ulm university universität
uulm

Statistical Computing 2009

Abstracts der 41. Arbeitstagung

HA Kestler, B Lausen, H Binder

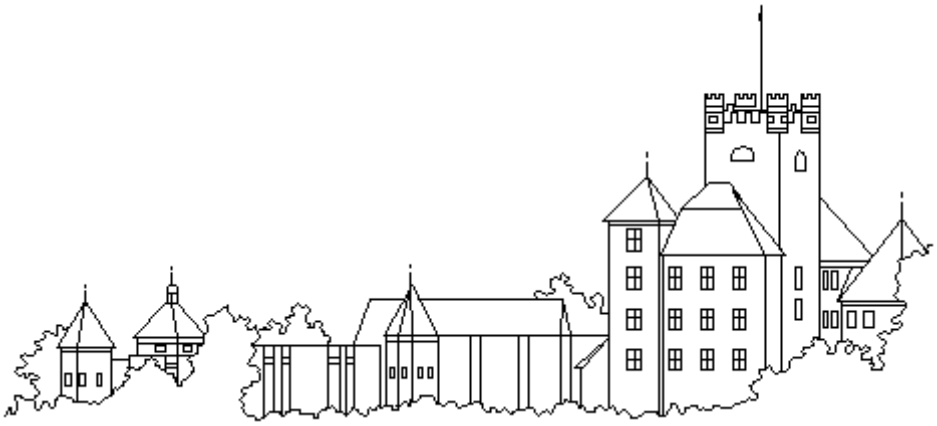
H-P Klenk, F Leisch, M Schmid (eds)

Ulmer Informatik-Berichte

Nr. 2009-07

Juni 2009

Statistical Computing 2009



41. Arbeitstagung

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

28.06.-01.07.2009, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, June 28, 2009

18:15-20:00		**Dinner**
20:00-21:00		**Chair: A. Benner (Heidelberg)**
20:00-21:00	Leonhard Held (Zürich)	INLA in action: Bayesian inference without (MCMC) tears

Monday, June 29, 2009

8:50		**Opening of the workshop**
09:00-12:00		**Chair: H.A. Kestler (Ulm)**
09:00-09:30	Birgit Schrödle (Zürich)	INLA in action: A practical introduction
09:30-10:00	Benjamin Hofner (Erlangen)	Biased model selection: Possible solutions for boosting
10:00-10:30	Nikolay Robinzonov (München)	Boosting techniques for nonlinear time series models
10:30-11:00		**Coffee break**
11:00-11:30	Werner Adler (Erlangen)	Classification of longitudinal data using tree-based ensemble methods
11:30-12:00	Friedhelm Schwenker (Ulm)	Ensemble methods and artificial neural networks for probability density function estimation
12:15-14:00		**Lunch**
14:00-18:00		**Chair: F Leisch (München)**
14:00-15:00	Georg Fuellen (Rostock)	Homology, Phylogeny, Evolution: 'old hats' at the core of biomedical investigation
15:00-15:30	Markus Göker (Braunschweig):	Methods for the phylogenetic inference from whole genome sequences and their use in prokaryote taxonomy
15:30-16:00	Johann Kraus (Ulm)	Multi-core parallelisation using transactional memory: A k-means case study
16:00-16:30		**Coffee break**
16:30-17:00	Markus Schmidberger (München)	State-of-the-art in parallel computing with R
17:00-18:00	Markus Schmidberger (München), Manuel Eugster (München), Christine Porzelius (Freiburg), Jochen Knaus (Freiburg):	Tutorial I: "Parallel Computing with R"
18:15-20:00		**Dinner**
20:00-21:00		Tutorial II: "Parallel Computing with R"

Tuesday, June 30, 2009

09:00-12:00		**Chair: B. Lausen (Essex)**
09:00-09:30	Uwe Ligges (Dortmund)	SVM based Classification of Instruments – Timbre Analysis
09:30-10:00	Wolfgang Lindner (Ulm)	Tolerant Learning with the Set Covering Machine
10:00-10:30	Günther Sawitzki (Heidelberg)	Computational Statistics: A Proposal for a Basic Course
10:30-11:00		**Coffee break**
11:00-11:30	Alfred Ultsch (Marburg)	Methods for the Identification of Differentially Expressed Genes
11:30-12:00	Esmeralda Vicedo (München)	Quality assessment of huge numbers of Affymetrix microarray data
12:15-14:00		**Lunch**
14:00-18:00		**Chair: H. Binder (Freiburg)**
14:00-15:00	Marco Grzegorczyk (Dortmund)	Bayesian networks and their applications in systems biology
15:00-15:30	Lars Kaderali (Heidelberg)	Reconstructing Signaling Pathways from RNAi data using Bayesian Networks and Markov Chain
15:30-16:00	Theodore Alexandrov (Bremen)	Feature extraction and classification in mass spectrometry using sparse coding algorithms
16:00-16:30		**Coffee break**
16:30-18:00		Working groups meeting on **Statistical Computing 2010** and other topics (all welcome)
18:15-20:00		**Dinner**
20:00-21:00		**Poster-Session**
	Stephan Gade (Heidelberg)	Challenges in the quantification and normalization of RPPAs
	Marc Johannes (Heidelberg)	A pipeline for the discovery of alternative splicing events with Affymetrix Exon Arrays

Wednesday, July 1, 2009

09:00-12:00		**Chair: M. Schmid (Erlangen)**
09:00-09:30	Christine Porzelius (Freiburg)	A general, prediction error based criterion for selecting model complexity for high-dimensional survival models
09:30-10:00	Manuela Zucknick (Heidelberg)	Independent screening approaches for Cox models with ultrahigh dimensionality
10:00-10:30	Arthur Allignol (Freiburg)	Empirical transition matrix of multistate models: The etm package
10:30-11:00		**Coffee break**
11:00-11:30	Esther Herberich (München)	Parametric simultaneous inference under test
11:30-12:00	Markus Maucher (Ulm):	On the influence of non-perfect randomness on probabilistic algorithms
12:15-14:00		**Lunch**

INLA in action: Bayesian inference without (MCMC) tears?

