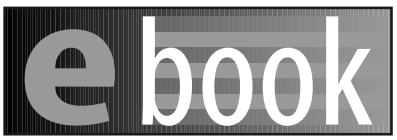
COMPSTAT 2002

Proceedings in

Computational Statistics

Wolfgang Härdle
Bernd Rönz

Humboldt-Universität zu Berlin
School of Business and Economics
Institute of Statistics and Econometrics
Spandauer Strasse 1
D-10178 Berlin
Germany



Preface

This COMPSTAT 2002 book contains the Keynote, Invited and Full Contributed papers presented in Berlin. A companion volume including Short Communications and Posters is published on CD.

The COMPSTAT 2002 is the 15th conference in a series of biannual conferences with the objective to present the latest developments in Computational Statistics and is taking place from August 24th to August 28th 2002.

Previous COMPSTATs were in Vienna (1974), Berlin (1976), Leiden (1978), Edinburgh (1980), Toulouse (1982), Prague (1984), Rome (1986), Copenhagen (1988), Dubrovnik (1990), Neuchâtel (1992), Vienna (1994), Barcelona (1996), Bristol (1998) and Utrecht (2000).

COMPSTAT 2002 is organised by CASE, Center of Applied Statistics and Econometrics at Humboldt-Universität zu Berlin in cooperation with Freie Universit¨t Berlin and University of Potsdam.

The topics of COMPSTAT include methodological applications, innovative software and mathematical developments, especially in the following fields: statistical risk management, multivariate and robust analysis, Markov Chain Monte Carlo Methods, statistics of E-commerce, new strategies in teaching (Multimedia, Internet), computerbased sampling/questionaires, analysis of large databases (with emphasis on computing in memory), graphical tools for data analysis, classification and clustering new statistical software and historical development of software.

A conference like COMPSTAT is not possible without an active scientific input from many high class researchers. The contributions were classified by the Scientific Programme Committee (SPC): Prof. Wolfgang H¨rdle (Berlin), Prof. Antony Unwin (Augsburg), Prof. Jaromir Antoch (Prague), Prof. Adrian Bowman (Glasgow), Prof. Michel Delecroix (Paris), Prof. Wenceslao Gonzalez Manteiga (Santiago de Compostela), Prof. Junji Nakano (Tokyo), Prof. Michael Schimek (Graz) and Prof. Peter van der Heijden (Utrecht).

The papers were refereed by the SPC and we would like to thank our colleagues for all their help. The resulting proceedings provide a broad overview of the current research areas in statistical computing.

We are also very grateful to the members of the Local Organizing Committee: Dr. Yasemin Boztug, Prof. Dr. Herbert Büning, Prof. Dr. Wolfgang Härdle, Prof. Dr. Lutz Hildebrandt, Dr. Sigbert Klinke, Prof. Dr. Uwe Küchler, Prof. Dr. Bernd Rönz, Dr. Peter Schirmbacher, Dipl.Ing. Benjamin Schüler, Prof. Dr. Vladimir Spokoiny, Prof. Dr. Hans Gerhard Strohe, Prof. Dr. Jürgen Wolters and Uwe Ziegenhagen.

We would also like to thank Patricia Ahrend and Luise Richter who together with Dr. Klinke did a great job in managing the papers.

Berlin, August 2002:

W. Härdle B. Rönz

Contents

1 Invited papers

An Implementation for Regression Quantile Estimation

T.W. Yee

Keywords: Age-reference Centile Analysis, LMS Method, Penalized Likelihood, Quantile Regression, R, S-Plus, Vector Generalized Additive Models

Of the many methods that have been developed for quantile regression (Wright and Royston, 1997) the LMS method (Cole and Green, 1992) can be easily understood, is flexible and based on splines. The basic idea is that, for a fixed value of the covariate, a Box-Cox transformation of the response is applied to obtain standard normality. The three parameters are chosen to maximize a penalized log-likelihood. One unpublished extension by Lopatatzidis and Green is to transform to a gamma distribution, which helps overcome a range-restriction problem in the original formulation. This paper proposes a new method based on the Yeo and Johnson (2000) transformation. It has the advantage that it allows for both positive and negative values in the response. R/S-PLUS software written by the author implementing the LMS method and variants are described and illustrated with data from a New Zealand workforce study.

Computational Methods for Time Series Analysis

G. Kitagawa, T. Higuchi, S. Sato

Keywords: General State Space Model, Nonlinear Filtering, Sequential Monte Carlo Method, Parallel Computation, Self-organizing State Space Model

By the progress of fast computing facilities and various computing technologies, it becomes realistic to apply computer intensive methods to statistical analysis. In time series analysis, sequential Monte Carlo methods was developed for general state space models which enables to consider very complex nonlinear non-Gaussian models. In this paper, we show algorithms, implementations and parameter estimation for Monte Carlo Filter

and smoother. Various ways of the use of parallel computer are also discussed. The usefulness of the general state space modeling is illustrated with several examples.

Forecasting PC-ARIMA Models for Functional Data

M.J. Valderrama, F.A. Ocaña, A.M. Aguilera

Keywords: ARIMA, *El Niño*, Functional Data, Principal Component

This paper introduces an improvement on the forecasting models previously developed by the authors for continuous time series based on the PCA of the stochastic process by cutting series in seasonal periods. The new approach consists of modelling principal components as ARIMA processes and then to formulate a mixed PC-ARIMA model for the time series. This methodology is then applied to the climatic phenomenon known as *El Niño*.

KyPlot as a Tool for Graphical Data Analysis

K. Yoshioka

Keywords: Data Visualization, Density Estimation, Graph Fitting, Nonparametric Regression, Smoothing Methods

KyPlot is a software package intended to provide an integrated environment for data analysis and visualization. It offers a broad range of statistical procedures on a spreadsheet interface and versatile graphing tools. KyPlot has the functionality of running a screen show to present the created graphs and schemes. It also supports an interactive system of graph fitting, including nonlinear regression, interpolation and a variety of smoothing methods for nonparametric regression and density estimation.

Mice and Elephants Visualization of Internet Traffic

J.S. Marron, F. Hernandez-Campos, F.D. Smith

Keywords: Heavy Tail Distribution, HTTP Flows, Visual Representation, Zooming Graphics

Internet traffic is composed of flows, sets of packets being transferred from one computer to another. Some visualizations for understanding the set of flows at a busy internet link are developed. These show graphically that the set of flows is dominated by a relatively few "elephants", and a very large number of "mice". It also becomes clear that "representative sampling" from heavy tail distributions is a challenging task.

Relativity and Resolution for High Dimensional Information Visualization with Generalized Association Plots (GAP)

S.-C. Chang, C.-H. Chen, Y.-Y. Chi, C.-W. Ouyoung

Keywords: Categorical Data, Color-coding, Data Mining, Matrix Gap, Multiple Correspondence Analysis, Seriation

Generalized association plots (GAP) (Chen, 1996; 1999; 2002) is an information visualization environment for high dimensional data structure without dimension reduction. There is no limit for sample size and variable number. Three matrix maps for raw data matrix, object proximity matrix, and variable proximity matrix are created for visually extracting grouping structures for objects and variables and the interaction information between object-clusters and variablegroups. Seriation algorithms are developed to permute objects and variables such that rows and columns with similar profiles are arranged at closer positions. Categorical generalized association plots (cGAP) (Chen, 1999; Chen et al., 2002) is an extension of GAP adapted for visualizing high dimensional categorical data structure. Optimal scaling (multiple correspondence analysis) is applied to compute the proximity matrices for objects as well as for variables and to obtain colors for coding all categories in the raw data matrix. Relativity and resolution are two related critical issues in conducting efficient GAP and cGAP analysis. This article discusses possible solutions when standard procedures fail in generating satisfactory relativity and resolution for GAP and cGAP. Color version of figures presented in this article together with software packages for GAP and cGAP can be obtained from our website at http://gap.stat.sinica.edu.tw.

Supervised Learning from Microarray Data

T. Hastie, R. Tibshirani, B. Narasimhan, G. Chu

Keywords: Classification, Discriminant Analysis, Expression, Microarrays

Gene expression arrays pose challenging problems for most traditional supervised learning techniques. We present a discussion of some of the issues involved. We then propose a simple approach to class prediction for DNA microarrays, based on a enhancement of the nearest centroid classi.er. Our technique uses soft-thresholded class centroids as prototypes for each class. The shrinkage improves signi.cantly prediction performance, and identifies a subset of the genes most responsible for class separation. The method performs as well or better than competitors from the literature, and is easy to understand and interpret. We illustrate the technique on data from three studies: small round blue cell tumors, leukemia and breast cancer.

Teaching Statistics with Electronic Textbooks

J. Symanzik, N. Vukasinovic

Keywords: ActivStats, CyberStats, Introductory Statistics, MM*Stat

In the first part of this paper, we will provide a general overview on electronic Introductory Statistics textbooks and summarize their individual strengths and weaknesses. Examples of such electronic textbooks are

- the Web-based CyberStats (http://www.cyberk.com),
- the Web-based MM*Stat (http://www.mm-stat.com/ and http://ise.wiwi.hu-berlin.de/ mdstat/scripts/mmeng/start.html),
- and the CD-ROM-Based ActivStats (http://www.datadesk.com/ActivStats/).

In the second part of this paper, we will present our experiences from teaching an undergraduate Introductory Statistics course at Utah State University, using CyberStats. We compare this Web-based course with similar textbook-based courses and report student viewpoints as well as instructor viewpoints.

Trans-Dimensional Markov Chains and their Applications in Statistics

S.P. Brooks

Keywords: Convergence Assessement, MCMC, Model Discrimination, Recovery-Recapture, Reversible Jump MCMC, Variable Selection

We discuss the application of trans-dimensional Markov chains to both Bayesian and Classical model discrimination.We discuss how to measure and improve the efficiency of the chains and illustrate the ideas via a classical variable selection problem and a Bayesian recovery-recapture analysis.

2 Full papers

A Bayesian Model for Compositional Data Analysis

M.J. Brewer, C. Soulsby, S.M. Dunn

Keywords: Bayesian Model, Compositional Data, MCMC

We introduce a Bayesian hierarchical model for the analysis of compositional data, where the source data are observed as indicator measurements rather than direct proportions.We focus on an application of the model to water quality data, where we apportion stream water to three sources.

