Interface '99: A Data Mining Overview

Arnold Goodman
The UCI Center for Statistical Consulting
University of California, Irvine
Irvine, CA 92697-5105
agoodman@uci.edu

ABSTRACT

This personal overview of Interface '99 is intended to communicate its meaning and relevance to SIGKDD, as well as provide valuable information on trends within the Interface for data miners seeking to learn more about statistics. In addition, it is the newest link in a bridge between the Interface and KDD begun by References 2-4 and the sessions on KDD at Interface '98 and Interface '99.

Keywords

Review of Interface'99 conference, statistics.

1. INTRODUCTION

KDD uses computer technology and statistics to discover knowledge buried deep within enormous databases that are usually heterogeneous and not from designed experiments. Although this situation is outside the bounds of statistical traditions, it is not outside the bounds of statistical techniques or thinking. KDD is actually part of the Interface of Computing and Statistics, and the Annual Symposia on the Interface are quite relevant to SIGKDD.

In fact, the Keynote Addresses at Interfaces '97, '98 and '99 all dealt with KDD: Jerry Friedman said, "Statistics is no longer the only data game in town"; David Rocke said, "The algorithm is the estimator"; and Leo Breiman said, "Accuracy vs. simplicity and high vs. low dimensionality should be analyzed for competing models of the data."

Reference 1 contains the Proceedings of Interface '98, while References 2-4 provide summaries of it and a short history of the Interface. Reference 5 contains the Proceedings of Interface '99, and this article provides a brief interpretive overview of it.

Interface '00 is in New Orleans on April 5-8, 2000 and features Modeling of the Earth's Systems from the Physical to the Infrastructural with Sessions on Archeology, Agriculture, Atmospherics, Business, Environment, Health and Information Technology. It is Co-Chaired by Sallie Keller-McNulty, Vicki Lancaster and Sally Morton (http://neptuneandco.com/interface/).

Interface '01 will be in Orange County, California on June 14-16, 2001 and will feature Challenges for a New Century with Sessions on Mega Datasets, Computational Finance, Mining of Web Data, Statistical Models for Text, Combining Models, Image Analysis, Graphics, Visualization, Beyond Correlation, Bayesian Analysis and Decisions, Evaluation of Models, Software Engineering, Trees, Clustering and Statistical Computing Pioneers plus a Mini Conference on Bioinformatics at the Interface. It is Co-Chaired by Padhraic Smyth and the author (http://www.ics.uci.edu/interface/).

2. SUMMARY

Interface '99 was the 31st Interface Symposium and featured Models, Predictions and Computing. It was Co-Chaired by Kenneth Berk and Mohsen Pourahmadi.

Session topics of most interest to SIGKDD were: Testing & Validation of Software, Classification & Regression Trees, Clustering & Classification, Bootstrap, Visualization, Markov Chain Monte Carlo, Image Analysis, Mixed-Effects Models, Data Mining and Computational Biology. In addition, John Elder and the author organized two KDD Sessions: The Best of KDD-98 with John Elder's "An Overview of KDD-98," Greg Ridgeway's "The State of Boosting" and Pedro Domingos' (Prize Paper) "Occam's Two Razors;" and Data Mining on the Interface with Usama's Fayyad's (Tutorial) "Scaling to Large Databases and the Link between Statistics and Databases."

3. PERSONAL PERSPECTIVE

It is interesting that the Sessions on Testing & Validation of Software and on Mixed-Effects Models echoed the author's solutions to problems in the early 1960's, and that the Sessions on Visualization echoed his major professor's solution to a problem of his in the early 1970's. Two recent talks by recognized experts in Bayesian analysis have also provided the author with new insights into both the past and future of this flowering area.

Evaluating statistical software has a long history, beginning with the 1962 checklist and scorecard that was developed by Robert McCornack and the author (Reference 6). This was the first scorecard for evaluating statistical software and perhaps the first scorecard for evaluating software in general.

The actual estimation of mixed-effects models with both fixed and random components: first estimates the fixed component assuming simplicity and a knowledge of the random component, then estimates the random component assuming knowledge gained about the fixed component, and repeats this sequence iteratively until the situation sufficiently stabilizes. This is the application of a procedure very similar to one developed by the author in 1964 for a somewhat easier least squares problem (Reference 7).

