bradley-Terry Models

Gunther Schauberger[1]

In paired comparison models, the inclusion of covariates can be an interesting tool to account for the heterogeneity of preferences and to investigate which characteristics determine the preferences of certain subjects for the compared objects. Nevertheless, the inclusion of covariates is rarely treated and most existing methods are restricted to only few variables. The main topic of this talk is to propose a coherent framework that allows for the inclusion of all possible types of covariates. There are three different types of covariates that can occur in paired comparisons, the covariates can either vary over the subjects, the objects or both the subjects and the objects of the paired comparisons. For each type of covariate, appropriate penalty terms are proposed. The penalty terms allow for sparser models and facilitate the interpretation of the covariate effects. For example, fusion penalties are applied to find clusters of objects with equal covariate effects or to eliminate covariates from the model. The whole framework is implemented in the R-package BTLLasso [3]. The framework is illustrated by different application, for example to German election data [1] or data from the German Bundesliga [2,4].

References

1 Schauberger, Gunther and Tutz, Gerhard (2017): Subject-Specific Modelling of Paired Comparison Data - a Lasso-Type Approach, Statistical Modelling, accepted

2 Schauberger, Gunther and Tutz, Gerhard (2016): Modelling Football Results in the German Bundesliga Using Match-specific Covariates, Department of Statistics, LMU Munchen, Technical Report 197

3 Gunther Schauberger (2017). BTLLasso: Modelling Heterogeneity in Paired Comparison Data. R package version 0.1-5.

4 Tutz G, Schauberger G (2015). Extended ordered paired comparison models with application to football data from German Bundesliga. Advances in Statistical Analysis, 99(2), 209 - 227

[1] Ludwig-Maximilians-Universität München, Germany

gunther@stat.uni-muenchen.de

Sparse kernel deep stacking networks

Thomas Welchowski[1] *and Matthias Schmid*[1]

Kernel deep stacking networks (KDSNs) [1] are a novel method for supervised learning in biomedical research. Belonging to the class of deep learning methods, KDSNs use multiple layers of non-linear transformations to derive abstractions of the input variables [2]. This architecture can efficiently represent complex nonlinear dependencies in the joint distribution of the inputs and the response variable [3]. While training of deep artificial neural networks usually involves the optimization of a non-convex problem, often implying local optima and slow convergence, KDSNs are characterized by an efficient fitting procedure that is based on a series of kernel ridge regression models with closed-form solutions.
Sparse kernel deep stacking networks (SKDSNs) extend the KDSN framework by tools for variable selection and dropout [4]. Variable selection is included into a two step procedure with the non-linear association measure randomized dependence coefficient [5]. Preselection searches for sets of variables by genetic algorithms. Internal selection is used within tuning to determine an optional cut-off based on rank scores of all remaining variables. Dropout randomly removes parts of the random Fourier transformation matrix, which results in more flexible sparse nonlinear transformations of the input space.
Tuning of SKDSNs is a challenging task due to the multiple hyper-parameters that have to be specified before network fitting. Specifically, we proposed a data-driven tuning strategy for SKDSNs that is based on model-based optimization (MBO) [6]. A meta-model estimates the relationship between the hyper-parameter space and a specified loss function. Simulation studies show that the MBO approach is substantially faster than traditional grid search strategies. Analysis of real data sets demonstrates that MBO-tuned SKDSNs are competetive to other state-of-art machine learning techniques in terms of prediction accuracy. The fitting and tuning procedures are implemented in the R package `kernDeepStackNet`.

[1] Department of Medical Biometry, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität Bonn, Sigmund-Freud-Str. 25, 53127 Bonn,Germany

`welchow@imbie.meb.uni-bonn.de, matthias.schmid@imbie.uni-bonn.de`

References

1 Po-Sen Huang, Li Deng, Mark Hasegawa-Johnson, et al., 2013, Random features for kernel deep convex network. In: Rabab Ward, Li Deng, IEEE International Conference on Acoustics, Speech and Signal Processing, Pages 3143-3147.

2 Li Deng, Dong Yu, 2014, Deep learning: Methods and Applications, Foundations and Trends in Signal Processing, Volume 7, Number 3-4, Pages 197-387

3 Yoshua Bengio, Olivier Delalleau, 2011, On the expressive power of deep architectures. In: Jyrki Kivinen, Csaba Szepesvari, Esko Ukkonen, et al., Algorithmic learning theory, Springer Berlin Heidelberg, Volume 6925, Pages 18-36.