A Comparison of Marginal Likelihood Computation Methods

C.S. Bos

Keywords: Bayesian Analysis, Marginal Likelihood

In a Bayesian analysis, different models can be compared on the basis of the expected or marginal likelihood they attain. Many methods have been devised to compute the marginal likelihood, but simplicity is not the strongest point of most methods. At the same time, the precision of methods is often questionable. In this paper several methods are presented in a common framework. The explanation of the differences is followed by an application, in which the precision of the methods is tested on a simple regression model where a comparison with analytical results is possible.

A Hotelling Test Based on MCD

G. Willems, G. Pison, P.J. Rousseeuw, S. van Aelst

Keywords: Hotelling, Minimum Covariance Determinant, One-sample Hypothesis Test, Robustness

Hypothesis tests and confidence intervals based on the classical Hotelling T^2 statistic can be adversely affected by outliers. Therefore, we construct an alternative inference technique based on a statistic which uses the highly robust MCD (Minimum Covariance Determinant) estimator of Rousseeuw (1984) instead of the classical mean and covariance matrix. Recently, a fast algorithm was constructed to compute the MCD (Rousseeuw and Van Driessen 1999). In our test statistic we use the reweighted MCD, which has a higher efficiency. The distribution of this new statistic differs from the classical one. Therefore, the key problem is to find a good approximation for this distribution. Similarly to the classical T^2 distribution, we obtain a multiple of a certain F-distribution. It is shown through a Monte Carlo study that this approximation is very accurate, both at the normal model and at contamination models.

A Resampling Approach to Cluster Validation

V. Roth, T. Lange, M. Braun, J. Buhmann

Keywords: Cluster Validation, Model Selection, Unsupervised Learning

The concept of *cluster stability* is introduced as a means for assessing the validity of data partitionings found by clustering algorithms. It allows us to explicitly quantify the quality of a clustering solution, without being dependent on external information. The principle of maximizing the cluster stability can be interpreted as choosing the most *self-consistent* data partitioning. We present an empirical estimator for the theoretically derived stability index, based on imitating independent sample-sets by way of resampling. Experiments on both toy-examples and real-world problems effectively demonstrate that the proposed validation principle is highly suited for model selection.

A Self Documenting Programming Environment for Weighting

W. Grossmann, P. Ofner

Keywords: Metadata, Programming Environments, Weighting

The paper presents a proposal for a statistical computing environment for weighting, which processes statistical data and description of data simultaneously. The developed model is of interest also for other types of computing in official statistics.

A State Space Model for Non-Stationary Functional Data

M. Ortega-Moreno, M.J. Valderrama, J.C. Ruiz-Molina

Keywords: B-splines, Brownian Motion, Kalman-Bucy Filter, Karhunen-Lòeve Expansion

A time dependent state space model with minimal dimension is introduced in this paper by approximating the stochastic process of continuous time nature by means of spline interpolation of its sample paths and then by differentiating its Karhunen-Lòeve expansion. A comparative study of forecasting, using the Kalman-Bucy Filter, with simulated data is presented from a well known non-stationary process, the Brownian motion, discussing its advantages.

A Wildlife Simulation Package (WiSP)

W. Zucchini, M. Erdelmeier, D. Borchers

Keywords: Abundance Estimation, Statistical Software, Wildlife, *WiSP*

WiSP is an ***R*** library of functions designed as a teaching tool to illustrate methods used to estimate the abundance of closed wildlife populations. It enables users to generate animal populations having realistically complex spatial and individual characteristics, to generate survey designs for a variety of survey techniques, to survey the populations and to estimate the abundance. It can be used to assess properties of estimators when the model assumptions are violated.

Algorithmical and Computational Procedures for a Markov Model in Survival Analysis

J.E. Ruiz-Castro, R. Pérez-Ocón, D. Montoro-Cazorla

Keywords: Markov Model, Matlab, Survival Data, Type-phase Distribution

A methodology for studying the homogeneous time-continuous Markov processes that allow the usual quantities of interest in survival studies to be expressed in a well-structured form is considered. This is performed by introducing the phase-type distributions, which enable algorithmic expressions for the performance measures used in survival analysis. A Markov model with several absorbent states is studied, incorporating the phase-type distributions. Following this procedure, we show that Markov models can be applied in a direct and tractable way. In this general model covariates are introduced and applied to analysing the behaviour of breast cancer. The computational implementation can be developed from the formulae proposed in a more convenient way than previous ones and this has been performed using the *Matlab* program.

An Algorithm for the Construction of Experimental Designs with Fixed and Random Blocks

P. Goos, A. N. Donev, M. Vandebroek

Keywords: BLKL Exchange Algorithm, Correlation Observations, D-optimality

This paper is concerned with the design of experiments where the relationship between a response variable of interest and a number of quantitative variables is studied. The observations have to be divided into blocks. Some of the block effects are regarded as random while others as fixed. An algorithmic approach to the construction of such designs is shown to be very successful.

An Algorithm to Estimate Time Varying Parameter SURE Models under Different Type of Restrictions

S. Orbe, E. Ferreira, J.M. Rodríguez-Póo

Keywords: Constrained Estimators, Cross Restrictions, Nonparametric Methods, Seasonal Restrictions

In this paper we propose an algorithm to estimate nonparametrically time varying coefficients in seemingly unrelated regression models. The estimation algorithm allows for restrictions both along time and across the different equations of the model. This procedure presents at least two advantages with respect to other estimation procedures in the literature. First, a closed form for the estimator is obtained and therefore there is no need to use iterative methods. Second, since the direct computation of the estimator requires to solve a linear system where the number of equations and coefficients increases with the sample size, an algorithm is derived to reduce substantially the computational cost. The asymptotic properties of the estimator are derived and a test procedure for the validity of the time varying constraints is implemented. Finally the proposed methodology is applied to estimate and test a log linear demand system.

Analyzing Data with Robust Multivariate Methods and Diagnostic Plots

G. Pison, S. van Aelst

Keywords: Cutoff Values, Diagnostic Plot, Influence Functions, Outliers, Robust Distances, Robustness

Principal Component Analysis, Canonical Correlation Analysis and Factor Analysis (Johnson und Wichern 1998) are three different methods for analyzing multivariate data. Recently robust versions of these methods have been proposed by Croux and Haesbroeck (2000), Croux and Dehon (2001) and Pison et al. (2002) which are able to resist the effect of outliers. However, there does not yet exist a graphical tool to display the results of the robust data analysis in a fast way. Therefore we now construct such a diagnostic tool based on empirical influence functions. These graphics will not only allow us to detect the influential points for the multivariate statistical method but also classify the observations according to their robust distances. In this way we can identify regular points, good (non-outlying) influential points, influential outliers, and non-influential outliers. We can downweigh the influential outliers in the classical estimation method

to obtain reliable and efficient estimates of the model parameters. Some generated data examples will be given to show how these plots can be used in practice.

Application of "Aggregated Classifiers" in Survival Time Studies

A. Benner

Keywords: Bagging, Boosting, Loss Function, Survival Trees

A gradient-descent boosting algorithm is presented for survival time data, where the individual additive components are regression trees.

Application of Hopfield-like Neural Networks to Nonlinear Factorization

D. Husek, A.A. Frolov, H. Rezankova, V. Snasel

Keywords: Binary Factorization, Hopfield Network, Sparse Encoding

The problem of binary factorization of complex patterns in recurrent Hopfield-like neural network was studied by means of computer simulation. The network ability to perform a factorization was analyzed depending on the number and sparseness of factors mixed in presented patterns. Binary factorization in sparsely encoded Hopfield-like neural network is treated as efficient statistical method and as a functional model of hippocampal CA3 field.

Bagging Tree Classifiers for Glaucoma Diagnosis

T. Hothorn, B. Lausen

Keywords: Discriminant Analysis, Error Rate Estimation, Method Selection Bias

The aggregation of multiple unstable classifiers leads to substantial reduction of misclassification error in many applications and bench mark problems. We focus on the

problem of classifying eyes as normal or glaucomatous based on measurements derived from laser scanning images of the optic nerve head. The performance of various aggregated classifiers is investigated for a clinical training sample and for a simulation model of eye morphologies.

Bayesian Automatic Parameter Estimation of Threshold Autoregressive (TAR) Models using Markov Chain MonteCarlo (MCMC)

E. Amiri

Keywords: Gibbs Sampler, MCMC, Non Linear Time Series, Threshold Autoregressive Models

In nonlinear time series analysis a Threshold Autoregressive model (TAR) is considered as an approximation to a Nonlinear autoregressive process (NLAR). A TAR model is a piece-wise linear model over the state space. It is linear in the space of the thresholds. In this article we confine our studies to a class of TAR models which is called Self-Exciting Threshold Autoregressive (SETAR). The first step in fitting a SETAR model to a data-set is to decide the number of regimes. The second step is to identify the delay, the threshold(thresholds) and the order of each regime. Our approach is to estimate automatically, the number of regimes, the delay, the threshold(thresholds), the order and parameters of each regime for this class of threshold models via a Bayesian approach using MCMC methods. Model selection is done through calculating mean of the root mean square error (MRMSE) of forecasts. Keywords. Non linear Time Series, Threshold Autoregressive models, Markov Chain Monte Carlo (MCMC), Gibbs Sampler.

Bayesian Semiparametric Seemingly Unrelated Regression

S. Lang, S.B. Adebayo, L. Fahrmeir

Keywords: Bayesian Semiparametric Models, Correlated Responses, MCMC, P-Splines

Parametric seemingly unrelated regression (SUR) models are a common tool for multivariate regression analysis when error variables are reasonably correlated. A weakness of parametric models is that they require strong assumptions on the functional form of possibly nonlinear e.ects of metrical covariates. In this paper, we develop a semipara-

metric SUR model based on Bayesian P-splines. Inference is fully Bayesian and uses recent Markov chain Monte Carlo techniques.

Blockmodeling Techniques for Web Mining

G. Schoier

Keywords: Blockmodeling, Cluster Analysis, log Files, Web Mining, Web Usage Mining

The aim of this paper is to introduce the concept of blockmodeling to Web data (log files), in order to obtain clusters of units (visited pages) which have similar patterns of relationships and to interpret the pattern of relationships among clusters. Doing this we can study the behaviour of the users (sessions) according to the relations among the visited pages. To explain these concepts an application to a Web site is performed.