Leonhard Held

Abteilung Biostatistik,
University of Zurich,
held@ifspm.uzh.ch

Integrated nested Laplace approximations (INLA) have been recently proposed for approximate Bayesian inference in latent Gaussian models (Rue, Martino and Chopin, 2009, JRSSB). The INLA approach is applicable to a wide range of commonly used statistical models, such as generalized linear mixed models, non- and semiparametric regression as well as spatial and spatio-temporal models. In this talk I will first review the methodology and contrast it with more established inference approaches such as Markov chain Monte Carlo (MCMC), empirical Bayes and hierarchical likelihood. In the second part of the talk I will discuss applications to bivariate random-effects meta-analysis, spatio-temporal disease mapping and age-period-cohort modelling and prediction. This is joint work with Michaela Paul, Andrea Riebler, Havard Rue and Birgit Schrödle.

INLA in action: A practical introduction

Birgit Schrödle

Biostatistics Unit, Institute of Social and Preventive Medicine,
University of Zurich,
birgit.schroedle@ifspm.uzh.ch

So far, Bayesian inference in structured additive regression models was done using Markov chain Monte Carlo techniques, which may be slow and difficult to apply due to highly correlated samples and the inherently large Monte-Carlo error present in MCMC estimates. An alternative approach which was proposed recently (Rue et al., 2009) uses integrated nested Laplace approximations (INLA) to obtain posterior marginals of the included parameters. This method runs remarkably fast concerning computational time and provides very accurate approximations. A C program called inla written by the authors of Rue et al. (2009) is available as an open source tool. From an users point of view there are two ways to run the inla program: Settings of the model, included effects and options for the INLA algorithm can either be specified by a file which is handed directly to the inla program or using the more intuitive R package INLA as an interface. One intention of this talk is to show by means of appropriate examples from practice how inla is run using both options. Additionally, pros and cons of both ways of using inla will be illustrated. A second issue will be to show how and why various options for the INLA algorithm can be set and how the obtained output can be used for a statistical analysis.

References

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, Journal of the Royal Statistical Society: Series B. To appear.

Biased Model Selection: Possible Solutions for Boosting

Benjamin Hofner

Institut für Medizininformatik, Biometrie und Epidemiologie,
Friedrich-Alexander-Universität Erlangen-Nürnberg,
benjamin.hofner@imbe.med.uni-erlangen.de

Variable selection and model choice are of major concern in many applications, especially in high-dimensional settings. Boosting (for an overview see Bühlmann and Hothorn (2007)) is a useful method for model fitting with intrinsic variable selection and model choice. However, a central problem remains: Variable selection is biased if the covariates are of very different nature. An important example is given by models that try to make use of continuous and categorical covariates at the same time. Especially if the number of categories increases, categorical covariates offer an increased flexibility and thus are preferred over continuous covariates (with linear effects). A closely related problem is model choice, where one tries to choose between different modelling alternatives for one covariate. The choice between linear or smooth effects is a classical example. The two competitors have different degrees of freedom (1 df for the linear effect and considerably more than 1 df for the smooth effect). Hence, smooth effects are preferably selected. To make categorical covariates comparable to linear effects in the boosting framework one could use ridge penalized base-learners (i.e, modelling components) with 1 df in this case. To overcome the problem of different degrees of freedom of, e.g., linear and smooth effects Kneib et al. (2008) proposed a model choice scheme which utilizes a decomposition of P-spline base-learners. An empirical evaluation of both approaches making use of the R package mboost (Hothorn et al., 2009) and some insights will be presented.

References

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting, Statistical Science 22: 477–505.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2009). mboost: Model-Based Boosting. R package version 1.1-1.

URL: http://cran.R-project.org/web/packages/mboost

Kneib, T., Hothorn, T. and Tutz, G. (2008). Variable selection and model choice in geoadditive regression models, Biometrics . (accepted).

Boosting Techniques for Nonlinear Time Series Models

Nikolay Robinzonov, Gerhard Tutz and Torsten Hothorn

Institut für Statistik,
Ludwig-Maximilians-Universität München,
Nikolay.Robinzonov@stat.uni-muenchen.de

Many of the popular nonlinear time series models require a priori choice of parametric functions which are assumed to be appropriate in specific situations. This approach is used mainly in financial applications, when there is suffcient knowledge to prespecify the nonlinear structure between the covariates and the response. One principal strategy to investigate a broader class on nonlinear time series is the Nonlinear Additive AutoRegressive (NAAR) model. The NAAR model estimates the lags of a time series as flexible functions in order to detect non-monotone relationships between current observations and past values. We consider two modifications of a numerical optimization rather than a traditional statistical model, called boosting. The first algorithm considers boosting of additive models, built on top of penalized B-Splines. The second strategy is boosting of linear models. Particularly, the componentwise boosting performs a built-in variable selection, as well as a model choice, both of which are facilitated simultaneously. Thus we address the major issues in time series modelling: lag selection and nonlinearity. An extensive simulation study compares the outcomes of boosting to the outcomes, obtained through alternative nonparametric methods. Boosting shows an overall strong performance in terms of precise estimations of highly nonlinear lag functions. The forecasting potential of boosting is examined on real data with target variable the German industrial production (IP). In order to improve the models forecasting quality we supply it with additional information through exogenous variables. Thus we address the second major aspect in this paper which concerns the issue of high-dimensionality in the models, i.e. models with many covariates. Allowing additional inputs in the model extends the NAAR model to an even broader class of models, namely the NAARX model. We show that boosting can cope with large models which have many covariates compared to the number of observations.