4 Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, et al., 2014, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research, Volume 15, Pages 1929-1958

5 David Lopez-Paz, Philipp Hennig, Bernhard Scholkopf, 2013, The randomized dependence coefficient, arXiv, Article-ID 1304.7717v2

6 Thomas Welchowski, Matthias Schmid, 2016, A framework for parameter estimation and model selection in kernel deep stacking networks, Artificial Intelligence in Medicine, Volume 70, Pages 31–40

The applications of FAM (fail assessment method) in dynamic data analysis

Albasher Shareif[1]

FAM (fail assessment method) is a modern technique that is useful in a decision making process, to evaluate the performance or the day-by-day functionality of objects (individuals, things, persons, resources, products, plants, etc.)[1]. Via the FAM we can use different assessment methods to measure the development and changes of objects and elements to use them via series of steps and integrated systems for analyzing each dynamic element of data depending on its characteristics which can be collected periodically from different sources, then analyzing these waves of dynamic data for the evaluating the performance, limits of fails, development process and other assessments.
This work presents the concept and uses of FAM and its systems -as an assistant method- in big data and dynamic analysis, with a case study about studying the stability of electricity prices for each period of time during one day (24 hours).

References

1 Mohamed Khlafalla Omar, Fail Assessment Method, Wroclaw University of Technology, Wroclaw 2004.

[1] Institute of Mathematics, Pedagogical University of Krakow,Podchorazych 2, 30-084 Krakow, Poland

shareif@up.krakow.pl

Extending joint models in terms of gradient boosting algorithms

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

Joint models turned out to be a powerful approach to analysing data where event times are measured alongside a longitudinal outcome and were first suggested by Wulfsohn and Tsiatis [3]. The idea is to combine a longitudinal and a survival model via a shared predictor ls used in both original models while a parameter α quantifies their relation. Thus the longitudinal part has the form

$$y = \eta_l(x) + \eta_{ls}(x,t) + \epsilon,$$

where η_l denotes a strictly longitudinal predictor and ϵ an error term. On the other hand, the survival part is described by the hazard function

$$\lambda(t|\alpha, \eta_s(x), \eta_{ls}(x,t)) = \lambda_0 \exp(\eta_s(x) + \alpha\eta_{ls}(x,t))$$

with constant baseline hazard λ_0 , a survival predictor η_s and the α-scaled predictor η_{ls} reappearing in the survival submodel. To fit a basic joint model efficiently, Waldmann et al. [2] presented for the case $\eta_s \equiv 0$ a gradient boosting algorithm based on statistical boosting methods for longitudinal data in multiple dimensions discussed in [1]. Aim of this work is to extend the algorithm suggested in [2] by incorporating the predictor η_s, hence a set of covariates, which are independent of the longitudinal structure, is added.

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. (2016): Boosting Joint Models for Longitudinal and Time-to-Event Data. Biometrical Journal. online first

3 Wulfsohn, M., Tsiatis, A. (1997): A joint model for survival and longitudinal data measured with error. Biometrics. 53: 330-339.

[1] Department of Medical Informatics, Biometry and Epidemiology, Waldstraße 6, 91054 Erlangen, Germany

colin.griesbach@fau.de, elisabeth.waldmann@fau.de, andreas.mayr@fau.de

Extensions of Mixture Models for Ordinal Responses

Micha Schneider[1] *and Gerhard Tutz*[1]

Mixture Models provide the opportunity to model discrete human choices as a combination of preference and uncertainty structure. In CUB models [2] the preference is represented by shifted binomial random variables and the uncertainty by a discrete uniform distribution.
We extend this approach to a much wider class of models by using the cumulative model for the preference structure [4] and the beta-binomial distribution for the uncertainty structure [5]. The resulting model allows to distinguish between a tendency to middle categories and a tendency to extreme categories. The model is estimated by using the EM algorithm [1].
It is demonstrated that severe bias might occur if inadvertently the uniform distribution is used to model uncertainty. An application to attitudes on the performance of health services for European countries illustrates the advantages of the more flexible model.

References

1 Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.

2 Iannario, M. and Piccolo, D. (2012). CUB models: Statistical methods and empirical evidence. In: Salini, S. and Kennett, R. (Ed.), Modern Analysis of Customer Surveys: with applications using R, New York: Wiley, 231 – 258.

3 McLachlan, G. J. and Peel, D. (2000). Finite Mixture Models. New York: Wiley.

4 Tutz, G., M. Schneider, M. Iannario, and D. Piccolo(2016). Mixture models for ordinal responses to account for uncertainty of choice. Advances in Data Analysis and Classification, published online.