Herman Chernoff introduced a clever graphical method to represent vectors of data by facial expressions, called Chernoff's Faces in the literature, around ten years before computer graphics and twenty years before computer visualization (Reference 8). This technique plus sequential analysis, another specialty of his, should prove quite useful to the explorations within KDD.

The perspective of Arnold Zellner, who took Bayes into novel areas of econometrics long before it became fashionable, contrasted with that of Eric Horvitz, who is now taking Bayes into

novel areas of computer science. They inspired the following conjecture: although the past of Bayesian analysis was enabled by methodology while bring static and non interactive, the future of Bayesian analysis might well be enabled by technology while being dynamic and interactive.

4. THE CONFERENCE CONTENT

4.1 Testing & validation of software

"AETGWeb - A Combinatorial Approach to Designing Test Cases for Testing Software" by Sid R. Dalal, A. Jam, G. Patton and M. Rathi at Telcordia Technologies (Formerly Bellcore)

Combinatorial design is a relatively new approach to test generation and AETGWeb uses a new set of combinatorial design algorithms to reduce the number of tests. In unit and system testing, it compares favorably to the standard experimental-design approaches

"Statistical Reference Datasets (StRD) for Assessing the Numerical Accuracy of Statistical Software" by W.F. Guthrie, J.E. Rogers, J.J. Filliben, L.M. Gill, E. Lagergren and M.G. Vangel at NIST.

NIST has developed a web site called Statistical Reference Datasets (StRD) that provides both reference data sets for a variety of statistical methods and the certified values of commonly computed statistics for a user to compare with. Described are the: sources of numerical inaccuracy in mathematical and statistical software, selection of reference data sets for statistical procedures, computation of error-free statistical results for each data set, and use of the StRD web site

"FDA and Software Validation" by Ghanshyam Gupta and Peter A. Lachenbruch at FDA.

The paper discusses things to validate, who is responsible for validation, and what the FDA needs to see in order to have assurance that the software is functioning as purported. Ability to trace subsetting, accuracy and missing values in data translation from one program to another and any output 'prettyfiers' are also important.

4.2 Classification & regression trees

"Hybrid CART-Logit and CART-Neural Nets for Classification and Regression" by Dan Steinberg at Salford Systems and CSUSD with Nicholas Scott Cardell

A method introduced here differs considerably from previous hybridization experiments in that the logistic component is not run within CART child nodes and it does include a complete summary of the CART tree in its covariate set. Monte Carlo tests are reported and data mining examples are discussed.

"Hybrid and Sequential Modeling with Trees" by Thomas W. Miller at University of Wisconsin-Madison

This paper considers modeling practices described as a sequence of simple modeling steps vs. traditional stepwise methods and practices that can be automated or partially automated by computer. The objective is to explore what effects modeling practices have upon estimation bias and variance, what statistical techniques work well in combination with other techniques, and how one can identify good modeling practices.

"A Comparison of Classifiers" by Wei-Yin Loh at University of Wisconsin – Madison, Hyunjoong Kim at Worcester Polytechnic Institute and Yu-Shan Shih at National Chung Cheng University.

Thirty-four classifications are compared on thirty-two datasets in terms of misclassification error rate, training time and classifier complexity. In general: classification trees with linear splits tend to have lower errors rates than trees with univariate splits, training times of classifiers range from seconds to days, and complexity varies with some algorithms yielding trees with very large numbers of leaves.)

4.3 Clustering & Classification

"Remarks on Fitting and Interpreting Mixture Models" by David W. Scott at Rice University

The paper introduces a least squares method of fitting normal mixture models using a nonparametric criterion, plus a case study comparing the new and EM algorithms for the analysis of household income distributions. Interpretability of components is the focus of this study and a particularly interesting extension to single component estimation is also described.

"Fluorescence Spectroseopy, Quantitative Pathology and Classification of Tissue" by Dennis D. Cox at Rice University and NCAR

Linear discriminant analysis, neural nets, naive Bayes, classification trees and ensemble methods are used to classify cervical tissue based on output from a spectroscopic probe. The main difficulty is dimensionality of the feature vector, but further complications include replicated measurements and confounding variables.