Bootstrapping Threshold Autoregressive Models

J. Öhrvik, G. Schoier

Keywords: AIC, AIC_c, AIC_u, AR-sieve Bootstrap, Bootstrap Model Selection Criteria, Moving Block Bootstrap

Threshold autoregressive (TAR) models have been widely used for periodic time series, as they are nonlinear models relatively simple to handle being linear in different regions of the state space, see e.g. Tong (1990). One of the main problems regards the decision of the selection of the correct order of a TAR model. This problem has been addressed by Tong (1983), Wong and Li (1998) and De Gooijer (2001). Wong and Li proposed the Akaike information criterion (AIC), the biased corrected version AIC_c, the AIC_u which is an approximately unbiased estimate of the Kullback-Leibler information and the Bayesian information criterion (BIC). De Gooijer proposed a cross validation criterion (CU_u*) corresponding to AIC_u. The main purpose of this paper is to study how bootstrap selection criteria perform. These criteria are based on a weighted mean of the apparent errors in the sample, and the average error rate obtained from bootstrap samples not containing the point being predicted. We also want to compare these new measures with the traditional ones based on AIC. Bootstrap methods assume that we resample from independent identically distributed (i.i.d.) observations. This is usually not the case in time series analysis. To overcome this problem we have used different approaches. First a parametric based on the TAR model suggested by AIC_u , then a

semiparametric - AR-sieve bootstrap motivated by Buehlmann (1997) and finally moving block bootstrap, see e.g. Kuensch (1989).

Canonical Variates for Recursive Partitioning in Data Mining

C. Cappelli, C. Conversano

Keywords: Canonical Variates, High Dimensional Data Analysis, Statistical Learning, Tree-Based Methods

This paper deals with the problem of dimension reduction in the general context of supervised statistical learning, with particular attention to data mining applications. The main goal of the proposed methodology is to improve tree based methods as prediction tool by introducing an alternative approach to data partitioning which is meant to handle large numbers of (possibly correlated) covariates. The key idea is to use suitable combinations of covariates recursively identi.ed.

CAnoVa©: a Software for Causal Modeling

O. Wüthrich-Martone, C. Nachtigall, M. Müller, R. Steyer

The new impulse given in the last decade to the theory of individual and average causal effects is mostly due to the approach developed by Steyer and others and resulted in valuable theoretical results such as, for example, the link between unconfoundedness and causal unbiasedness. This approach has also allowed for the development of different practical procedures for both testing for confounding and for testing for (average) causal effects. The scope of this contribution is to present CAnoVa, a software for causal modeling that allows a straightforward use of these two methods. CAnoVa is fully compatible with SPSS and allows to test for confounding and for causal effects even in case of designs with multiple treatment variables and multiple confounders.

Classification Based on the Support Vector Machine, Regression Depth, and Discriminant Analysis

A. Christmann, P. Fischer, T. Joachims

Keywords: Data Mining, Discriminant Analysis, Logistic Regression, Overlap, Regression Depth, Separation, Support Vector Machine

The minimum number of misclassifications achievable with affine hyperplanes on a given set of labled points is a key quantity in both statistics and computational learning theory. We compare the modern approaches the regression depth method and the support vector machine with discriminant analysis. Summarizing, the regression depth method using currently available algorithms yields often better classifications results for small to moderate data sets, say for sample sizes less than 1000 and dimension up to 10, whereas the support vector machine is often more appropriate for larger or higher dimensional data mining problems.

Clockwise Bivariate Boxplots

A. Corbellini

Keywords: Bivariate Boxplot, B-spline, Convex Hull, Robust Centroid

In this paper we suggest a simple way of constructing a robust non parametric bivariate contour based on the rotation of the univariate boxplot which does not necessarily have to use a bivariate generalization of the univariate depth measures. The suggested approach is based on the projection of bivariate data along the round angle. When the angle is a multiple of $\pi/2$ we obtain the traditional univariate boxplot referred to each variable. In all the other cases we obtain univariate boxplots which keep into account in a different way the correlation between the two original variables. We apply the suggested approach to some datasets and exploit the properties of the different choices of defining the inner region and the outer contour. The final result is a simple and easy to construct bivariate boxplot which enables to visualize the location, spread skewness and tails of the data.

Combining Graphical Models and PCA for Statistical Process Control

R. Fried, U. Gather, M. Imhoff, M. Keller, V. Lanius

Keywords: Dimension Reduction, Online Monitoring, Pattern Recognition, Time Series Analysis

Principal component analysis (PCA) is frequently used for detection of common structures in multivariate data, e.g. in statistical process control. Critical issues are the choice of the number of principal components and their interpretation. These tasks become even more difficult when dynamic PCA (Brillinger, 1981) is applied to incorporate dependencies within time series data. We use the information obtained from geographical models to improve pattern detection based on PCA.

Comparing Two Partitions: Some Proposals and Experiments

G. Saporta, G. Youness

Keywords: Jaccard Index, K-means, Latent Class, Partitions, Rand Index

We propose a methodology for finding the empirical distribution of the Rands measure of association when the two partitions only differ by chance. For that purpose we simulate data coming from a latent profile model and we partition them according to 2 groups of variables. We also study two other indices: the first is based on an adaptation of Mac Nemar's test, the second being Jaccard's index. Surprisingly, the distributions of the 3 indices are bimodal.

Comparison of Nested Simulated Annealing and Reactive Tabu Search for Efficient Experimental Designs with Correlated Data

N. Coombes, R. Payne, P. Lisboa

Keywords: A-efficiency, Correlated Data, Design Optimisation, Reactive Tabu Search

More complex analysis of correlated data has created a need for algorithms to find

efficient designs for known correlation structures. Reactive Tabu Search (RTS) is adapted to the design of experiments with correlated error. It uses a scavenging descent local search and repeat returns to designs in the search path are used to invoke diversification strategies. Nested Simulated Annealing (NSA) is used in block design and adapted here for correlated data. NSA uses supercooling and reannealing to speed and continue the search. NSA is shown to be a less efficient algorithm than RTS as measured by interchanges tested and computing time.

Computational Connections between Robust Multivariate Analysis and Clustering

D.M. Rocke, D.L. Woodruff

We examine relationships between the problem of robust estimation of multivariate location and shape and the problem of maximum likelihood assignment of multivariate data to clusters and we offer a synthesis and generalization of computational methods reported in the literature. These connections are important because they can be exploited to support effective robust analysis of large data sets. Recognition of the connections between estimators for clusters and outliers immediately yields one important result that we demonstrate in this paper; namely, the ability to detect outliers can be improved a great deal using a combined perspective from outlier detection and cluster identification. One can achieve practical breakdown values that approach the theoretical limits by using algorithms for both problems. Computational results are reported that demonstrate the effectiveness of this approach.

Computer Intensive Methods for Mixed-effects Models

J.A. Sanchez, J. Ocaña

Keywords: Bootstrap, Mixed-effects Models, Simulation, Statistical Software

This work focuses on the development of a framework that implements computer intensive methods for mixed-effects models in an easy and efficient way, and special attention is given to flexible bootstrap techniques. Here we developed an S-Plus library, *nlme.bootstrap*, as an extension of the standard *nlme* library in S. This new library is designed for Monte Carlo and bootstrap methods for mixed-effects models. Some design and implementation aspects are described. To illustrate the use of the software, a simulation study on the performance of parametric bootstrap in the presence of non-normality is discussed.

Construction of T-Optimum Designs for Multiresponse Dynamic Models

D. Uciński, B. Bogacka

Keywords: Dynamic Systems, Experimental Design, Nonlinear Regression, Process Engeneering

The paper aims at developing the underlying theory and constructing an efficient procedure for determining optimal experimental conditions for discriminating between several rival multivariate statistical models where the expected response is given by ordinary differential equations. The method elaborated is validated on a simulation example.

Data Compression and Selection of Variables, with Respect to Exact Inference

J. Läuter, S. Kropf

Keywords: Data Compression, Multivariate Test, Selection of Variables

In many applications of statistics, we are confronted with a large number p of different variables whereas the number n of independent individuals remains limited. The classical multivariate tests like Wilks' Λ or Hotelling's T^2 test do not attain sufficient power under these circumstances because they cannot take into account special parameter structures. Often overfitted parameter estimations and unstable behaviour arise, and confirmatory data analysis becomes difficult. We will nevertheless intend to investigate the multivariate data by exact statistical tests. Methods of dimension reduction and data compression are preferentially used.

We start from a sample of n independent p-dimensional observation vectors

$$X = \begin{pmatrix} x'_{(1)} \\ \vdots \\ x'_{(n)} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ & \vdots & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}. \tag{2.1}$$

We will compress the data into so-called scores: $Z = XD$. For this, a $p \times q$ coefficient matrix D is used ($p \leq q$). The original p-dimensional vectors $x'_{(j)}$ are transformed into q-dimensional vectors $z'_{(j)} (j = 1, \ldots, n)$. The compression is based on principal components, that is, the matrix D is obtained as the solution of the eigenvalue problem

$$GD = D\Lambda \quad \text{with} \quad G = \sum_{j=1}^{n}(x_{(j)} - \bar{x})(x_{(j)} - \bar{x})' = (X - \bar{X})'(X - \bar{X}), \tag{2.2}$$

where $\bar{X} = 1_n\bar{x}'$ is the $n \times p$ mean matrix, $D = (d_1 \ldots d_q)$ is the matrix of the eigenvectors d_h pertaining to the q largest eigenvalues $\lambda_h \quad (\lambda_1 \geq \lambda_2 \geq \ldots, \quad D'D = I_q)$, and $\Lambda = D'GD = (Z - \bar{Z})'(Z - \bar{Z})$ is the diagonal matrix of these eigenvalues (1_n represents the vector consisting of n ones, I_q is the $q \times q$ identity matrix). In the following, we will treat different testing problems. We can show that it is sufficient to work exclusively with the compressed data Z of the diminished dimension q.