5 Tutz, G., Schneider, M. (2017): Mixture Models for Ordinal Responses with a Flexible Uncertainty Component, Department of Statistics: Technical Reports, Nr. 203.

[1] Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 Munchen

Micha.Schneider@stat.uni-muenchen.de

Sample Size Calculation with Unequal Group Sizes in Dunnett's Testing

Marcus Vollmer[1]

Background:
Dunnett's T3 procedure[1] is a standard statistical test when comparing multiple treatment groups to the same reference group, e.g. in animal experiments by comparing different immunized mice with a control group. Interestingly, in animal experiments equal sample sizes have been frequently proposed. However, the same statistical power can be achieved by unequally distributed group sizes with a reduction of the total sample size. Currently, two packages are available in R [2, 3] to perform the special testing problem with unequal group sizes. The computation of the p-values includes the consideration of a multidimensional t-distribution and the adjustment for multiple testing. Unfortunately, a procedure for sample size estimation is missing.
Methods & Results:
For sample size estimation, we developed a method to derive unequal group sizes while assuming different effect sizes (different means and unequal variances). The method minimizes the number of animals needed in such experiments while performing Dunnett's T3 procedure. The minimal set of group sizes was derived using a genetic algorithm on Monte Carlo experiments. A necessary statistical power of 80% has been used within our calculations and a practical example is given. Furthermore, for quick assessment in sample size estimation, a table of optimal sets of group sizes has been generated involving different effect sizes and numbers of comparative groups.

References

1 Charles W. Dunnett. Pairwise multiple comparisons in the unequal variance case. Journal of the American Statistical Association, 75(372):796–800, 1980.

2 Torsten Hothorn, Frank Bretz, Peter Westfall, Richard M. Heiberger, Andre Schuetzenmeister, and Susan Scheibe. Package multcomp. Website, 2016. https://cran.r-project.org/package=multcomp.

3 Matthew K Lau. Package DTK. Website, 2013. https://cran.r-project.org/package=DTK.

[1] Institute of Bioinformatics, Universtity Medicine Greifswald

marcus.vollmer@uni-greifswald.de

Mathematical modelling of the behaviour of cofilin-1 in pancreatic cancer

Silke D. Kühlwein[1,2], Julian Schwab[1,2], Malte Buchholz[3], Hans A. Kestler[1]

Mathematical modelling can support biological research by generating hypotheses. Here, moderate size models are faster and cheaper than laboratory experiments and they can give a first hint about phenotypes, cell populations or the outcome of overexpression or knockdown experiments. However, the challenge in modelling biological pathways is that often only qualitative knowledge is available.
Boolean network models are the simplest kind of dynamic models. They can be applied, when only qualitative knowledge is available. Nevertheless, these *in silico* models are able to recapitulate complex biological signalling pathways. The simplicity of Boolean network models arises through the assumption that genes are considered as either expressed (ON/1) or not expressed (OFF/0). All regulatory influences on a gene coming from intensive literature search are summarised in its Boolean function. The set of Boolean functions can further be used to determine the subsequent state of all components within the model [1]. Cyclic recurring states, called attractors, describe the long-term behaviour of the model. In biological context attractors can be related to phenotypes [2].
Pancreatic cancer is one of the most lethal cancer in developed countries with a five year survival rate of less than 5 % [3]. The actin severing protein cofilin-1 is overexpressed in pancreatic cancer. This correlates with high invasivness and poor prognosis [4]. Here, we introduce a Boolean network model describing the behaviour of cofilin-1 in pancreatic cancer cells to identify potential causes leading to the progression and invasiveness of pancreatic cancer.
Biological networks have an error-tolerant architecture with a scale-free topology [5]. Therefore, they are considered to be robust. Computer-intensive studies compared the robustness of the cofilin-1 model with randomly generated models. Thereby, we could show that the cofilin-1 model is more robust against perturbation in Boolean functions as well as perturbations of states. The scale-free topology of biological networks is comparable to a fast concentration of states from different state transitions in a Boolean network and results in a high Gini index. This was the case for the cofilin-1 model and further underlines the biological representative of the cofilin-1 model. In addition, first wet lab experiments could already support our model.