Classification of Longitudinal Data Using Tree-Based Ensemble Methods

Werner Adler and Berthold Lausen

Department of Biometry and Epidemiology,
University of Erlangen,
Department of Mathematical Sciences,
University of Essex, United Kingdom,
Werner.Adler@imbe.med.uni-erlangen.de

In many medical applications, longitudinal data sets are available. These data show a dependency structure similar to examinations of paired organs, e.g. eyes. Adler et al. (2009) examined various bootstrapping schemes for ensemble methods based on classification trees in the situation of paired data. These schemes have shown to improve the classification performance compared to the traditional approach, where only one observation per subject is used. We investigate the performance of the proposed methods in the situation of longitudinal data and extend the methodology to the situation, where more than two observations per individual are available. Furthermore, we account for the temporal aspect of the data by weighting the observations according to their examination dates. The examined clinical data set consists of morphological examinations of eyes of glaucoma patients and healthy controls over a time period of up to thirteen years. The performance of our modified classifier is evaluated by bootstrap based ROC analysis (Adler & Lausen 2009). As a reference baseline, we examine the performance of bagging classification trees using only the newest valid examination of the subjects.

References

Adler, W., Brenning, A., Potapov, S., Schmid, M. and Lausen, B. (2009): Ensemble classification of paired data. submitted.

Adler, W., and Lausen, B. (2009): Bootstrap estimated true and false positive rates and ROC curve. Computational Statistics & Data Analysis, 53(3), 718-729.

Ensemble methods and artificial neural networks for probability density function estimation

Friedhelm Schwenker

Institute of Neural Information Processing,
University of Ulm,
friedhelm.schwenker@uni-ulm.de

In this paper probability density function (PDF) estimation is considered. PDF estimation is an important field in reasearch in statistics and pattern recognition. Many different approaches have been made to solve such problems, e.g. parametric methods have been applied in cases where an assumption on the underlying PDF can be made. Then the goal is to estimate (the few) parameters of this assummed PDF. Nonparametric PDF estimation is different, here the PDF is assumed to be unknown but is from a family of more general functions, such as polynomials, kernel density functions, functions defined by artificial neural nets (ANN) etc. Ususally a large set of parameters must be computed to achieve sufficiently accurate estimates. Here we focus on nonparamteric estimation of PDF utilizing ANN models, these models are trained by ensembles of kernel density estimators. The behaviour of this approach is demonstrated by numerical experiments on the approximation one- dimensional PDF, and in addition an elementary proof of the general Haar approximation property is shown for Gaussian kernel density estimators.

Homology, Phylogeny, Evolution: 'old hats' at the core of biomedical investigation

Georg Fuellen

Institut für Biostatistik und Informatik in Medizin und Alternsforschung,
Medizinische Fakultät Universität Rostock,
fuellen@alum.mit.edu

The analysis of the ever-increasing amount of biological and biomedical data can be pushed forward by comparing the data within and among species. For example, an integrative analysis of data from the genome sequencing projects for various species traces the evolution of the genomes and identifies conserved and innovative parts. Here, I review the foundations and advantages of this "historical" approach and evaluate recent attempts at automating such analyses. Biological data is comparable if a common origin exists (homology), as is the case for members of a gene family originating via duplication of an ancestral gene. If the family has relatives in other species, we can assume that the ancestral gene was present in the ancestral species from which all the other species evolved. In particular, describing the relationships among the duplicated biological sequences found in the various species is often possible by a phylogeny, which is more informative than homology statements. Detecting and elaborating on common origins may answer how certain biological sequences developed, and predict what sequences are in a particular species and what their function is. Such knowledge transfer from sequences in one species to the homologous sequences of the other is based on the principle of 'my closest relative looks and behaves like I do', often referred to as 'guilt by association'. To enable knowledge transfer on a large scale, several automated 'phylogenomics pipelines' have been developed in recent years, and seven of these will be described and compared. Overall, the examples in this review demonstrate that homology and phylogeny analyses, done on a large (and automated) scale, can give insights into function in biology and biomedicine... Up to this point, the abstract text you're reading is taken from a recent review, it's an old hat! But based on the foundation just described, I will outline some novel work, developments and insights, aimed at investigating (systematically) the transferability of results from model species to man, and the implications of SNP analyses to clinical studies. That the latter has a lot to do with Homology, Phylogeny & Evolution, that's just a hypothesis I have at the time of writing

this text. How much all this has to do with statistics I cannot estimate with confidence. I'm usually feeling uneasy whenever I read or hear about statistics that goes beyond the very basic, but I admire those who do not have these problems. And I assume statistics is useful always everywhere, so I'm keen on feedback!

Methods for the phylogenetic inference from whole genome sequences and their use in Prokaryote taxonomy

Markus Göker, Alexander F. Auch, Mathias von Jan
and Hans-Peter Klenk

DSMZ – German Collection of Microorganisms and Cell Cultures GmbH,
Braunschweig,
Center for Bioinformatics,
Eberhard Karls University of Tübingen,
markus.goeker@dsmz.de