"Simplifying Mixture Models with Applications" by William Szewczyk at NSA with David W. Scott at Rice University

A primary challenge is to determine an appropriate number of components, which often involves identifying those components that are close enough to be considered the same. The paper introduces a new easily-calculated measure of similarity between distributions and illustrates its use in collapsing components.

"Building Ensembles of Classifiers for Loss Minimization" by Dragos Margineantu at Oregon State University

This paper studies approaches to modifying existing methods for constructing ensembles to incorporate arbitrary loss functions. New algorithms are compared with ensemble learning methods like boosting and bagging, in which only the weak learner has been modified to incorporate the loss function, and with single hypotheses built by the weak learner.

"A Pseudo-Nearest~Neighbor Combined Classifier" by Majid Mojirsheibani at Carleton University

The paper considers a pseudo nearest-neighbor procedure for combining different classifiers in order to construct more effective classification rules. Effectiveness of the proposed procedure is assessed via simulation and theoretical studies.

"Ideal Bootstrap Estimation of Expected Error Rate for k-Nearest Neighbor Classifiers" by David Patterson and Brian Steele at University of Montana

Analytic formulae are presented for the ideal leave-one-out bootstrap estimate (Efron and Tibshirani 1997) of expected error rate for k-nearest neighbor (k-NN) classifiers, and a new weighted k-NN classifier is proposed based on these calculations. A simulation study shows that the resampling-weighted classifier compares favorably to unweighted and distance-weighted k-NN classifiers.

"The Number of Clusters in Binary Data Sets" by
Andreas Weingessel, Evgenia Dimitriadou and Sara Dolnicar at Vienna University of Technology.

This paper compares and analyzes the performance of several index measures on binary data sets. To evaluate performance, artificial binary data sets with known properties that are generated to resemble real data sets are used, as well as some empirical data sets

"A Part-Time Parallel Processing Environment for Statistical Computation" by David G. Whiting and Quinn Snell at Brigham Young University

Since current research has shown that disparate personal computers can be connected to create a parallel computer with capabilities rivaling those of a supercomputer, a model is demonstrated having idle computers join in an ad-hoc cluster with the user's machine for part-time parallel processing. This is illustrated using the statistical package R and showing that components of such a package can be written with this parallelization in mind.

4.4 Bootstrap

"Bootstrap Tilting" by Tim Hesterberg at MathSoft

Bootstrap-tilting confidence intervals and hypothesis tests offer both statistical and computational advantages, with a variation related to empirical likelihood methods being more accurate in finite-sample applications. Combining ideas from bootstrap-tilting diagnostics with bootstrap-t confidence intervals yields a promising new interval called the 'bootstrap-tilting-t.'

"Bootstrapping the Two-Stage Shrinkage Estitnators" by Makarand V. Ratnaparkhi at Wright State University and Vasant B. Waikar and Frederick J. Schuurman at Miami University

The paper discusses problems of bootstrapping two-stage estimators for the mean of a normal distribution. New estimators are constructed based on bootstrapping the shrinkage factor and simulation studies are performed to show how bootstrapping increases efficiency of estimators.

"Effects of Data Splitting on Prediction Error and its Estimation" by Craig A. Rolling at Infoworks

Three common splitting strategies are discussed and evaluated on both simulated and real examples: don't split the data at all and use a bias corrected resubstitution estimator of predictive performance, split the data into a modeling set to estimate the model and a validation set to estimate the model's predictive performance, and use 10-fold cross validation. Although cross validation is usually seen as a way to estimate unconditional prediction accuracy averaged over many training data sets, it can also be interpreted as the conditional accuracy of a certain "averaged" model, putting it on the same footing as the other strategies.

"Resampling Methods for Spatial Prediction" by
Soumendra N. Lahiri at Iowa State University

Suppose that Z is a spatially distributed random process and it is observed at a finite number of locations yielding the observations Z(1),..., Z(n). This paper formulates a spatial subsampling method and a spatial bootstrap method for making inference in such prediction problems, and shows that the proposed resampling methods yield consistent estimators of the mean square prediction error and other population characteristics of a large class of predictors, without any specific model assumptions on the spatial process.