[1] Institute of Medical Systems Biology, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany
[2] International Graduate School in Molecular Medicine, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany
[3] Department of Gastroenterology, Endocrinology and Metabolism, Philipps-University, Hans-Meerwein-Strasse 3, 35043 Marburg, Germany

silke.kuehlwein@uni-ulm.de

References

1 Müssel C., Hopfensitz M., Kestler H.A. Boolnet an R package for generation, reconstruction and analysis of boolean networks. *Bioinformatics* 26, 10 (2010), 1378-1380.

2 Kauffman S. The origins of order. Self-organization and selection of evolution. *Journal of Evolutionary Biology* 7, 4 (1994), 518-519

3 Blum R., Kloog Y. Metabolism addiction in pancreatic cancer. *Cell Death and Disease* 5, e1065 (2014)

4 Satoh M., Takano S., Sogawa K., Yoshitomi H., Isibashi M., Mogushi K., Takizawa H., Otuka M, Shimizu H. Miyazaki M., Nomura F. Immune-complex level of cofilin-1 in sera is associated with cancer progression and poor prognosis in pancreati cancer. *Cancer Science*. (2017)

5 Jeong H., Tombor B., Albert R., Oltvai Z.N., Barabási A.-L. The large-scale organization of metabolic networks. *Nature* 407, (2000)

Simulation of regulatory processes with qualitative networks their acceleration on specilaized hardware

Julian D. Schwab[1,2], Michael Kühl[3], Malte Buchholz[4], Hans A. Kestler[1]

In Systems Biology mathematical models are often used to gain insights into cellular pathways and regulatory networks. If only qualitative knowledge is available, Boolean networks (BN) can provide important insights into the dynamic behavior of complex regulatory systems. BN, introduced by Stuart Kauffman in 1969 [1], describe regulatory processes in biology using two-valued logic. A regulatory factor is represented by a Boolean variable $x \in \{true, false\}$. The dependencies between different regulatory factors can be expressed using Boolean functions (AND, OR, NOT). Synchronous BN are defined by a set of Boolean variables $X = \{x_1, ..., x_n\}$ and a set of Boolean functions $F = \{f_1, ..., f_n\}$ corresponding to the n factors in the modeled regulatory network. The state of a BN at a time t is defined as $\mathbf{x}(t) = (x_1(t), ..., x_n(t))$ - the values of all n factors in the network at that point in time. This results in 2^n possible states of a BN with n regulatory factors [2]. A state transition is performed by applying all transition functions at the same time.

One key aspect in the analysis of BN is the simulation of the systems longterm behavior. State transitions from each state in the network eventually lead to a recurrent cycle of states. These attractors describe the longterm behavior of BNs and are associated to biological phenotypes [2, 3]. To search the network for all attractors, the search algorithm has to be executed from each of the 2^n possible states in the network. This leads to an exponential growth $\mathcal{O}(2^n)$ of computation time and memory consumption for an exhaustive attractor search [4].

Field programmable gate arrays (FPGA) are reprogrammable integrated circuits. These chips are based on configurable logic blocks and programmable interconnects. The logic blocks encapsulate Flip-Flops and lookup table circuits (LUTs) for implementing Boolean functions [5].

We implemented an attractor search algorithm on an FPGA to accelerate the search process. The acceleration is based on 1) the ability to program Boolean functions on register-transfer-level allows for fast state transitions and 2) the FPGAs capability for massive parallelization accelerates the attractor search process in comparison to a software based solution. We report on an acceleration of the attractor search process in orders of magnitude compared to software-based solutions.

[1] Institute of Medical Systems Biology, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany
[2] International Graduate School in Molecular Medicine, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany
[3] Institute of Biochemistry and Molecular Biology, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany
[4] Department of Gastroenterology, Endocrinology and Metabolism, Philipps-University, Hans-Meerwein-Strasse 3, 35043 Marburg, Germany

julian.schwab@uni-ulm.de

References

1 Kauffman, S. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology 22*, 3 (1969), 437 – 467.

2 Kauffman, S. The origins of order. Self-organization and selection in evolution. *Journal of Evolutionary Biology 7*, 4 (1994), 518–519.

3 Thomas, R. Regulatory networks seen as asynchronous automata: a logical description. *Journal of Theoretical Biology 153*, 1 (1991), 1–23.

4 Hopfensitz, M., Müssel, C., Maucher, M., and Kestler, H. A. Attractors in boolean networks: a tutorial. *Computational Statistics 28*, 1 (2013), 19–36.

5 Kuon, I., Tessier, R., and Rose, J. FPGA architecture: Survey and challenges. *Foundations and Trends in Electronic Design Automation*, 2 (2008), 135–253.