Since the pioneering work of Carl Woese and his coworkers in the late Seventies, three main branches of the tree of life have been established: Bacteria, Archaea, and Eukaryotes. While the Prokaryotes (Bacteria and Archaea) belong to the most inconspicuous living beings on earth, they by far outclass the Eukaryotes regarding their biochemical abilities, and most likely also regarding the total number of species and the total biomass. Because of their small size and the frequent lack of morphological distinctions, the classification of Prokaryotes is dominated by chemotaxonomical and molecular techniques and has been considerably standardized. For instance, to establish a new Prokaryotic species it is required to demonstrate that the similarity between its genomic DNA and the DNA of closely related reference species (represented by their type strains), as inferred using DNA-DNA hybridization methods, is lower than 70%. However, this technique is cumbersome and cannot easily be made reproducible and thus is currently carried out in only a few molecular labs in the world. Due to the recent staggering advances in DNA sequencing technology, Prokaryotic genomes can be obtained in steadily decreasing time and at steadily decreasing costs. Hence, it is likely that routine sequencing of nearly complete genomes will become an integrated part of biodiversity research on Prokaryotes within the next few years. However, beyond sequencing techniques, two further conditions are necessary to fully replace DNA-DNA hybridization approaches by the comparison of genomes. First, algorithms to calculate distances and similarities between partial or full genome sequences need to be devised that are both reasonably fast and

able to reflect biologically sensible differences. Second, a regularly updated database including the genome sequences all type strains (i.e., taxonomically relevant strains) must be established. We here describe several approaches to infer genome-genome distances and discuss them regarding their correlation with DNA-DNA-hybridization values, their computational speed, and their robustness regarding the use of only partially sequences genomes. A web server to calculate the similarity between reference and query genomes is introduced. We briefly present the GEBA (Genomic Encyclopedia of Bacteria and Archaea) project, which aims at systematically filling the gaps in genome sequencing along the bacterial and archaeal branches of the tree of life, and emphasize the importance of public culture collections to achieve such goals. Finally, we put approaches based on similarity thresholds in the broader context of taxonomic methods and discuss their specific advantages and limitations.

Multi-core parallelization using transactional memory: A K-means case study

Johann M. Kraus and Hans A. Kestler

AG Bioinformatics and Systems Biology,
Institute of Neural Information Processing,
University of Ulm,
hans.kestler@uni-ulm.de

In recent years, the demand for computational power in bioinformatics has increased due to rapidly growing data sets from microarray and other high-throughput technologies. This demand is likely to increase. Currently, the field of bioinformatics is confronted with data sets containing thousands of samples and up to millions of features, e.g. genome-wide association studies using SNP chips, etc. Standard algorithms need to be parallelized for fast processing. Unfortunately, most approaches for parallelizing algorithms require a careful software design and mostly rely on network communication protocols, e.g. Message Passing Interface (MPI). In contrast, a shared memory architecture allows parallelization via threads on a single multi-core computer, where several tasks have access to a shared memory. In shared memory architectures developers have to deal with concurrency control. Simultaneously running threads can process the same data and might also try to change the data in parallel. Concurrency control ensures that software can be parallelized without violating data integrity. Currently, the most popular approach for managing concurrent programs is the use of locks. Locking and synchronizing ensures that changes to the states of the data are coordinated. But implementing thread-safe programs using locks can be fatally error-prone. We outline the process of multi-core parallelization of the K-means cluster algorithm using Clojure. The rationale behind Clojure is combining the industry-standard JVM with functional programming, immutable data structures, and a built-in concurrency support via software transactional memory. This makes it a suitable tool for parallelization and rapid prototyping in many areas. Using the design principle of transactional memory can efficiently improve the performance and maintainability of multi-core applications.

State-of-the-art in Parallel Computing with R

Markus Schmidberger and Ulrich Mansmann

IBE,

Ludwig-Maximilians-Universität Munich,

schmidb@ibe.med.uni-muenchen.de

R is a mature open-source programming language for statistical computing and graphics. Many areas of statistical research are experiencing rapid growth in the size of data sets. Methodological advances drive increased use of simulations. A common approach is to use parallel computing. This presentation introduces to parallel computing and presents an overview of techniques for parallel computing with R on computer clusters, on multi-core systems, and in grid computing. Sixteen different packages were reviewed, comparing them on their state of development, the parallel technology used, as well as on usability, acceptance, and performance. Two packages (snow, Rmpi) stand out as particularly useful for general use on computer clusters. Packages for grid computing are still in development, with only one package currently available to the end user. For multi-core systems four different packages exist, but a number of issues pose challenges to early adopters. The presentation concludes with ideas for further developments in high performance computing with R.

References

M. Schmidberger, M. Morgan, D. Eddelbuettel, H. Yu, L. Tierney, U. Mansmann (2009). State-of-the-art in Parallel Computing with R; Journal of Statistical Software; submitted. Preprint: http://epub.ub. uni-muenchen.de/8991/

SVM based Classification of Instruments - Timbre Analysis

Uwe Ligges and Sebastian Krey

Fakultät für Statistik,

TU Dortmund,

ligges@statistik.tu-dortmund.de

An application of Timbre Analysis, or speaking in terms of statistics, classification of different voices or instruments will be presented. Corresponding methods are used, for example, as tools in (polyphonic) music transcription, by singing teachers and students who try to improve voices, and by modern music recommender systems. The latter are already widely used on server infrastructure that support music listeners who download music from the web to their home computers and even their mobile devices. In timbre classification, after extracting separate tones, variables that contain relevant information of the timbre must be derived or constructed. Typical variables are derived from models in time and particularly frequency domain. We present some additional techniques of data preprocessing and construction of variables that are used in speech recognition. Since variables from very different spaces are used, the resulting classification problem might be rather non-linear. Hence there was some need to try various classification methods with a focus on SVMs with different kernels. As in so many other classification tasks, it turns out that the choice and construction of appropriate variables is much more important than the particular classification method or kernel that is finally used.