Data Extraction from Dense 3-D Surface Models

M. Bock, A. Bowman, J. Bowman, P. Siebert

Keywords: Curves, Facial Modelling, Mesh Conformation, Principal Components, Procrustes, Shape, Surfaces

Stereo-photogrammetry involves the use of pairs of cameras to extend standard photographic methods by recovering depth information through triangulation. The resulting data consist of a point-cloud of three-dimensional observations, providing a digital representation of the surface of the object of interest. The number of observations can typically be many thousands, providing a rich source of information on object shape. This can be represented efficiently as an irregular triangular mesh, in Virtual Reality Modelling Language form.

This technology is currently being used in a study of the growth of childrenś faces. There have been very few quantitative studies of growth, especially in three dimensions. This study involves an additional longitudinal component in that repeated measurements are taken over time, from 3 months to 2 years for one cohort of children and from 3 years to 5 years for another. Stereo-photogrammetry provides a practical means of gathering these data. In particular, the very fast capture time involved is important in minimizing the effects of movement in the subject. The clinical aims of the study are to contrast the facial shape and growth of healthy children with those from children who have been born with a cleft lip and/or palate and who have subsequently undergone surgical repair.

The project has developed a computing tool which is able to interrogate the three-dimensional surface information to identify anatomical landmarks. The ability to superimpose a photographic image on the three-dimensional surface information is an important feature of accurate landmark identification. Methods of analyzing landmark shape data are well developed, as described by Dryden & Mardia (1998). Procrustes methods of shape matching provide standard tools for defining mean shapes. Standard

methods of multivariate analysis can then be brought to bear in investigating the residual shape variation. In particular, principal components provide a suitable descriptive tool for displaying the key features of high-dimensional data of this type.

Initial investigations suggested that around 30 to 40 anatomical landmarks could be realiably extracted from each facial surface. This is a useful and meaningful summary but it clearly does not adequately represent the very much richer information present in each digitized face. As a first step in exploiting more of the information available, in a form suitable for analysis, facial curves with clear anatomical meaning have also been extracted. These can be represented as 'pseudo-landmarks' lying along the facial surface between two anatomical landmarks. In order to analyze groups of subjects it is necessary that there is a correspondence in these curve representations across individuals. Functional data analysis, as described by Ramsay & Silverman (1997), provides an appropriate suite of techniques to summarize and analyze differences in data of this type. These methods have been adapted to the special features of this study, namely the three-dimensional nature of the curves and the need to respect the anatomical anchors. An essential feature is the use of smoothing techniques to recover a smooth curve from the underlying irregular mesh.

In order to exploit the full extent of the information present in the irregular mesh, and to apply methods for analyzing statistical variation, it is necessary to create a mesh whose nodes correspond across individuals. Extensive use of this approach is made in computer animation and some standard computing tools are available. The relative merits of applying this approach in the context of the present study, compared to the use of curve information, will be discussed.

Although these methods have been discussed in the context of a specific application to the growth of children's faces, the generic nature of the tools makes them applicable to a wide range of other applications where the surface shape of objects is of prime interest.

Detection of Locally Stationary Segments in Time Series

U. Ligges, C. Weihs, P. Hasse-Becker

Keywords: Chattering Prediction, Note Classification, Segmentation, Time Series

In many applications it is required to segment a time series into its locally stationary parts. Two applications are presented: As a first example consider online monitoring of a BTA Deep-Hole-Drilling process. Here chattering and spiralling of the drilling tool should be avoided by process control. A second example is the analysis of vocal sound signals. It may be required to analyze only specific tones instead of a whole song. Three new algorithms are introduced in this paper, all based on the theory of Dahlhaus (1997)

and the analysis of the spectrum of time series, but with different methods to distinguish locally stationary parts of the signals.

Detection of Outliers in Multivariate Data: A Method Based on Clustering and Robust Estimators

C.M. Santos-Pereira, A.M. Pires

Keywords: Clustering, Multivariate Analysis, Outlier Detection, Robust Estimation, Supervised Classification

Outlier identification is important in many applications of multivariate analysis. Either because there is some specific interest in finding anomalous observations or as a preprocessing task before the application of some multivariate method, in order to preserve the results from possible harmful effects of those observations. It is also of great interest in supervised classification (or discriminant analysis) if, when predicting group membership, one wants to have the possibility of labelling an observation as "does not belong to any of the available groups". The identification of outliers in multivariate data is usually based on Mahalanobis distance. The use of robust estimates of the mean and the covariance matrix is advised in order to avoid the masking effect (Rousseeuw and Leroy, 1985; Rousseeuw and von Zomeren, 1990; Rocke and Woodruff, 1996; Becker and Gather, 1999). However, the performance of these rules is still highly dependent of multivariate normality of the bulk of the data. The aim of the method here described is to remove this dependence.

Development of a Framework for Analyzing Process Monitoring Data with Applications to Semiconductor Manufacturing Process

Y.-H. Yoon, Y.-S. Kim, S.-J. Kim, B.-J. Yum

Keywords: Artificial Neural Networks, Data Mining, Decision Tree, Semiconductor Manufacturing, Taguchi Parameter Design

A semiconductor manufacturing process consists of hundreds of steps and produces a large amount of data. These process monitoring data contain useful information on the behavior of a process or a product. After semiconductor fabrication is completed, dies on a wafer are classified into bins in the EDS (Electrical Die Sorting) process.

Quality engineers in semiconductor industry are interested in relating these bin data to the historical monitoring data to identify those process variables that are critical to the quality of the final product. Data mining techniques can be effectively used for this purpose. In this article, a framework for analyzing semiconductor process monitoring and bin data is developed using the data mining and other statistical techniques.

Different Ways to See a Tree – KLIMT

S. Urbanek

Keywords: Classification and Regression Trees, Exploratory Data Analysis, Interactive Software, Visualization

Recursive partitioning trees offer a valuable tool to analyze structure in datasets. There are many ways to display various structures contained in a tree. This paper describes different means of visualization of a tree with our prototype software, KLIMT (Klassification – Interactive Methods for Trees), for interactive graphical analysis of trees.

e-stat: A Web-based Learning Environment in Applied Statistics

E. Cramer, K. Cramer, U. Kamps

Keywords: Applied Statistics, Interactive, Learning Environment, Teaching Environment, Web-based

Within the "New Media in Education Funding Programme" the German Federal Ministry of Education and Research (bmb+f) supports the project "e-stat" to develop and to provide a multimedia, web-based, and interactive learning and teaching environment in applied statistics called EMILeA-stat. The structure of EMILeA-stat, its scope and objectives is sketched, and some screen-shots of interactive applets are shown. In the Compstat 2002 proceedings several other contributions on EMILeA-stat can be found.

e-stat: Automatic Evaluation of Online Exercises

K. Bartels

Keywords: Automatic Evaluation, Classification of Free Text, Natural Language Processing, Online Exercises

The implementation of tools for online exercises into the e-stat environment is described here. Besides the usual types of online exercises special emphasis is put on the automatic evaluation of free text answers. This can be formulated as a problem of classification which is based on an analysis of the free text with statistical methods of computational linguistics.

e-stat: Basic Stochastic Finance at School Level

C. Mohn, D. Pfeifer

Keywords: Arbitrage, Binomial-Tree-Model, Combination of Options, Discrete Mathematics of Finance, e-learning, e-stat, Financial Markets, Java-Applets, Option Pricing Theory, Options, Stochastic Education at School

This text shows possibilities how basic facts of mathematics of finance can be introduced at school level by using e-stat. The basic facts ground on the discrete option pricing theory in the one-period binomial-tree-model. A simple numerical example is used to explain the term of "arbitrage" and its importance to financial mathematics. The example leads into two interactive Java-applets which give an impression of the ability of computers to bring complex contents to a better understanding by interactivity.

This text will also give an idea of how to teach basic elements of mathematical stochastics by using mathematics of finance such as probability and expectation.

e-stat: Development of a Scenario for Statistics in Chemical Engineering

C. Weihs, M. Kappler

Keywords: Chemical Engineering, Design of Experiments, e-learning, Learning-Software, Multimedia, Scenario, Statistical Education

In the environment of e-stat, a learning software for statistical methods, we develop a scenario for statistics in chemical engineering. In this scenario, data analysis methods and design of experiments are applied to the optimization of the styren process. According to Moore (1997), learning by means of individual action is the most important new model of pedagogy. In particular, the combination of statistical content, learning in realistic data situations, and the use of multimedia with various interaction facilities supports effective and modern learning best. All these concepts are realized in the presented scenario utilizing modern internet technology.

e-stat: Web-based Learning and Teaching of Statistics in Secondary Schools

C. Pahl, P. Lipinski, K. Reiss

Keywords: Interactive Learning, Learning Environment, Multimedia, Statistical Education, Teaching Environment, Web-based Learning

The student course of the e-stat project, which is currently under development, will use multimedia support in order to implement an interactive, internet-based learning and teaching environment. This educational environment for learning statistics will include a variety of tools in order to support self-initiated learning processes for approaching statistical problems and finding their solutions.

EMILeA-stat: Structural and Didactic Aspects of Teaching Statistics through an Internet-based, Multi-medial Environment

U. Genschel, U. Gather, A. Busch

Keywords: Interactivity, Internet-based Learning Environments, Multimedia, Pedagogy, Statistical Education

We will illustrate the design of the written content of a new interactive software for statistical education, called EMILeA-stat (currently under development). Through our work, we suggest a way of teaching statistical theory/methodology via an internet-based, multimedial environment. To foster a greater understanding and appreciation of statistics among students, EMILeA-stat aims at achieving and maintaining a maximum level of student activity. In this talk, we choose the nonparametric Wilcoxon Rank Sum Test to demonstrate the structure and didactic features embodied by EMILeA-stat. The assumed level of instruction, called level B here, is a medium level of difficulty, to address college students not majoring in Statistics. For completeness reasons we mention that level A intends to illustrate the basic ideas, where as level C is written for mathematically orientated, more advanced students.

There have been initiatives to reform college level statistical education in the United States. In particular, the ASA has recommended that introductory statistics courses rely more on computers and automated computations (Cobb, 1992). Moore (1997) has advocated a constructive model of learning, in which students are encouraged to learn through their own activities and teachers are to de-emphasise lecture-based classroom teaching. Velleman and Moore (1996) stress that multimedial instruction can be highly interactive and provide immediate feedback, while allowing students to determine their own pace of learning.