"Balanced Confidence Regions Based on Tukey's Depth and the Bootstrap" by Arthur Yeh at Bowling Green State University with Kesar Singh

Based on the normalized or Studentized statistic formed for an i.i.d. random sample obtained from some unknown distribution, bootstrap points are deleted based on Tukey's depth until the desired confidence level is reached and the proposed confidence regions are shown to be second-order balanced in the context of Beran (1988). Applicability of the proposed method is demonstrated in a simulation study.

"Exact Bootstrap Moments of an L-estimator" by Michael Ernst and Man D. Hutson at University of Florida

It is shown that exact analytic expressions of the bootstrap mean and variance can be obtained for the broad class of statistics which are linear combinations of order statistics (L-estimators), eliminating the error due to bootstrap resampling. The non-negligible error of this resampling approach for estimating the bootstrap variance is examined using some classical L-estimators, such as the trimmed mean and the median, on real data.

4.5 Visualization

"Visualizing Attribute Relationships with DataSpheres" by Tamraparni Dasu and Theodore Johnson at AT&T Labs – Research

The paper attempts to capture how attribute distributions vary with each other by using DataSphere: marginal distributions of attributes are plotted within each class of the partition, and partition boundaries place distance and directional constraints on the attribute values that are used to visualize the linkage between variables. Both real and artificial data illustrate this technique.

"Visualizing Keyword Lists and Other High-Dimensional Binary Response Variables" by Mark A. Johnson and Martin Schulz at Pharmacia & Upjohn and Cheng Cheng at Johns Hopkins University

Lexicographically ordering the keyword list representations produces a histogram that provides a "forest" view of the structure in the associated high dimensional point cloud and exposes the dominant high-dimensional interactions, while a more detailed insight into this structure is obtained by coloring the bars according to the number of keywords in each list and subsetting the data by important keywords. These ideas are illustrated by examining the pharmacological profiles of 1800 compounds taken from an early version of the Derwent Standard Drug File, with the visual operations being executed using a dynamic data-visualization package called Spotfire.

"ORCA: A Project in Platform-Independent, Distributed Data Visualization" by Thomas Lumley, Peter Sutherland and Tony Rossini at University of Washington and Dianne Cook at Iowa State University.

ORCA is a project in platform independent statistical visualization tools for data analysis and research with objects that are designed to function as stand alone aids, as display devices for popular statistical systems, and as components in a large distributed statistical computing environment. The software uses Java for platform independence and to simplify 3D graphics programming, as well as to provide facilities for sockets and CORBA that will be used for distributed communication.

"A Template for Interactive Conditioned Choropleth Maps" by Daniel B. Carr and John F. Wallin at George Mason University.

The paper describes a template for conditioned choropleth (CC) maps, while extending the static coplots of Cleveland, Groose and Shyu (1992) to dynamic classed choropleth maps and emphasizing a 3 x 3 layout. CC maps provide a simple way of thinking about three geo-referenced variables, a dependent variable and two explanatory variables: interactivity facilitates exploration, sliders control selection of class boundaries and conditioning for each row and column, regions satisfying constraints appear in foreground colors, a small tab at the lower right of each panel provides a statistical summary, cognostics algorithms help select interesting views, and a fixed aspect ratio (FAR) pan and zoom widget simultaneously controls all panels.

"New Methods and Tools for Visualizing Graphs" by Ted Mihalisin at Temple University and John Timlin at Mihalisin Associates.

A new method for visualizing graphs is presented that facilitates understanding the graphs' clique structure and edges between the cliques. Possible methods to visually discover the clique structure of larger graphs are discussed, and new tools for visually discovering the clique structure of a graph visualized via the well-known matrix representation are also presented.

"Deep Regression" by Peter J. Rousseeuw at University of Antwerp with Stefan Van Aelst.

This paper introduces the notion of depth in a regression setting with a nonparametric error term, and provides the 'rank' of any line rather than the ranks of observations or residuals. Using computational geometry, algorithms are devekoped to compute regression depth and the deepest regression.