Noise-Tolerant Learning with the Set Covering Machine

Wolfgang Lindner and Hans A. Kestler

Institute of Neural Information Processing,
University of Ulm,
Department of Internal Medicine I,
University Hospital Ulm,
hans.kestler@uni-ulm.de

The Set Covering Machine (SCM) introduced by Marchand and Shawe-Taylor (2002) can be regarded as a generalization of Hausslers greedy algorithm for attribute-efficient learning of conjunctions with few boolean variables, where the boolean variables are replaced by an arbitrary (possibly data-dependent) set of features mapping the given examples to boolean values. When used with so-called data-dependent rays (simple threshold functions depending only on a single attribute with thresholds taken from the given dataset) the SCM effectively performs feature selection and produces very simple classiers which admit a meaningful interpretation in the given application context. This is especially important in the clinical setting where the SCM can be applied to micro-array data to distinguish between benign and malign cases. Furthermore it is possible to prove rather tight bounds on the generalization error which in turn can be used as an alternative to cross-validation based model selection or directly as the optimization criterium in the greedy selection step of the SCM algorithm. We consider possibilities to extent the SCM to deal with data corrupted by noise. We give corresponding bounds on the generalization error and apply techniques developed in the context of learning from statistical queries, a popular technique to design noise-tolerant learning algorithms in Valiants PAC-setting.

Computational Statistics: A Proposal for a Basic Course

Günther Sawitzki

StatLab Heidelberg,

Universität Heidelberg,

gs@statlab.uni-heidelberg.de

The course "Computational Statistics: An Introduction to R" has been designed for a mixed audience. Part of the audience comes from applied areas (in particular from clinical departmentsm and from the DKFZ, the German cancer research center), with some working knowledge in statistical methods and with considerable laboratory experience. The other part of the audience are students from mathematics or computer science, with a basic knowledge in (mathematical) stochastics. As one of the participants from the applied field said "We can lookup the methods ourselfes. What we need is a guide to the underlying concepts." The course tries to give a concise introduction to R, together with a selected introduction to basic concepts of statistics. Designed as a five day compact course (or a two hour lecture for one term), the statistical topics are

- distribution diagnostics
- linear models and regression diagnostics
- non-parametric comparisons
- multivariate analysis.

The course may be presented as an introduction to R. But actually it is an invitation to statistical data analysis. The talk will give an outline of the course, and discuss the underlying choices.
See http://sintro.r-forge.r-project.org/

Benchmarking Methods for the Identification of Differentially Expressed Genes

Alfred Ultsch

Databionics,
Universität Marburg,
ultsch@Mathematik.Uni-Marburg.de

The identification of differences in the expression of genes between two groups, e.g. healthy and cancerous patients, is an important task in DNA microarray analysis. Several algorithms for this task are published using different approaches from statistics and machine learning. The result of these algorithms is a set of genes which are supposed to be differentially expressed. The calculation of the correct number of truly differentially expressed genes is a very difficult task. New and surprising hypotheses of the genetic mechanisms which explain the differences between the two groups are sought. Therefore many genes should be considered. With the thousands of measurements on a typical DNA microarray, however, many false positives are also to be expected. A comparison of the algorithms is proposed which is independent of the prediction of the correct number of truly differentially expressed genes.

Quality Assessment of huge numbers of Affymetrix Microarray Data

M. Esmeralda Vicedo Jover, Markus Schmidberger
and Ulrich Mansmann

Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie,
Ludwig-Maximilians-Universität München,
e.vicedo@gmx.de

Microarray experiments have become more popular since the manufactured arrays are cheaper, with more quality and easier to manipulate and therefore the amount of data generated has been incremented in order of hundreds of arrays per experiment. Hence, one package affyPara, has been developed in Bioconductor repository to facilitate parallel preprocessing of microarray data on a Computer Cluster. The graphical representation of the affy and arrayQualityMetrics packages is not optimized for parallel and huges numbers of microarray data. Therefore, in this work, the algorithms boxplotPara() and MaplotPara() have been implemented in R for the affyPara package to detect bad quality samples (or outliers), when the samples to be analyzed are more than 200 arrays. They have been implemented using several functions of affyPara and snow and statistical graphical representations in R like box plot, Bagplot, MA-plot and histogram. Therefore, the definition of bad quality array for each function depends on the applied graphic. BoxplotPara() includes two different methods: Method 1 - Median and IQR Boxplots, based on the statistic generated by a boxplot and the Bagplot representation and Method 2 - Cutoff Histogram, uses a cutoff value in a histogram of difference between intensities. MAplotPara(), which has been based on MA-plot properties, includes three methods: Method 1 - Limit of Sigma, Method 2 - Oscillates Loess and Method 3 - Absolute Sum of M. Both functions classify the detected bad quality arrays in levels, give the user the possibility to manipulate some parameters and generate graphics with the results. A statistic analysis of agreement among the both functions (Cohens kappa and Fleiss kappa methods) has been applied at the final results to measure how many systematic concordance exists between both approaches.

References

M. Schmidberger, M. Morgan, D. Eddelbuettel, H. Yu, L. Tierney, U. Mansmann; State-of-the-art in Parallel Computing with R; Journal of Statistical Software (2009); submitted.

M. Schmidberger, U. Mansmann; Parallelized preprocessing algorithms for high-density oligonucleotide array data; 22th International Parallel and Distributed Processing Symposium (IPDPS 2008), Proceedings, ISBN: 978-1-4244-1693-6, 14-18 April 2008, Miami, FL, USA.

Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber; arrayQualityMetrics a bioconductor package for quality assessment of microarray data; Bioinformatics Advance Access published on February 1, 2009, Bioinformatics 25: 415-416.

R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit; Bioinformatics and Computational Biology Solutions Using R and Bioconductor; 2005, Springer.