When introducing statistics to students (especially non-majors in the subject), teachers often lament of encounters with student apathy and resentment toward statistics. Many students experience "math anxiety" or "symbol shock" (described by Tanur, 1997), often as a result of discouraging past coursework in mathematics/statistics. Success in teaching statistics requires an educator to create a level of greater accessibility and relevance in their instruction; a tenant of often overlooked "humanistic" educational psychology is that students learn best what they want and need to know (Gage and Berliner, 1991). This educational environment can be realized with EMILeA-stat. Within each presented statistical procedure, we value and attempt to incorporate "real-world" problem scenarios and data to make statistical tools more appealing to students. We design the presentation of the educational material (eg. the Wilcoxon Rank Sum Test) to allow and encourage individual student activity in learning statistics (intended for use both in and outside the classroom). The format, which we suggest, enables students to explore material at their own pace and allows for a two-way flow of information (eg. interva-

tive applet or multiple choice questions). Hence, the student users can actively process information and construct meaning for the statistical method, which is better for memory retention as well (Rumelhart and McClelland, 1986, connectionic learning model). An advantage of the multimedial environment is the capacity to provide significantly more examples than are possible in a classroom setting, where time is a practical constraint. This increases the opportunities for student exposure to statistical applications. In our internet-based teaching approach, presented material is intended and formulated to be as concise as possible. The internet provides the possibility to illustrate statistical methods, without requiring students to search through and read numerous text book pages. At level B, introduced statistical procedures are intended to give students a basic working knowledge of the statistical methodology, enabling the students to apply what they learn and to use formulas. Once an understanding of application is established for a statistical method, the student/teacher is given the possibility to explore the theoretical justification for the method at level C. We propose an interactive Java applet to enhance student understanding of the Wilcoxon Rank Sum Test and give further suggestions on effectively structuring multimedial material for students. We wish to emphasise that EMILeA-stat (though written to be a self-contained, interactive learning tool) is intended to supplement, not replace, a classroom instructor. A high level of communication between an instructor and students, allowing for the immediate exchange of ideas, is in our opinion indispensable for success in statistics eduation (cf. Moore, 1997).

Evaluating the GPH Estimator via Bootstrap Technique

S. Golia

Keywords: Bootstrap, Discrete Wavelet Transform, GPH Estimator, Long Memory Process

In this work a bootstrap method for long-memory processes is presented. The method is based on a combination of the discrete wavelet transform and the stationary bootstrap and is a development of the one presented by Percival et al. (2001). This bootstrap scheme is empirically tested calculating the GPH estimator for fractional gaussian noise time series. The set of bootstrap estimations is compared with the set of estimations obtained via Monte Carlo simulation procedure.

Evolutionary Algorithms with Competing Heuristics in Computational Statistics

J. Tvrdík, I. Křivý, L. Mišík

Keywords: Evolutionary Algorithms, Global Optimization, Heuristics, Non-linear Regression, Robust Estimates, Time Series Analysis

The paper presents a new class of evolutionary algorithms based on the competition of different heuristics. The algorithm was applied to solving some optimization problems of computational statistics, namely to estimating the parameters of non-linear regression models, constrained M-estimates and optimizing the smoothing constants in the Winters exponential smoothing. The results showed that the evolutionary algorithm with competing heuristics can be successfully used in solving some global optimization problems of computational statistics.

Exact Nonparametric Inference in R

T. Hothorn, K. Hornik

Keywords: Confidence Interval, Exact Conditional Inference, Permutation Test

For many of the classical tests, including the Wilcoxon signed rank and rank sum and the Ansari-Bradley tests, exact unconditional distributions of the test statistics can be obtained using recursion formulae provided that the underlying distribution functions are continuous. For every score function special algorithms are needed. Moreover, they are not valid for tied scores. However, the classical tests can be viewed as special cases of permutation tests. We use the shift algorithm introduced by Streitberg & Röhmel (1986) for the computation of the conditional distribution of a permutation test for integer valued scores. Implementation details and generalizations to situations with rational or real scores are given.

Exploring the Structure of Regression Surfaces by using SiZer Map for Additive Models

R. Raya Miranda, M.D. Martínez Miranda, A. González Carmona

Keywords: Additive Model, Backfitting, Bandwidth, Binning, SiZer Map

In this work we study the structure of regression surfaces by using a graphic tool called SiZer Map. This kind of graph was introduced by Chaudhuri and Marron (1999) for density and regression estimation. To estimate the regression model we use the backfitting algorithm, Buja et al. (1989), with local lineal smoothers, Opsomer and Ruppert (1997), implemented with binning methods, Fan and Marron (1994).

Fast and Robus Filtering of Time Series with Trends

R. Fried, U. Gather

Keywords: Level Shifts, Linear Regression, Outliers, Signal Extraction

Fast and robust methods are needed for preprocessing data measured with high sampling frequencies. In intensive care e.g. we observe physiological variables like the heart rate in short time intervals. Systematic changes in these data have to be detected quickly and distinguished from clinically irrelevant short term fluctuations and measurement artifacts. Median filtering works well if there is no substantial trend in the data but improvements are possible by approximating the data by a local linear trend.

Functional Principal Component Modelling of the Intensity of a Doubly Stochastic Poisson Process

A.M. Aguilera, P.R. Bouzas, N. Ruiz-Fuentes

Keywords: Doubly Stochastic Poisson Process, Functional Principal Component Analysis

An estimator for the intensity process of a doubly stochastic Poisson process is presented, having no statistical previous knowledge of it. In order to give a statistical structure of the intensity, a functional Principal Component Analysis is applied to k estimated

sample paths of the intensity built from k observed sample paths of the point process.

Growing and Visualizing Prediction Paths Trees in Market Basket Analysis

M. Aria, F. Mola, R. Siciliano

Keywords: Assortment Choices, Factorial Representation, Latent Basket, Mixing Parameter, Neural Network, Segmentation, Tree-based Models

This paper provides a new approach to Market Basket Analysis taking as measure unit the monetary value of each choice in the transaction. Furthermore, instead of association rules suitable prediction rules are defined in order to consider the causal links between different items. Two methodologies will be presented in order to grow predictive paths through oriented graphs (either trees and neural networks) which facilitate the visual comparison among different basket typologies.

Improved Fitting of Constrained Multivariate Regression Models using Automatic Differentiation

T. Ringrose, S. Forth

Keywords: Automatic Differentiation, Multivariate Regression, Response Surface Methods

Regression models for multivariate response surfaces are proposed, and an illustration is given of numerical maximization of likelihoods using automatic differentiation being superior to that using finite differences.

Imputation of Continuous Variables Missing at Random using the Method of Simulated Scores

G. Calzolari, L. Neri

Keywords: Estimation/Imputation, Missing Data, Reduced Form, Simulated Scores, Structural Form

For multivariate datasets with missing values, we present a procedure of statistical inference and state its "optimal" properties. Two main assumptions are needed: (1) data are missing at random (MAR); (2) the data generating process is a multivariate normal linear regression. Disentangling the problem of convergence of the iterative estimation/imputation procedure, we show that the estimator is a "method of simulated scores" (a particular case of McFadden's "method of simulated moments"); thus the estimator is equivalent to maximum likelihood if the number of replications is conveniently large, and the whole procedure can be considered an optimal parametric technique for imputation of missing data.

Induction of Association Rules: Apriori Implementation

C. Borgelt, R. Kruse

Keywords: Apriori Algorithm, Association Rule, Item Coding, Prefix Tree

We describe an implementation of the well-known *apriori algorithm* for the induction of association rules [Agrawal et al. (1993), Agrawal et al. (1996)] that is based on the concept of a *prefix tree*. While the idea to use this type of data structure is not new, there are several ways to organize the nodes of such a tree, to encode the items, and to organize the transactions, which may be used in order to minimize the time needed to find the frequent itemsets as well as to reduce the amount of memory needed to store the counters. Consequently, our emphasis is less on concepts, but on implementation issues, which, however, can make a considerable difference in applications.

Intelligent WBT: Specification and Architecture of the Distributed Multimedia e-Learning System e-stat

C. Möbus, B. Albers, S. Hartmann, J. Zurborg

Keywords: Architecture, Authoring Systems, Cognitive Approaches, Content-Engineering, Distributed Learning Environment, Extensible Markup Language (XML), Intelligent Distance Learning, Intelligent Web Based Training (I-WBT), Internet Environments, Learning Environments, Specification of e-Learning Systems, Unified Modeling Language (UML)

Modern e-Learning Systems (eLS) are expected to be innovative not only concerning comprehensive representation of content enriched by multimedia, but also in the integration of learning situations in contexts suitable for students. Suitable, motivating contexts can be "fun" as found in strategic games or business simulations or of a more "serious" variety in the form of virtual data labs. In the new BMBF Project EMILeA-stat (e-stat) 13 partners from different organisations are cooperating to construct such an innovative intelligent web based training (I-WBT) system for applied statistics. The German Federal Ministry of Education and Research finances e-stat by means of the NMB funding program "Neue Medien in der Bildung" (New Media in Education).

Special emphasis is placed on reuse and sharing of contents, clean separation of factual contents and its didactical motivated presentation, as well as the avoidance of proprietary solutions. E-stat is an attempt to go beyond the scope of existing WBT systems by using a strong integrating concept in combining content with a high diversity of methodical and didactical approaches. This ambitious approach creates the need for new research and evaluation. For example, a method for the presentation of coherent and user-adaptive content (learning objects) supplied by a variety of sources has to be found. The development of an eLS can be approached from different angles. This paper describes the architecture of e-stat from a knowledge and content engineering point of view, while applying cognitive-science criteria where necessary. All components were developed in a close cooperation with the statistical content providers.

Interactive Exploratory Analysis of Spatio-Temporal Data

J.M. Dreesman

Keywords: Brushing, Graphical Data Analysis, Space-Time Interaction, Spatial Modelling, Surveillance Data

An approach for the exploratory analysis of the space-time interaction in spatio-temporal-data is presented. The approach uses the interactive brushing technique, known from exploratory data analysis (EDA), in order to connect time series with maps. By applying this approach to infectious disease surveillance data, we gain further insight into the space-time-relationship of these data and obtain a tool for detecting outbreaks of infectious diseases.