Transfer Learning for Invariant Biomarker Selection

Gunnar Völkel[1], Ludwig Lausser[1], Hans A. Kestler[1]

Age-related diseases can be caused by a multi-factorial combination of biomarkers which can only be detected with bioinformatic techniques. The major ingredient in this context are biomarker selection strategies that allow for extracting interpretable decision rules from high-dimensional marker profiles. Selecting valuable biomarkers, these strategies suggest potential hypotheses on the molecular background of a phenotype or disease. Often, these hypotheses will not be unique – several candidate biomarker combinations and explanations exist.

In this work, we provide an ensemble feature selection technique and a parameter tuning setup for its adaptation to different selection problems. The ensemble feature selection technique is based on a genetic algorithm with diversity preserving methods and aims at sparsely overlapping marker combinations. The fitness of an individual marker combination is determined by a correlation-based measure. The final ensembles of marker combinations are evaluated as multi-classifier systems.

Population-based optimization techniques like genetic algorithms allow a large degree of customisation. The choice of suitable operators or parameter values is not evident. In this situation automatic tuning methods are favoured over manual selection. We employ irace for parameter tuning of the genetic algorithm and parallelise the algorithm runs on remote computation servers via the Sputnik library.

Finding a configuration for the genetic algorithm that achieves good performance for the general marker selection problem is a main goal. Hence, the approach is evaluated on multiple cancer datasets using a cross-validation scheme over datasets. The tuning is carried out on the training datasets. The best-found algorithm configuration is used to find and evaluate biomarker combinations on the remaining test datasets in a cross-validation setup.

[1] Institute of Medical Systems Biology, Ulm University, 89081 Ulm, Germany

{gunnar.voelkel,ludwig.lausser,hans.kestler}@uni-ulm.de

References

1 López-Ibáñez, M., Dubois-Lacoste, J., Stützle, T., Birattari, M.: The irace package, Iterated Race for Automatic Algorithm onfiguration. Technical Report TR/IRIDIA/2011-004, IRIDIA, Université Libre de Bruxelles, Belgium, 2011.

2 Völkel, G., Lausser, L., Schmid, F., Kraus, J., Kestler, H. A.: Sputnik: ad hoc distributed computation. Bioinformatics **31**(8), 1298–1301 (2015)

3 Lausser, L., Schmid, F., Schirra, L.-R., Wilhelm, A. F. X., Kestler, H. A.: Rank-based classifiers for extremely high-dimensional gene expression data. Advances in Data Analysis and Classification (2016).

4 Burkovski, A., Schirra, L. R., Schmid, F., Lausser L., Kestler, H. A.: Ordinal prototype-based classifiers. Archives of Datascience Series A (2016), accepted.

Identifying conserved gene signatures among species

Lea Siegle[1,2], Ludwig Lausser[1], Hans A. Kestler[1]

Animal models are typically applied as surrogate test subjects in drug development. An assumption made when experimenting with animals is that the molecular processes of an animal react in similar ways as their human equivalents so the results can be linked back to the molecular processes in humans. Thus, the identification of conserved molecular processes and comparable geno- and phenotypes can improve not only the efficacy of experimental test procedures but also help in linking similar or identical processes (and their modifications under experimental conditions) between different species.
In this project, we analyze gene expression profiles of humans patients and rodent animal models in parallel that show comparable phenotypes. These experiments are interlinked via a combined feature selection process that provides a common gene signature to both individual setups. The gene signature is adapted according to classification experiments on both species. It therefore can be seen as a candidate list of conserved biomarkers.
The feature selection method will be tested on publicly available pairs of microarray datasets [1]. The accuracies achieved in paired 10×10 cross validation experiments and the stability of the identified signatures [2] will be characterized.

References

1 Edgar R, Domrachev M, Lash A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30(1):207-10

2 Lausser L., Müssel C., Maucher M., Kestler H. A. (2013). Measuring and visualizing the stability of biomarker selection techniques. Computational Statistics 28(1), 51-65

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

lea.siegle@uni-ulm.de, ludwig.lausser@uni-ulm.de, hans.kestler@uni-ulm.de

Detecting drug targets by the functional analysis of ligand binding effects on protein thermostability

Dorothee Childs[1], Karsten Bach[2] , Holger Franken[3] , Nils Kurzawa[1], Mikhail Savitski[1] and Wolfgang Huber[1]