Bayesian networks and their applications in sytems biology

Marco Grzegorczyk

Fakultät Statistik,

Technische Universität Dortmund,

Grzegorczyk@statistik.tu-dortmund.de

Bayesian networks are a flexible tool for reverse engineering gene regulatory networks and protein signalling pathways from expression data. Instead of detailed mathematical descriptions of the regulatory relationships in terms of systems of coupled differential equations, Bayesian networks are based on conditional probability distributions of a standard form what results in a scoring function ('marginal likelihood') of closed form that depends only on the network and avoids overfitting problems. We present two probability models which have been employed in the past: the multinomial distribution with the Dirichlet prior (BDe) and Gaussian distribution with the normal-Wishart prior (BGe). These models are restricted in that they either require the data to be discretised or can only learn linear relationships. We present a non-linear generalization of the BGe model based on a mixture model, using latent variables to assign measurements to different classes, and we apply the new approach to two gene expression data sets. The focus of the Arabidopsis thaliana experiment is on nine circadian genes which have been measured in two different plants under constant light condition after different dark:light entrainments. BGM infers a two stage process and a phase shift, which can be explained by the different entrainments. The focus of the mouse macrophage experiment is on the interactions between three Interferon Regulatory Factors under different cellular conditions: infection with Cytomegalovirus, treatment with Interferon Gamma, and infection after pre-treatment.

References

Grzegorczyk, M.; Husmeier, D.; Edwards, D.E.; Ghazal, P.;and Millar, A.J. (2008): Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. Bioinformatics, 24, 2071-2078.

Grzegorczyk, M.; and Husmeier, D. (2008): Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. Machine Learning, 71, 265-305.

Reconstructing Signaling Pathways from RNAi Data using Bayesian Networks and Markov Chains

Lars Kaderali, Eva Dazert, Ulf Zeuge, Michael Frese, and Ralf Bartenschlager

Viroquant Research Group Modeling, Bioquant BQ26,
University of Heidelberg,
Department of Molecular Virology,
University of Heidelberg,
lars.kaderali@bioquant.uni-heidelberg.de

The qualitative and quantitative characterization of interactions between genes in signal transduction and genetic regulatory networks is a major research goal in systems biology. RNA interference (RNAi) offers an approach to systematically screen for genes associated with a particular phenotype or cellular pathway of interest (Fire, 1998). However, while RNAi is well suited to identify genes associated with a particular phenotype, the temporal and spatial placement of these genes in the respective cellular pathways remains a challenging problem (Moffat and Sabatini, 2006). We here focus on the data-driven inference of underlying networks directly from observed phenotypes after perturbation or knockdown of individual genes.

Prior work on this problem has focused on clustering phenotypes (Sacher, 2008), on observing the nested structure of effects of different knockdowns (Markowetz, 2007), and recently in the context of drug pair treatments instead of RNAi perturbations also on detailed deterministic modeling based on differential equations (Nelander, 2008). However, while nested effect models offer a stochastic framework that may be well suited to handle noisy data, they require high-dimensional phenotype readouts, which are often not available. Furthermore, nested effect models do not take the dynamic aspects of signal transduction into account, and cannot be used for predictive simulations. Differential equation models, on the other hand, are not suitable to deal with stochastic effects, and in particular nonlinear models may be prone to overfitting due to their large number of parameters.

Due to the combinatorial explosion of possible network topologies for increasing number of pathway components, an iterative procedure that weighs alternative topologies explaining the data with their respective probabilities, and allows for the design of additional experiments to resolve the topology further, would be highly desirable.

We propose to address these issues using a dynamic Bayesian network model, using Boolean variables to represent genes, and activation probabilities for each gene described by sigmoid functions. To deal with the problem of overfitting, we use strict regularization of the model parameters using a prior distribution driving network inference to sparse networks. We simultaneously infer model topology and model parameters using a Markov chain Monte Carlo approach, by sampling from the posterior distribution over model parameters given the knockdown data. We first present an approach to compute the exact transition probabilities between different network states, taking into account the effect of single- or combinatorial knockdowns. This leads to an exact equation for the likelihood. We address the problem of incomplete observations by marginalization over unobserved nodes. We then present a likelihood approximation using a strategy based on simulation with the underlying stochastic model. By embedding this simulation approximation into the Markov chain Monte Carlo approach, we can sample from the posterior without explicit evaluation of the likelihood. Mode hopping steps integrated in the sampling algorithm furthermore allow an efficient exploration of alternative, possibly very distinct, network topologies.

We evaluate our approach on a small simulated 5-Gene network, showing that the original topology is accurately reconstructed. We provide an extensive discussion of identifiability and robustness of the inference on this simulated example. We furthermore present results on the Jak/Stat signal transduction pathway in a hepatoma cell line with a subgenomic hepatitis C virus genotype, where we carried out systematic knockdowns of 11 genes after IFN stimulation. We show that we can correctly reconstruct the core jak/stat pathway topology, and can provide confidence intervals for parameters and topologies. Different topologies consistent with the experimental data are identified, with their associated probabilities. We conclude by discussing promises this approach holds for experiment design, and discuss open questions and ongoing work.

Feature extraction and classification in mass spectrometry using sparse coding algorithms

Theodore Alexandrov

Center for Industrial Mathematics,
University of Bremen,
theodore@math.uni-bremen.de

Mass spectrometry (MS) is an important technique for chemical profiling which calculates for a sample a high dimensional histogram-like spectrum. A crucial step of MS data processing is the peak picking which selects peaks containing information about molecules with high concentrations which are of interest in an MS investigation. Another important problem of MS data processing is classification of spectra. This problem frequently arise e.g. in proteomics where the aim of research is to build and to interpret a model discriminating ill patients from healthy controls. First, we present a new procedure of the peak picking based on a sparse coding algorithm. Given a set of spectra of different classes, i.e. with different positions and heights of the peaks, this procedure can extract peaks by means of unsupervised learning. Instead of an L1-regularization penalty term used in the original sparse coding algorithm we propose using an elastic-net penalty term for better regularization, see [Alexandrov et al., 2009b]. The evaluation is done by means of simulation. We show that for a large region of parameters the proposed peak picking method based on the sparse coding features outperforms a mean spectrum-based method. Moreover, we demonstrate the procedure applying it to a real-life dataset of colorectal cancer and control spectra. Second, the sparse coding algorithm used provides a sparse representation of the set of spectra in terms of a few basis vectors. Each original spectrum is a linear combination of the basis vectors where the coefficients used are known explicitly. We show that using these coefficients as features, one can successfully classify spectra. Being applied to the real-life dataset, the proposed approach outperforms Linear Discriminant Analysis, but provides lower total recognition rates compared to an advanced classification approach, namely combination of Discrete Wavelet Transformation and Support Vector Machines with statistical feature selection [Alexandrov et al., 2009a]. However, the number of coefficients used in representation of the spectra (a spectrum of length 4731 points is represented by less than 10 coefficients) is significantly smaller than the number of wavelet coefficients exploited, even after the