"Robust Analysis of Large Data Sets" by Peter J. Rousseeuw at University of Antwerp with Katrien Van Driessen

A new algorithm called FAST-LTS is developed involving order statistics and sums of squared residuals, and techniques that we call 'selective iteration' and 'nested extensions'. FAST-LTS typically finds the exact LTS for small data sets, and gives more accurate results for large data sets than existing algorithms for LTS and is faster by orders of magnitude

"Visualization Via Figural Animation" *(Animated and Colored Chernoff's Faces plus Sound)* by

Kurt Pflughoeft, Ehsan S. Soofi and Mariam Zahedi at University of Wisconsin – Milwaukee

A visualization technique called Figural Animation (FA) combines size, shape, orientation, color and sound to display a multidimensional data point or summaries of the multivariate data distribution. The static figural represents a single multidimensional point with each feature summarizing one to four dimensions, and the dynamic figural traverses through data points to animate their variation.

4.6 Markov Chain Monte Carlo (MCMC)

"Fully Model Based Approaches for Spatially Misaligned Data" by Bradley Carlin at University of Minnesota

(No abstract is available.)

"Weighted Monte Carlo Estimates For Computing Posterior Quantities" by Ming-Hui Chen at Worcester Polytechnic Institute with Qi-Man Shao

The weighted Monte Carlo estimate is a natural generalization of the sample mean estimate. After a brief overview of weighted estimates is provided, general strategies to construct good weights are discussed.

"Simulation Based Optimal Design and some Applications in Clinical Trials" by Peter Mueller at Duke University with Don Berry

The paper proposes forward simulation to approximate the integral expressions, and a reduction of the allowable action space to avoid problems related to an exponentially exploding number of possible trajectories in the backward induction. A proposed approach is illustrated by the application to an optirnal stopping problem in a clinical trial.

"Analysis of Cross-Section and Clustered Data Treatment Models" by Siddhartha Chib Washington University in St. Louis with Barton Hamilton

This paper is concerned with determining the effect of a categorical treatment variable on the response given that a treatment is non-randomly assigned and the response is observed for one setting of the treatment. It considers classes of models that are subjected to a fully Bayesian analysis based on Markov Chain Monte Carlo methods, with the analysis of treatment effect being based on the posterior distribution of potential outcomes at the subject level that is obtained as a by-product of the MCMC simulation procedure.

"Statistical Supercomputing for Gibbs Sampling of Massive Spatio-Temporal Models" by Chris Wikle at University of Missouri – Columbia with Tim Hoar, Ralph Milliff and Doug Nychka at NCAR

Conditionally-specified spatio-temporal models are advantageous in that physical and dynamical constraints are easily incorporated into a conditional formulation: the series of relatively-simple-yet-physically-realistic conditional models leads to a much more complicated joint space-time covariance structure than can be specified directly, but the models are computable with a Gibbs Sampler for reasonably large data sets on a standard workstation Focus is on a problem concerning spatio-temporal prediction of near surface wind fields over the tropics, given "data" from weather center deterministic models and remotely sensed satellite observations on the 24 processor, 1.5 GFLOPS, 1024 Mword Cray J924se super-computer at NCAR.

"Reversible Jump MCMC Analysis of Spatial Poisson Cluster Processes with Bivariate Normal Displacement" by John Castelloe at SAS Institute and University of Iowa

The reversible jump Markov Chain Monte Carlo (RJMCMC) technique for bivariate normal mixtures with common covariance matrix is developed and applied to the spatial Poisson cluster process with bivariate normal offspring dispersal. A new convergence assessment method, applicable to any RJMCMC situation where distinct models can be identified, is designed and theoretically justified.

"Combining Expert Beliefs, Nongaussian Continuous Random Variables and Decision Option Variables with a Simulation-Based Influence Diagram" by Timothy C. Haas at University of Wisconsin - Milwaukee

The theory of influence diagrams (ID's) is extended to allow inclusion of systems of stochastic differential equations and/or birth-death processes. Since the joint probability density-probability function (PDPF of Koopmans 1969) often has no analytical form, simulation is used to approximate the PDPF for purposes of finding constrained minimum distance estimates of the distribution's parameters.