Detecting the targets of a drug is an important challenge in drug research. Thermal proteome profiling (TPP) addresses this need by inferring drug engagement from changes in protein thermostability in a high-throughput manner [1, 2]. Until now, TPP data analysis has been limited by the requirement to summarise temperature-dependent denaturation by a single parameter, the melting point. Drug engagement effects are typically identified based on changes in this parameter estimate. However, this strategy misses targets for which treatment effects are not well represented by melting point shifts, or which follow non-monotonic response curves. To overcome these limitations, we present a functional approach that compares curves instead of summary parameters. It projects the data to a space of smooth functions and then infers treatment effects by a moderated F-statistic [3,4,5]. Our method outperformed the traditional parameter-centric approach with regard to specificity and sensitivity on three independent datasets. We could correctly detect cancer drug targets for which ligand binding was not reflected by melting point shifts, or whose melting curves deviated from the commonly applied sigmoid model in the cellular environment. We hope that the proposed approach will aid the detection of novel targets and off-targets of drugs with to date unexplained mechanisms of action or side effects.

References

1 Savitski, M. M. et al. (2014). Tracking cancer drugs in living cells by thermal profiling of the proteome. Science, 346(6205), 1255784–1255784.

2 Franken, H. et al. (2015). Thermal proteome profiling for unbiased identification of drug targets and detection of downstream effectors. Nature Protocols, 10(10), 1567–1593.

3 Ramsey, J. O. and Silverman, B. W. (2005). Functional Data Analysis (2nd ed.)

4 Storey, J. D. et al. (2005). Significance analysis of time course microarray experiments. PNAS, 102(36), 12837–12842.

5 Berk, M. et al. (2011). A statistical framework for biomarker discovery in metabolomic time course data. Bioinformatics, 27(14), 1979–1985.

[1] EMBL Heidelberg + Meyerhofstraße 1, 69117 Heidelberg
[2] University of Cambridge + 11 Tennis Ct Rd, Cambridge CB2 1QJ
[3] Cellzome, GlaxoSmithKline + Meyerhofstraße 1, 69117 Heidelberg

childs@embl.de, kb615@cam.ac.uk, holger.j.franken@gsk.com, nils.kurzawa@embl.de, savitski@embl.de, whuber@embl.de

Cell-type Deconvolution Optimization

Franziska Görtler[1]

Immune infiltrates of tumors affect progression of disease and response to treatment. Thus profiling the cellular composition of immune infiltrates might yield novel biomarkers to support treatment decisions. Statistical deconvolution methods have been suggested to reverse engineer specific cell proportions from gene expression profiles of cell mixtures (Gong et al. 2013, Newman et al. 2015, Altboum et al. 2014)). Given a reference matrix X of methylation profiles from sorted immune cell populations one is interested in and a tissue cell mixture profile y, we have $y = bX + R$, where b is the vector of relative cellular proportions and R a residuum. R integrates signals from cells that are not represented in X, regulation of genes in the immune cells, and measurement noise. Solving for b or minimizing the residuum is straight forward, but does not yield a correct deconvolution because the signals in the residuum should not be expected to be small for the majority of genes. The expression of genes expressed in tumor cells, fibroblasts or any other cell type not covered by X can not be explained by b, nor can the expression of genes that are strongly regulated in immune cells. Therefore we face the challenge to select a small set of genes with the property that they are not expressed in non-modeled cells and constantly expressed in the modeled immune cells. For these genes the residuum R is dominated by unbiased noise and the deconvolution model should yield valid estimates of the immune cell composition of a tumor.
Our approach to gene selection is model based. Gene sets that have the desired properties yield correct predictions of changes of immune cell fractions. The optimal gene set for deconvolution is numericaly calculated by maximizing the correlation between real and calculated cell-type proportions.
For its calculation I use a simulated annealer for optimizing the gene set which is best for deconvolution of the cell-types of interest. Using continuous gene weights between zero, gene is not useful for deconvolution, and one, means gene is very useful, leads to better results compared with just using gene is on or off. Furthermore, we take into account that cells are usually not in isolation. They are embedded in e.g. tissue. This is achieved by simulations that digitally mix cells with bulk. Based on this data we deduce a deconvolution model with improved predictive performance. This we could also validate on experimental data from Altboum et al.

[1]Statistical Bioinformatics, Institute of Functional Genomics, University of Regensburg, Regensburg, Germany

Franziska.Goertler@klinik.uni-regensburg.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
Statistical Computing 2017 - Abstracts der 49. Arbeitstagung

Ulmer Informatik-Berichte
ISSN 0939-5091

Herausgeber:
Universität Ulm
Fakultät für Ingenieurwissenschaften und Informatik
89069 Ulm