statistical feature selection (200-300 depending on the statistical test). This significantly reduces classification runtime and may lead to more general classifiers. What is more, in contrast to the DWT approach, the features used can be easily interpreted. In the ideal case each basis vector represents a class-specific prototype similar to an in-class mean but produced in an unsupervised manner.

References

[Alexandrov et al., 2009a] Alexandrov, T., Decker, J., Mertens, B., Deelder, A. M., Tollenaar, R. A. E. M., Maass, P., & Thiele, H. (2009a). Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. Bioinformatics, 25(5), 643–649.

[Alexandrov et al., 2009b] Alexandrov, T., Keszoecze, O., Lorenz, D. A., Schiffer, S., & Steinhorst, K. (2009b). An active set approach to the elastic-net and its applications in mass spectrometry. In Proc. SPARS09.
available at http://hal. archives- ouvertes.fr/docs/00/36/93/97/PDF/19.pdf.

A general, prediction error based criterion for selecting model complexity for high-dimensional survival models

Christine Porzelius, Martin Schumacher and Harald Binder

Institut für Medizinische Biometrie und Informatik,
Universität Freiburg,
cp@imbi.uni-freiburg.de

Fitting predictive survival models to high-dimensional data typically requires regularized estimation techniques. An adequate criterion for selecting the amount of regularization, i.e. the model complexity, is needed, to avoid overfitting. Usually, the predictive partial log-likelihood is used, which is estimated via cross-validation. As an alternative criterion we propose a relative version of the integrated prediction error curve, which is based on the Brier score, estimated via bootstrap resampling. This criterion has the advantage that it is also applicable for models and fitting techniques where the partial log-likelihood is not available, as e.g. random forests. To investigate the performance of the two criteria, a simulation study is carried out, mimicking microarray survival data. For regularized estimation, we apply a boosting technique, in which model complexity corresponds to the number of boosting steps. Model selection by bootstrap estimates of the integrated prediction error curve is seen to perform not consistently better or worse compared to selection by cross-validation or bootstrap estimates of the partial log-likelihood. Therefore it is expected to be a reasonable alternative in cases where there is no partial log-likelihood. Similar results were found for the analysis of microarray survival data from patients with diffuse large-B-cell lymphoma and breast cancer.

Independent screening approaches for Cox models with ultrahigh dimensionality

Manuela Zucknick, Axel Benner and Thomas Hielscher

Division of Biostatistics,
German Cancer Research Centre Heidelberg,
m.zucknick@dkfz-heidelberg.de

Sparse penalised likelihood methods like lasso or SCAD can perform variable selection in very high-dimensional applications. However, the good model consistency properties, that these methods enjoy in low dimensions and/or under certain restrictions on the correlation structure, do not always translate into good model selection performance in ultrahigh-dimensional real data applications. Applying simple independence screening methods as a first step can help to reduce the dimensionality enough (i.e. below sample size) to establish good performance of penalised likelihood methods, which are applied in a second step. However, such a two-step approach requires that the initial step has the sure screening property (Fan and Lv, 2008), i.e. that all important variables survive the screening step. For linear models, several independence screening methods have been proposed recently, that have the sure screening property under certain weak assumptions (Fan and Lv, 2008; Bühlmann et al., 2009). However, we are here interested in applications of the Cox model for censored data. We therefore investigate existing screening approaches for censored data (e.g. Tibshirani 2009) with respect to the sure screening property. In addition, we adapt the sure independence screening (SIS) and iterative SIS (ISIS) methods by Fan and Lv for the Cox model, using existing SIS and ISIS modifications for likelihood-based screening approaches. Since the ISIS method is related to lasso via the LARS algorithm and its connection to componentwise L2 - boosting, lasso regression might show good screening performance in situations where ISIS works well. We will therefore include the lasso in our comparison of screening methods. We present results of simulations and an application to gene expression data, in which we investigate the impact of these initial screening methods in the context of Cox regression. We compare the screening approaches with respect to their ability to keep the important variables as well as variable selection performance and prediction accuracy of the final fitted models.

References

Bühlmann, P., Kalisch, M., and Maathuis, M.H. (2009). Variable selection in high-dimensional models: partially faithful distributions and the PC-simple algorithm. URL ftp://ftp.stat.math.ethz.ch/Research-Reports/ Other-Manuscripts/buhlmann/buhl-kal-maath.pdf

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society, Series B: Statistical Methodology, 70(5), 849–911.

Tibshirani, R. (2009). Univariate shrinkage in the cox model for high dimensional data. Depts. of Health, Research & Policy, and Statistics, Stanford Univ, Stanford CA, 94305.