Interactive Graphics for Data Mining

D. Di Benedetto

Keywords: Data Mining, Interactive Graphics, Parallel Co-ordinates, Trellis

Today information analyses require investigation of very large raw data sets and the study of relationships between numerous and different kinds of variables. Data Mining techniques have been developed for large data sets, but they produce very many results and these can be difficult to interpret. Adding interactive controls to the standard graphics can help a lot to sort through the results and understand them. This paper reviews also some tools which could give new and interesting ideas on how to develop new interactive representations, and investigates how they scale up for large data sets, with particular regard to multivariate continuous data and tools such as Trellis Displays and Parallel Coordinates, especially as presented in Manet and Cassatt interactive software.

Least Squares Reconstruction of Binary Images using Eigenvalue Optimization

S. Chrétien, F. Corset

Keywords: Combinatorial Optimization, Eigenvalue Optimization, Least Squares Estimation, Semi-definite Programming

The goal of this paper is to present an application to binary image least-squares estimation of some recent results in the semi-definite programming approximation theory of some combinatorial problems due to Goemans and Williamson, Yu Nesterov and others. In particular, we show in a very simple fashion that a good suboptimal solution may be obtained via eigenvalue optimization.

Locally Adaptive Function Estimation for Categorical Regression Models

A. Jerak, S. Lang

Keywords: Adaptive Smoothing, MCMC, Nonparametric Bayesian Regression, Random Walk Priors, Unsmooth Functions, Variable Smoothing Parameter

In this paper we present a nonparametric Bayesian approach for fitting unsmooth or highly oscillating functions in regression models with categorical responses. The approach extends previous work by Lang et al. (2002) for Gaussian responses. Nonlinear functions are modelled by first or second order random walk priors with locally varying variances or smoothing parameters. Estimation is fully Bayesian and uses latent utility representations of categorical regression models for efficient block sampling from the full conditionals of nonlinear functions.

Maneuvering Target Tracking by using Particle Filter Method with Model Switching Structure

N. Ikoma, T. Higuchi, H. Maeda

Keywords: Bayesian Modeling, Multiple Model, Non-Gaussian Distribution, Particle Filter, Switching Structure, Target Tracking

Tracking problem of maneuvering target is treated with assumption that the maneuver is unknown and its acceleration has abrupt changes sometimes. The cope with unknown maneuver, Bayesian switching structure model, which includes a set of possible models and switches among them, is used. It can be formalized into general (nonlinear, non-Gaussian), state space model where system model describes the target dynamics and observation models represents a process to observe the target position. Heavy-tailed unimodal distribution, e.g. Cauchy distribution, is used for the system noise to accomplish good performance of tracking both for constant period and abrupt changing time point of acceleration. Monte Carlo filter, which is a kind of particle filter that approximates state distribution by many particles in state space, is used for the state estimation of the model. A simulation study shows the efficiency of the proposed model by comparing with Gaussian case of Bayesian switching structure model.

mathStatica: Mathematical Statistics with *Mathematica*

C. Rose, M.D. Smith

Keywords: Computational Statistics, *Mathematica*, mathStatica

This paper presents **mathStatica** (2002), a completely general toolset for doing mathematical statistics with *Mathematica* (Version 4). **mathStatica** defines statistical operators for taking expectations, finding probabilities, deriving transformations of random variables and so on. Importantly, **mathStatica** is not tied to a set of pre-specified statistical distributions. Rather, it is designed to derive statistics such as moments, cumulative distribution functions, characteristic functions, and other generating functions for user-defined distributions. **mathStatica** supports discrete and continuous distributions – univariate and multivariate. Applications to inference include: estimation (moment unbiased, minimum variance unbiased, best unbiased, maximum likelihood: symbolic and numeric), curve-fitting (Pearson and Johnson systems, non-parametric kernels), asymptotics, decision theory, and moment conversion formulae (for conversion between cumulants, raw moments, and central moments: univariate and multivariate). **mathStatica** accompanies the book: Rose and Smith (2002), Mathematical Statistics

with *Mathematica* (Springer Texts in Statistics).

MCMC Model for Estimation Poverty Risk Factors using Household Budget Data

E. Käärik, E.-M. Tiit, M. Vähi

An important task for analysts of official statistical data is the determination of social risk groups, especially in poverty risk. As the data used in such analysis (Household Income and Expenditures surveys) have rather high non-response rate and may be affected by different sources of errors, the problem of model stability is important. One possibility to check the stability of a model is to use different generation procedures.

MD*Book online & e-stat: Generating e-stat Modules from LaTeX

R. Witzel, S. Klinke

Keywords: Conversion from LaTeXto HTML, Conversion from LaTeXto XML, E-Books, Internet, MD*Book, MD*Book online, MM*Stat, MM*Stat

We describe the basics of MD*Book and MD*Book online which we use together with XploRe to create interactive electronic books. We generate different output formats (Postscript, PDF, HTML and MM*Stat-like formats) for the internet as well as for CD and/or local installation. The e-stat project for creating a knowledge base in statistics needs statistical input from various authors to cover a larger area in statistics. The teaching material, which we want to incorporate in the e-stat project, is already available in LaTeXand we have used MD*Book to convert it from LaTeXto HTML. Interactivity in our electronical books is incorporated via a applet which connects to XploRe – a statistical software. Thus we extended our tool MD*Book, and therefore also MD*Book online, such that we can generate the XML file (and further files) which are the base for an e-stat module. We first describe MD*Book and MD*Book online. Finally we describe the conversion process and show an reused example from MM*Stat as e-stat module.

Missing Data Incremental Imputation through Tree Based Methods

C. Conversano, C. Cappelli

Keywords: Data Mining, Lexicographic Order, Missing Data, Nonparametric Imputation, Tree-Based Methods

Conditional mean imputation is a common way to deal with missing data. Although very simple to implement, the method might suffer from model misspecification and it results unsatisfactory for non linear data. We propose the iterative use of tree based models for missing data imputation in large data bases. The proposed procedure uses *lexicographic order* to rank missing values that occur in different variables and deals with these incrementally, i.e, augmenting the data by the previously filled in records according to the defined order.

Missing Values Resampling for Time Series

A.M. Alonso, D. Peña, J.J. Romo

Keywords: Blockwise Bootstrap, Blockwise Jackknife, Missing Values, Time Series

In this work we are interested in the moving block jackknife (MBJ) introduced in Künsch (1989) and independently in Liu and Singh (1992). This method is valid to estimate the variance of statistics defined by functionals of finite dimensional marginal distributions. As it is usual in the jackknife methods, the variance estimator is obtained by a weighted sample variance of the statistic evaluated in a sample where some observations (blocks of consecutives observations, in this case) are deleted or downweighted.

For independent data deleting observations is equivalent to assume that these observations are missing, but for autocorrelated data, as shown in Peña (1990), both procedures are very different. Treating the block as missing values implies to substitute the observations in the block by their conditional expectations given the rest of the data. This is the procedure that we propose in the paper. In our approach, the observations left out in the MBJ are considered as missing observations and they are substituted by a missing value estimates which takes into account the data's dependence structure. Then, the variance estimator is a weighted sample variance of the statistics evaluated in a "complete" series. This procedure could be interpreted as an smooth transition between the two parts with full weight in the MBJ. We also extend this missing values replacement

to the moving blocks bootstrap (MBB), in such a case the procedure resemble to a block joining engine similar to the matched-block bootstrap of Carlstein et al (1996).

We establish the consistency of the proposed methods as variance and distribution estimators of the sample mean assuming some moment and mixing restrictions. Also we perform an extensive Monte Carlo study comparing the proposed methods with the MBJ and MBB of Künsch (1989). The results point out that missing values approaches generally improve the preceding methods.

ModelBuilder – an Automated General-to-specific Modelling Tool

M. Kurcewicz

Keywords: General-to-specific modelling, Money Demand

The general-to-specific methodology (also known as the London School of Economics, LSE, methodology) is one of the most widely used methods of econometric model construction. Recently, specification search algorithms that automate some stages of the general-to-specific methodology have been proposed. In this paper we present an enhanced specification search algorithm. Extensions include cointegration analysis and automated construction of Error Correction Models. Although, specification search algorithms were originally developed to analyse how well does the LSE modelling approach work in controlled conditions, they are also a valuable tool for empirical research. As an empirical example, an automated construction of a demand function for narrow money in Austria is presented. Results are similar to those obtained in a more traditional way – a stable money demand function is found.

On the Use of Particle Filters for Bayesian Image Restoration

K. Nittono, T. Kamakura

Keywords: Bayesian Framework, Image Restoration, Particle Filter, Sequential Monte Carlo Method

Image restoration is one of the wide variety of branches in image analysis. In recent work, the Bayesian approach to the restoration has attracted interest and much of this work involves the use of statistical modeling for images assuming Markov random fields (MRF), the stochastic technique based on Monte Carlo methods and maximum a poste-

riori (MAP) estimation. The Bayesian approach ordinarily needs relatively large amount of computation time, especially for the type of problems of image processing because of high dimensionality and its massive configuration space. Geman and Geman (1984) proposed an approach to find MAP estimate for any given images. Due to the feature of Gibbs sampler along with simulated annealing, the method enables to find global maxima as MAP estimate, however, in practical situations the optimal annealing schedule for its convergence is not definite and in general it needs much iterative update calculation. Besag (1986) proposed an opposite scheme of iterated conditional modes (ICM). The scheme can considerably reduce computation time with the simple criterion for updating calculation, however, the restored images tend to retain relatively higher misclassification depending on the initial configuration or given images. In this paper, we aim to study the application of particle filters to Bayesian image restoration. The particle filter is a Monte Carlo method for computing complex posterior distributions, and it provides simple and flexible algorithms for the computing (Doucent et al., 2001). Involving the virtue of the convenient algorithms of the particle filters, we attempt to compose a new method for Bayesian image restoration while using tractable algorithm.

Optimally Trained Regression Trees and Occam's Razor

P. Savický, J. Klaschka

Keywords: Bottom-up Algorithms, Dynamic Programming, Generalization, Occam's Razor, Optimization, Recursive Partitioning, Regression Trees

Two bottom-up algorithms growing regression trees with the minimum mean squared error on the training data given the number of leaves are described. As demonstrated by the results of experiments with simulated data, the trees resulting from the optimization algorithm may have not only better, but also worse generalization properties than the trees grown by traditional methods. This phenomenon is discussed from the point of view of Occam's razor principle.