"The Art of Data Augmentation: From EM to MCMC" by David van Dyk at Harvard University and Xiao-Li Meng at University of Chicago

An effective strategy is introduced that combines the ideas of marginal Data Augmentation (DA) and conditional DA with regard to a working parameter, together with a deterministic approximation method for selecting prior distributions for the working parameter that result in good DA schemes. This strategy is then applied to mixed-effects models to obtain efficient Markov Chain Monte Carlo algorithms for posterior sampling.

"Transformation Group and MCMC" by Jun S. Liu at Stanford University and Ying Nian Wu at University of Michigan

A generalization of the Gibbs sampler is proposed, where conditional moves along the coordinate axes are generalized to conditional moves along the orbits of transformation groups. This scheme can lead to more efficient MCMC algorithms with more flexible conditional moves and two methods for accelerating MCMC using this scheme are studied.

"The Use of Multiple-Try Method and Loeal Optimization in Metropolis Sampling" by Jun S. Liu at Stanford University, Faming Liang at National University of Singapore and Wing Hung Wong at UCLA

The paper proposes two novel approaches for general Metropolis-type sampling: the Multiple-Try Metropolis (MTM) uses a nontrivial twist of the original Metropolis-Hastings algorithms to accommodate moves with very large stepsize, and a second one finds that the Adaptive Direction Sampling framework (Guks, Roberts, and George 1994) can be extended to rigorously incorporate local deterministic moves into a Markov Chain Monte Carlo sampler for simulating random variables in continuous state-space. Numerical studies show that the new method performs significantly better than the traditional Metropolis-Hastings sampler, and the MTM with minor tailoring can be used to explore a fixed or random direction along the sample space without having to know how to draw from the required conditional distribution.

4.7 Image Analysis

"Feature Detection in Spatial Point Processes with Application to Minefield Detection" by Adrian Raftery at University of Washington

This paper considers the "reconnaissance" minefield detection problem in which a large area is surveyed and imaged and potential mines identified: the model is a finite mixture of bivariate normal distributions with possible cross-cluster constraints on the covariance matrices, and a Poisson process component is added to the mixture model to represent noise. Simulation results based on specifications from the Naval Coastal Systems Station are good and the MCLUST software has been widely used in many applications.

"Probabilistic Methods in Image Analysis with Applications in Automatic Target Recognition" by Alan Willsky at MIT

The paper describes two different approaches to nonlinear image analysis and segmentation: one can be thought of as a limiting form of so-called anisotropic diffiisions that result in a coupled set of differential equations with discontinuous right-hand sides and a demonstrated algorithm robustness to severe image degradations such as speckle, and the other represents a new curve evolution algorithm with the general idea of devising a dynamical system for taking an initial curve and then evolving it into yielding meaningful segmentations. It begins from a global statistical formulation that looks for curves to maximize the distance between a statistic calculated over pixels inside the curve and a statistic calculated over pixels outside the curve, and it results in a flow that is statistically based and makes explicit use of smoothed global statistics while yielding remarkably robust segmentations of noisy images.

"Statistical Learning and Coarse-to-Fine Object Detection" by Donald Geman at University of Massachusetts

This paper is about research in computational vision that is motivated by parlor games such as "Twenty Questions" and involves a mixture of methods from statistics, information theory and image analysis where the variables are selected based on purely statistical criteria, without a priori semantic or geometric interpretation. In addition, image regions are explored according to their information content and the spatial distribution of computation is very skewed.

"Importance Sampling for Spatial Scan Analysis: Computing Scan Statistic p-Values for Marked Point Processes" by Carey E. Priebe and Daniel Q. Naiman at The Johns Hopkins University

An importance sampling method based simultaneously on multiple scan geometries is presented for deciding if a scan statistic provides significant evidence of increased activity in some localized region of time or space given an observed marked point pattern. The paper develops an importance sampling p-value estimator for marked spatial Poisson processes, and illustrates the approach via application to minefield detection and to the detection of microcalcification clusters in digital mammography.