Empirical Transition Matrix of Multistate Models: The etm Package

Arthur Allignol, Martin Schumacher and Jan Beyersmann

Freiburg Center for Data Analysis and Modeling,
University of Freiburg,
Institute of Medical Biometry and Medical Informatics,
University Medical Center Freiburg,
arthur.allignol@fdm.uni-freiburg.de

When dealing with data in which patients can experience more than one single event type, multistate models provide a pertinent modelling framework. Well known examples include the competing risks model in which individuals can die from one of several possible causes, and the illness-death model that allows to study the impact of an intermediate event on a terminal event. In this framework, one key quantity is the matrix of transition probabilities that can be estimated by the empirical transition matrix, also referred to as the Aalen- Johansen estimator. In this talk, we present the R-package etm that computes and displays the matrix of transition probabilities. The etm package also features a Greenwood-type estimator of the covariance matrix, which has recently been found to be the preferable estimator in the competing risks situation. The use of the package is illustrated through a prominent example in bone marrow transplant for patients suffering from leukaemia.

References

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). Statistical Models Based on Counting Processes. Springer-Verlag, New York.

Klein, J., Szydlo, R., Craddock, C. and Goldman J. (2000). Estimation of current Leukaemia-Free Survival Following Donor Lymphocyte Infusion Therapy for Patients with Leukaemia who Relapse after Allografting: Application of a Multistate Model. Statistics in Medicine, 19, 3005–3016.

Putter, H., Fiocco, M. and Geskus, R. B. (2007). Tutorial in Biostatistics: Competing Risks and Multi-State Models. Statistics in Medicine, 26, 2389–2430.

Parametric Simultaneous Inference Under Test

Esther Herberich and Torsten Hothorn

Institut für Statistik,
LMU München,
Esther.Herberich@stat.uni-muenchen.de

Multiple testing problems occur in many areas of application. Hothorn, Bretz and Westfall (2008) introduced a framework for simultaneous inference in general parametric models, which allows for an arbitrary number of null hypotheses to be tested simultaneously with an overall type I error rate below the nominal level alpha. Each null hypothesis is specified by a linear combination of model parameters. The test procedure is based on the asymptotic or exact distribution of the linear functions set up in the hypotheses; a reference distribution which is obtained under little restrictive conditions. As normality and homoscedasticity are not assumed, the framework allows for simultaneous inference in various parametric models such as linear regression and ANOVA models, generalized linear models, Cox proportional hazard models, linear mixed effects models, and robust linear models. In ANOVA models, multiple comparisons can be considered not only of contrasts of means, but of arbitrary contrasts specified by a linear function of the model parameters. In a simulation study the size and power properties of this test procedure were investigated in various parametric models (Herberich, 2009). Furthermore, the performance of a robust variant of simultaneous inference using sandwich estimators was investigated in ANOVA models with heterogeneous variances and compared to the properties of other post-hoc tests which assume homoscedasticity.

References

Herberich, E. (2009). Niveau und Güte simultaner parametrischer Inferenzverfahren. Diploma Thesis, Ludwig-Maximilians-Universität München, Institut für Statistik.

Hothorn, T., Bretz, F. and Westfall, P. (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal, 50(3):346–363.

On the influence of non-perfect randomness on probabilistic algorithms

Markus Maucher, Uwe Schöning and Hans A. Kestler

AG Bioinformatics and Systems Biology,
Institute of Neural Information Processing,
Institute of Theoretical Computer Science,
University of Ulm,
markus.maucher@uni-ulm.de

Randomization is a valuable tool in computer science. The randomized QuickSort algorithm, for example, uses randomness to avoid worst case inputs and achieves a running time of order n*log(n) even for a worst case input, while its deterministic version has a running time of order n^2 in the worst case. Optimization problems represent another area where randomness is widely used. Many of these problems, for example the Traveling Salesman Problem, can be solved or approximated with the help of probabilistic search heuristics (see [1]) like Simulated Annealing or genetic algorithms (for descriptions, see [2] and [3]). Run-time analysis as well as error analysis of probabilistic algorithms is usually based on the assumption that the random numbers that the algorithms use are independent and uniformly distributed. Since computers are deterministic devices, that assumption does usually not hold - in order to involve real randomness, external data has to be collected. Such data is only gained relatively slowly, so it is used sparsely - usually, it is used as a seed for a pseudo random number generator that constructs longer sequences from few random numbers. The behavior of the randomized QuickSort algorithm with a linear congruential generator has been investigated in [4]. There, it was shown that the use of a linear congruential generator can lead to a worst case running time of order n^2, while the choice of a different generator can lead to a running time of order n*log(n). We will present theoretic bounds for the worst case running time of the randomized version of QuickSort and show how decreasing the quality of the source of randomness can gradually increase its worst case running time from n*log(n) to n^2. We will then compare Simulated Annealing and a genetic algorithm with respect to their behavior when bad random numbers are used. Although these two probabilistic local search heuristics are both based on random local perturbations, they show different susceptibility to the use of non-perfect random numbers. We will present experimental

results observed when solving the Traveling Salesman Problem with these heuristics, using flawed random numbers, i.e random numbers that are biased or have a short period length.

References

[1] Gerhard Reinelt. The Traveling Salesman. Springer Berlin Heidelberg, 1994.

[2] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. Science, 4598:671–680, 1983.

[3] J. H. Holland, Adaptation in Natural and Articial Systems, MIT Press 1975

[4] H.J. Karloff and P. Raghavan. Randomized algorithms and pseudorandom numbers. Journal of the ACM (JACM), 40(3):454–476, 1993.

Liste der bisher erschienenen Ulmer Informatik-Berichte

Einige davon sind per FTP von `ftp.informatik.uni-ulm.de` erhältlich
Die mit * markierten Berichte sind vergriffen

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

95-12 *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^{cc} 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 *HA Kestler, B Lausen, H Binder, H-P Klenk, F Leisch, M Schmid (eds)*
Statistical Computing 2009, Abstracts der 41. Arbeitstagung

Ulmer Informatik-Berichte
ISSN 0939-5091

Herausgeber:
Universität Ulm
Fakultät für Ingenieurwissenschaften und Informatik
89069 Ulm