Parallel Algorithms for Inference in Spatial Gaussian Models

M. Whiley, S.P. Wilson

Keywords: Bayesian Inference, Matrix Inversion, MCMC, Parallel Algorithms, Spatial Modelling

Markov chain Monte Carlo (MCMC) implementations of Bayesian inference for latent spatial Gaussian models are very computationally intensive, and restrictions on storage and computation time are limiting their application to large problems. Here we propose three parallel algorithms for linear algebra calculations in these implementations. The algorithms' performance is discussed with respect to a simulation study, which demonstrates the increase in speed and feasible problem size as a function of the number of processors. We discuss how parallel algorithms may be useful more generally in MCMC schemes for these models.

Parameters Estimation of Block Mixture Models

M. Nadif, G. Govaert

Keywords: Block Mixture Model, EM Algorithm, Fuzzy Block Clustering

While most of the clustering procedures are aimed to construct an optimal partition of objects or, sometimes, of variables, there are other methods, called block clustering methods, which consider simultaneously the two sets and organize the data into homogeneous blocks. Recently, we have proposed a new mixture model called *block mixture model* which takes into account this situation. This model allows to embed simultaneous clustering of objects and variables in a mixture approach. We have considered this probabilistic model under the classification likelihood approach and we have developed a new algorithm for simultaneous partitioning based on the Classification EM algorithm. In this paper, we consider the block clustering problem under the maximum likelihood approach and the goal of our contribution is to estimate the parameters of this model. To reach this aim, we use an approximation of the likelihood and we propose an alternated-optimization algorithm to estimate the parameters, and to illustrate our approach, we study the case of binary data.

Pattern Recognition of Time Series using Wavelets

E.A. Maharaj

Keywords: Randomization Test, Stationary and Non-Stationary Time Series, Wavelet Coefficients

In this paper a pattern recognition procedure for time series using wavelets is developed. This is done by means of a randomization test based on the ratio of the sum of squared wavelet coefficients of pairs of time series at different scales. A simulation study using pairs of stationary and non-stationary time series and using the *Haar* and *Daubechies* wavelets reveals that the test performs fairly well at scales where there are a sufficient number of wavelet coefficients. The test is applied to a set of financial time series.

Representing Knowledge in the Statistical System Jasp

I. Kobayashi, Y. Yamamoto, T. Fujiwara

Keywords: Class Based Object, Jasp, Rule, Statistical Knowledge

We describe a framework to assist processes of analyzing data in our statistical system Jasp. Recent statistical systems provide so many statistical methods that most users have di.culties to master how to use them properly. In addition, users are sometimes in danger of swallowing results from systems without thinking deeply. In order to prevent these problems, we have implemented rules of "condition - action" forms in Jasp classes to express heuristic knowledge for statistical analysis. By using these rules, Jasp can give advice to users about possible statistical analysis procedures and notify problems when they appear. This ability is useful for users, especially for students and novices in statistics or the Jasp system.

Robust Estimation with Discrete Explanatory Variables

P. Čižek

Keywords: Discrete Explanatory Variables, Least Trimmed Squares, Linear Regression, Robust Statistics

The lest squares estimator is quite sensitive to data contamination and model misspecifi-

cation. This sensitivity is addressed by the theory of robust statistics which builds upon parametric specification, but provides methodology for designing misspecification-proof estimators by allowing for various "departures" of subsets of the data. Unfortunately, most of highly robust estimators developed within robust statistics cannot be easily applied to models containing binary and categorical explanatory variables. Therefore, we design a robust estimator based on least trimmed squares that can be used for any linear regression model no matter what kind of explanatory variables the model contains. Additionally, we propose an adaptive procedure that maximizes the efficiency of the proposed estimator for a given data set while preserving its robustness.

Robust Principal Components Regression

S. Verboven, M. Hubert

Keywords: High-Dimensional Data, Principal Component Analysis, Principal Component Regression, Robust Methods

We consider the multivariate linear regression model with p explanatory variables X and $q \geq 1$ response variables Y. Moreover we assume that the regressors multicollinear. This situation often occurs in the calibration of chemometrical data, where the X-variables correspond with spectra that are measured at many frequencies. It is well known that the classical least squares estimator has a large variance in the presence of multicollinearity. Therefore many biased estimators have been proposed. A very appealing method is principal components regression (PCR) since it is easy to understand and to compute.

PCR first constructs a new set of uncorrelated explanatory variables, which are called the principal components. They correspond to the eigenvectors of the sample covariance matrix of the X-variables. The response variables are then regressed on these components using the (multivariate) Least Squares Estimator.

Both stages of this procedure are however very sensitive to the presence of outliers in the data. We present a robust principal components regression method which also consists of two steps. First a robust principal components analysis is applied to the X-variables (Hubert and Rousseeuw, 2002), yielding a smaller set of k orthogonal regressors. We then regress the response variable Y on these regressors using the multivariate MCD-regression method (Rousseeuw et al. 2000).

To select the number of regressors, we propose a robust R^2. It is roughly defined as the proportion of the robust variance of Y which is explained by the robust fit. Moreover we propose several diagnostic plots, which allow us to visualize and to distinguish the outliers in the data.

Robust Time Series Analysis through the Forward Search

L. Grossi, M. Riani

The forward search (FS) is a powerful general method for detecting masked multiple outliers and for determining their effect on models fitted to the data (Atkinson and Riani, 2000). This method was originally introduced for models which assumed independent observations: linear and non linear regression, generalized linear models and multivariate analysis. In this paper we extend the forward search technique to the analysis of time series data. The basic ingredients of the FS are a robust start from an outlier-free subset of observations, a criterion for progressing in the search, which allows the subset to increase by one or more observations at each step, and a set of diagnostic tools that are monitored along the search. The robustness of the FS stems from the very definition of its algorithm, starting from "good" data points and including outliers at the end of the procedure. Computation of high-breakdown estimators is not required, except possibly at the starting stage. Indeed, the application of efficient likelihood or moment based methods at subsequent steps of the FS provides the analyst with more powerful tools than those obtained via traditional high-breakdown estimation.

The flexibility of the FS makes this procedure suited for extensions to areas other than multiple regression and multivariate analysis. This is especially true in time series, where it is often necessary to detect and model sudden or unexpected events. It is well known that outliers or structural changes in the observed time series may seriously damage identification and estimation of the suggested ARIMA or structural model (e.g. Koopman and Harvey, 1992; Chen and Liu, 1993), because they can introduce serious bias in the sample autocorrelation function. In time series the standard procedures for automatic outlier detection and correction consider four types of outliers, namely: additive (AO), innovational (IO), level shift (LS), and transitory change (TC) (Tsay, 1986; Chen and Liu, 1993). The AO represents a single spurious observation, the IO a pulse shock to the noise sequence which propagates to the observed time series, the LS a step function and the TC a spike that takes a few periods to disappear.

The main drawback of the procedures for detecting outliers in time series is that they start with the specification of a model for the observed series as if there were no outliers. In the ARIMA approach these procedures repeatedly use $T \times 4$ times the filter $\Pi = \phi(B)/\theta(B)$ based on the parameters estimated using all the observations, to obtain a statistic λ for each observation $yt, t = 1, \ldots, T$ and each type of outlier AO, IO, LS and TC. The maximum value of λ is compared with a predetermined critical value C to decide whether a potential outlier is present in the time series. Once an outlier is found, the filter Π is revised and reapplied for each observation and for each type of outlier in an iterative way. These procedures do not detect the k outliers at once, but proceed in several iterations detecting them one by one. Given that these approaches start with estimated parameters based on all T observations, as if no outlier were present in the data, it is clear that they may suffer from the well known masking and swamping

effects. In order to avoid these problems and to be able to detect stretches of over influential observations, Bruce and Martin (1989) suggested leave-k-out diagnostics. The parameters of the model fitted to the full data set are compared with those generated by fitting the model to the data when a stretch of k points are taken to be missing. This method however, becomes computationally infeasible when k is large and may still suffer from masking and swamping effects if the number of outliers is greater than k.

In this paper we show how the forward search, free from masking and swamping problems, can detect masked multiple outliers in time series and their effect on the model fitted to the data.

Rough Sets and Association Rules – Which is Efficient?

D. Delic, H.-J. Lenz, M. Neiling

Keywords: Apriori-Algorithm, Association Rules, Bitmaps, Hybrid Association Rules Induction Scheme, Rough Sets

We evaluate the rough set and the association rule method with respect to their performance and the quality of the produced rules. It is shown that despite their different approaches, both methods are based on the same principle and, consequently, must generate identical rules. However, they differ strongly with respect to performance. Subsequently an optimized association rule procedure is presented which unifies the advantages of both methods.

Skewness and Fat Tails in Discrete Choice Models

R. Capobianco

Keywords: Dichitomous Models, Fat Tails, Skewness, Student t Distribution

In discrete choice models, the probability that the dependent variable will assume value 0 or 1 depends on a set of explanatory variables through a function F. In this paper we propose the class of skew Student t distribution as a function F in order to have a more flexible model that can simultaneously account for asymmetry and thick tails and such that probit and logit models can be considered as special cases. Two examples illustrate the performance of the proposed model.

Standardized Partition Spaces

U. Sondhauß, C. Weihs

Keywords: Classification, Experimental Design

We propose a standardized partition space (SPS) that offers a unifying framework for the comparison of a wide variety of classification rules. Using SPS, one can define measures for the performance of classifiers w.r.t. goodness concepts beyond the expected rate of correct classifications of the objects of interest. These measures are comparable for rules from so different techniques as support vector machines, neural networks, discriminant analysis, and many more. In particular, we are interested in assessing the reliability of classification rules when used for interpretation of the relationship between the values of predictors and the membership in classes.

We will demonstrate the high potential of SPS for the comparison of classification methods in a simulation study to analyse the following problem:

Given a medium number of predictors, (10-20), and a potentially complex relation between classes and predictors, one would expect flexible classification methods like support vector machines or neural networks to do better than simple methods like e.g. the linear discriminant analysis or classification and regression trees. Nevertheless, one often observes on real data sets, that the simple procedures do pretty well. Our assumption is, that simple methods are more robust against instability, and that the effect of instability superposes the effect of complexity of the relation. By instability we mean the deviation from the assumption that the collected data is some independent and identically distributed sample from some joint distribution of predictors and classes.