"A Divide and Conquer Approach to Fisher's Exact Test" by Edmund Staples at University of Massachusetts – Dartmouth

A form f(T) of an r x c contingency table T is the list of subtotals for rows and columns and the approach is to efficiently enumerate splits of forms f into pairs (f1, f2) having fixed shapes r x c1 and r x c2 (with c1+c2 = c) via linear programming. Since the mathematical partition function grows more slowly than functions of binomial coefficients, the space and time requirements of the histograms are less.

"A Statistical Model for Computer Recognition of Handwritten ZIP Codes" by Steve C. Wang at Williams College

Bayesian principles are used to build a statistical computer model for recognizing handwritten ZIP codes by producing a posterior distribution that optimizes both segmentation and recognition, and then sums it to obtain a ZIP code estimate that is the marginal mode for recognition alone. To make this optimization feasible, a generalized dynamic programming algorithm is implemented.

"A Model-Based Approach to Handling Missing Values in S-PLUS using EM, Iterative Simulation, and Multiple Imputation" by Tim Hesterherg and James Schimert at MathSoft

The paper discusses S-PLUS software that supports the model based missing data methods of EM (Little and Rubin 1987) and data augmentation (Schafer 1997). It may be applied to handle a wide variety of missing data problems.

"A Note on the Minimization of the Trimmed Sum of Absolute Deviations as an Integer Program" by Andrew A. Neath and Edward C. Sewell at Southern Illinois University – Edwardsville

The trimmed sum of absolute deviations criterion (Tableman, 1994) is useful for diminishing the effects of large random error terms on estimated regression coefficients, and for aiding the detection of outliers by writing the minimization problem as an integer program. Thus, the requirements to perform a least trimmed absolute deviations regression analysis are met for any practitioner with access to a computer package capable of solving integer programming problems.

4.8 Mixed-Effects Models

"Nonlinear Mixed Models: A Future Direction" by Russell D. Wolfinger at SAS Institute

Statistical models in which both fixed and random effects parameters enter nonlinearly are becoming increasingly popular, and have a wide variety of applications such as nonlinear growth curves and overdispersed binomial data. A new SAS procedure NLMIX£D fits these models using likelihood-based methods.

"Maximum Likelihood for Generalized Linear Mixed Models via High Order Multivariate LaPlace Approximation" by Steve Raudenbush at University of Michigan and Meng Li Yang at Nan Tai Institute of Technology in Taiwan

Since both evaluation of the required integral and likelihood inference are difficult, a multivariate Taylor series approximation is proposed for the integrand's log and a high order Laplace approximation is then applied to the integral in order to maximize the approximately integrated likelihood via Fisher scoring. The approach is applied to generalized linear models with random coefficients and illustrated in the special case of binary outcomes.

"Computing REML or ML Estimates for Linear or Nonlinear Mixed-Effects Models" by Douglas Bates at University of Wisconsin - Madison

This paper describes three methods for streamlining the optimization of REML and/or ML parameter estimation criteria for linear mixed-effects models: it conditions on the ratio of the variance components or the relative precision factors, it uses matrix-logarithm parameterization for the relative precision factors to provide a more stable optimization problem, and it derives both the EM and Newton-Raphson updates so optimization can begin with EM iterations and switch to Newton-Raphson updates when near the optimum. All of these can also be used in the optimization of estimation criteria in nonlinear mixed-effects models.

4.9 Data Mining

"A Modified Simulated Annealing Algorithm and its Application to the Satisfiability Problem" by G. Sampath at University of Massachusetts - Dartmouth

A modified version of the simulated annealing algorithrn is discussed in which the neighborhood of the current state is sampled and the sample distribution so obtained is used instead of a Boltzman distribution. Computational experiments are discussed.

"Minimizing the Costs of Screening Designs" by Queritin F. Stout at University of Michigan and Janis P. Hardwick at Purdue University

Assuming that tests have Bernoulli outcomes, a program for optimizing screening designs is provided that has an integrated framework for accurately modeling a situation and giving the user a great deal of flexibility, while incorporating test costs and constraints with the costs of incorrect decisions. It allows for grouped allocation of tests, ranging from a single round of tests to fully sequential designs.