We analyse this problem with a simulation study using experimental design.

StatDataML: An XML Format for Statistical Data

D. Meyer, F. Leisch, T. Hothorn, K. Hornik

Keywords: Data Design, Data Exchange, XML

In order to circumvent common difficulties in exchanging statistical data between heterogeneous applications (format incompatibilities, technocentric data representation), we introduce an XML-based markup language for statistical data, called StatDataML. After comparing *StatDataML* to other data concepts, we detail the design which borrows from the language S, such that data objects are basically organized as recursive

and nonrecursive structures, and may also be supplemented with meta-information.

Statistical Computing on Web Browsers with the Dynamic Link Library

A. Takeuchi, H. Yadohisa, K. Yamaguchi, C. Asano, M. Watanabe

Keywords: DLL, DLLSA/QC, Statistical Software, World Wide Web

A statistical computing system using a user-friendly graphical user interface (GUI) to a pre-defined statistical engine is proposed. The GUI, which is constructed using a Web browser interface, provides access to a statistical engine program that has been implemented as a dynamic link library. Since a browser is used, the GUI can be constructed using popular and free tools such as HTML. Using HTML, we can describe the various information appropriate for the user's level and purpose, together with analysis tools.

Statistical Inference for a Robust Measure of Multiple Correlation

C. Dehon, C. Croux

Keywords: Confidence Interval, F-Statistic, Multiple Correlation, R-Squared Statistic, Regression Analysis, Robustness

In regression analysis it is standard practice to report an R-statistic. This measure of multiple correlation is based on the Least Squares estimator, which is known to be extremely sensitive to outliers. The associated R-squared suffers from the same lack of robustness. Also the value of the F-statistic, used to see whether the explanatory variables have jointly a significant influence on the dependent variable, exhibits similar problems. Many robust estimation procedures are available now, but there has been less emphasis on the development of the inference part and in particular on the study of robust R-squared measures. It is however clear that applied statisticians desire to have the same tools to validate their model as when using the classical least squares estimator. In this note we will study the stability properties of robust measures of multiple correlation, and also investigate the size and power of F-statistics based on them.

Statistical Software VASMM for Variable Selection in Multivariate Methods

M. Iizuka, Y. Mori, T. Tarumi, Y. Tanaka

Keywords: Factor Analysis, Principal Component Analysis, Statistical Tool, Web-based Program

A statistical software package VASMM (VAriable Selection in Multivariate Methods) has been developed for selecting a subset of variables in multivariate methods without external variables. The current version is fully implemented for variable selection in principal component analysis and factor analysis. The system has been constructed with interactive architecture on Internet. The users can not only use the system via a web browser but can also obtain information related to variable selection in multivariate techniques of their choice. It allows for us to perform variable selection easily in a variety of practical applications.

Structural Equation Models for Finite Mixtures – Simulation Results and Empirical Applications

D. Temme, J. Williams, L. Hildebrandt

Keywords: Finite Mixtures, Model-based Clustering, Monte Carlo Simulation, Structural Equation Modeling, Unobserved Heterogeneity

Unobserved heterogeneity is a serious, often neglected problem in structural equation modeling (SEM). Recently, a finite mixture approach to SEM has been proposed to solve this problem but until now only a few studies analyse the performance of this approach. The contribution of this paper is twofold: First, results from a Monte Carlo study into the properties of the software program MECOSA are presented. Second, an empirical application to data from a large-scale consumer survey in the fast moving consumer goods industry is described.

Sweave: Dynamic Generation of Statistical Reports using Literate Data Analysis

F. Leisch

Keywords: Integrated Statistical Documents, Literate Statistical Practice, R, Reproducible Research, S

Sweave combines typesetting with LaTeXand data analysis with S into integrated statistical documents. When run through R or S-plus, all data analysis output (tables, graphs,...) is created on the y and inserted into a final LaTeXdocument. Options control which parts of the original S code are shown to or hidden from the reader, respectively. Many S users are also LaTeXusers, hence no new software has to be learned. The report can be automatically updated if data or analysis change, which allows for truly reproducible research.

Testing for Simplification in Spatial Models

L. Scaccia, R.J. Martin

Keywords: Autoregressive Process, Axial Symmetry, Doubly-Geometric Process, Lattice Process, Separability, Spatial Process

Data collected on a rectangular lattice occur frequently in many areas such as field trials, geostatistics, remotely sensed data, and image analysis. Models for the spatial process often make simplifying assumptions, including axial symmetry and separability. We consider methods for testing these assumptions and compare tests based on sample covariances, tests based on the sample spectrum, and model-based tests.

The Forward Search

A. Atkinson

Keywords: Box-Cox Transformation, Fan Plot, Forward Search, Masked Outliers, Robustness

This paper summarises joint research with Marco Riani on the forward search, a powerful

general method for detecting multiple masked outliers and for determining their effect on models fitted to the data. Atkinson and Riani (2000) describe its use in linear and nonlinear regression, response transformation and in generalized linear models. These examples are here extended to include multivariate analysis. Riani and Atkinson (2001) describe an application to multivariate transformations and discriminant analysis.

The MISSION Client: Navigating Ontology Information for Query Formulation and Publication in Distributed Statistical Information Systems

Y. Bi

Keywords: Dissemination, Graphical User Interface, Integration of Distributed Statistical Information System, Metadata, Ontology, XML

This paper describes the technologies developed for the summary and publication of large volumes of statistical information, such as produced by National Statistical Institutes (NSIs), in the European fifth framework project, MISSION (Multi-Agent Integration of Shared Statistical Information Over the [inter]Net). We review the MISSION system architecture that is built on a three-tier client/server architecture, and then focus on novel methods and techniques in designing and implementing different client components, and data model for ontologies within the library.

Time Series Modelling using Mobile Devices and Broadband Internet

A. Prat

Keywords: ARIMA Models, Forecasting, Internet2, Seasonal Adjustment, Security, Smart Cards, UMTS

In this paper we present one application included in the Project @DAN, an AdvANced and high secure mobile platform to support the digital economy, that started on November 2001, and is financed by the *EU IST-2001-32634*.

This application will provide all types of services necessary to model and obtain forecasts of time series data by the professionals of any kind of organisations. The user of this application will be able to obtain two different types of services: demand forecasts and

consultancy using mobile terminals, broadband internet and internet payments.

The background system installed in a server, is FOREtess*, which is an evolution of TESS (see Prat et al. (2000)).

Unbiased Partial Spline Fitting under Autoregressive Errors

M.G. Schimek

Keywords: AR Errors, Partial Spline, Semiparametric Regression, Smoothing Parameters, Time Series, Unbiased Estimation

A special variant of semiparametric regression models are partial spline models. Schimek (2000) provides a cheap algorithm for unbiased estimation under an iid assumption. Here we introduce a similar algorithm for the case of autoregressive (AR) errors. Further we address the problem of smoothing parameter choice in this situation. Finally the results are illustrated on time series data from environmental epidemiology.

Unobserved Heterogeneity in Store Choice Models

I.R. del Bosque, A. Suárez-Vázquez, I. Moral-Arce, J.M. Rodríguez-Póo

Keywords: Heterogeneity, Random Effect Models, Shopping Centre Choice

Discrete choice models are a tool which is shared by a large variety of scientific areas, and for this reason, different specifications have been developed. Of all these specifications, the so-called multinomial logit model (McFadden, 1974) is the one that has been used most to analyze situations in which there are multiple choice alternatives, specifically in retail studies (Arentze y Timmermans, 2001). The multinomial logit model has a strong theoretical base and is also easy to apply. However, a potential limitation of the studies that use this specification is the fact that consumer heterogeneity is not taken into account (Severin, Louviere and Finn, 2001). In this paper we present the results of a random effects model. The distribution of the heterogeneity is assumed to be unknown, and it is approximated through a piecewise constant function, as was proposed originally in Heckman and Singer (1984).

*FOREtess is in process of being registered as a trademark by UPC.

Using the Forward Library in S-plus

K. Konis

Keywords: Forward Library, Forward Search, Robustness

The Forward Library is an S language implementation of the forward search described in Atkinson and Riani (2000). Software is provided for conducting the forward search in linear regression, in generalised linear models, and in analysing the effect of outliers on the Box-Cox transformation of the response in linear regression. The examples provided in this paper demonstrate how to conduct a forward search using both the graphical user interface (Windows) and the command line interface (Windows and Unix/Linux). S-plus version 6 is assumed.

Variance Stabilization and Robust Normalization for Microarray Gene Expression Data

A. von Heydebreck, W. Huber, A. Poustka, M. Vingron

Keywords: Gene Expression, Microarray, Robust Regression, Variance Stabilization

We introduce a statistical model for microarray gene expression data that comprises data calibration, the quantification of differential gene expression, and the quantification of measurement error. In particular, we derive a transformation h for intensity measurements, and a difference statistic $\triangle h$ whose variance is approximately constant along the intensity range. The parametric form $h(x) = \mathrm{arsinh}(a + bx)$ is derived from a model of the variance-versus-mean dependence for microarray intensity data, using the method of variance stabilizing transformations. The parameters of h together with those of the calibration between experiments are estimated with a robust variant of maximum-likelihood estimation.

Weights and Fragments

S. Morgenthaler

Keywords: Model Break, Parsimony, Robustness, Structural Change

This talk explores two separate but related ideas. The first one is to let the data guide us in downweighting observations in order to improve the fit of the model. Such procedures can, for example, be used to identify subsets of the data that closely fit a particular regression model and as such provide an alternative to robust methods. The second idea consists in fragmenting the data into pieces, analyzing the pieces separately and constructing a global analysis by putting the fragments back together. In analyzing a two-way table, for example, different interaction patterns may apply to different parts of the table.

XQS/MD*Crypt as a Means of Education and Computation

J. Feuerhake

Keywords: Distributed Computing, Education, MD*Crypt, XQS

This paper introduces MD*Crypt as a means of distributing statistical methods and computing power. This paper explores MD*Crypt regarding topics of education. The role of MD*Crypt in projects associated with teaching statistics is examined. In an outlook, the potential of the MD*Serv/MD*Crypt architecture in frameworks of distributed computing and method providing is mentioned.

Author index

Keyword index