"Visual Data Mining of Brain Cells" by Giorgio Ascoli, Jeffrey L. Krichmar and Stuart D. Washington at George Mason University

Since little is known about the influence of morphological variablity on the physiological response of brain cells and it is difficult to separate the physiological differences in the cell from the morphological differences in laboratory experiments,

computer simulations of 3-D neuroanatomical data on hippocampal while keeping the physiology across different cells constant cells have been used. Data from neuroanatomical archives on hippocampal pyramidal cells has been obtained and a visual data mining approach with packages such as XGobi has been found to be a very effective way of detecting some structure in the data.

"Data Mining" by Philip Pretorius at Potchefstroomse Universitiet vir CHO

A logical strategy based on statistical methods is proposed for the analysis of large data sets. The focus is on an example from the auditing environment.

"Run Rules" by Philip Pretorius at Potchefstroomse Universitiet vir CHO

Run Rules are used in statistical process control to indicate persistent shifts away from targets. This paper is concerned with adapting Run Rules for the data-rich computerized environment of today.

"Tukey-Kramer Multiple Comparison Pruning of Neural Networks" by Donald E. Duckro, Dennis W. Quinn, and Samuel J. Gardner at AFIT

The method proposed uses statistical multiple comparison procedures to remove multiple parameters in the model when no significant difference exists. While computationally intensive, this method compares well with Optimal Brain Surgery in pruning and network performance.

"Quantifying Operational Risks through MC Simulations and Bayesian Networks" by Prabhat Ojha, Christopher Lee Marshall, and Vishrut Jain at National University of Singapore

Operational managers needing to act on events emphasize the importance of causation over correlation as the representation of interdependence between risks by incorporating events into a causal structure using a Bayesian Network. After obtaining a preset level of consistency, Monte Carlo simulation is performed over the Bayesian Network to generate a loss distribution from which are derived the expected, unexpected and catastrophic loss components of Operational Risk.

"Clustering with Finite Data from Semi-Parametric Mixture Distributions" by Yong Wang and Ian Witten at University of Waikato in New Zealand

The paper proposes a new distance-based clustering method that overcomes the problem by avoiding global constraints. Experimental results confirm its superiority over standard methods when small clusters are present in finite data sets, and they also suggest that it is more accurate and stable than other methods even when there are no small clusters.

"Mathematical Models of Epidemics for Public Health Applications" by John J. Hsieh at University of Toronto

This paper develops both deterministic and stochastic models for microparasitic infection to study: (1) the static behaviour of epidemics by deriving the basic reproductive rate, incidence rate, transmission rate, recovery rate and death rate in terms of the mean age at infection, mean age at death and mean duration of the infectious period under two assumed distributions of lifelength in the stationary population; (2) the dynamic behaviour of epidemics by using deterministic models to analyze periodicities in the incidence of endemic infection; and (3) the effects of immunization on epidemiological and demographic parameters by using stochastic models to analyze the effects on persistence of the epidemic and on eradication of endemic infection. The degree of realism in these models is demonstrated by adequately fitting the models to the incidence curve, prevalence curve and curve of proportion seropositive arising from the epidemiological data.

"Robust Estimation of the SUR Model" by Martin Bilodeau at Universite de Montreal

The paper proposes robust regression to solve a problem of outliers in seemingly unrelated regression (SUR) models by adapting S-estimators to SUR models. A classical example on U.S. corporations is revisited, and it appears that the procedure gives an interesting insight into the problem.

"Genome Data: Finding Trait Genes Using a Dense Marker Map" by Katy Simonsen at Purdue University

Markers and disease genes are simulated with historical correlation built into the data according to evolutionary models, and the data are analyzed by a Monte Carlo procedure to assess significance and then contrasted with the results of a Bonferroni correction. As expected, the Monte Carlo procedure is found to be more powerful when markers are correlated, and the change in power can be related to various evolutionary parameters.

"Models of Protein Evolution Incorporating Correlation" by Kay S. Tatsuoka at NISS, Francoise Seillier-Moiseiwitsch at University of North Carolina and Alex Stark at NISS

A computationally tractable model is introduced that allows for correlated mutations in two or more positions, for the substitution matrix of a model to be estimated from primary sequences, and for MCMC to be used for exploring tree space and computing the posterior probabilities of trees. A test statistic is proposed for testing the independence of positions and taking into account the phylogeny, and the methods are illustrated on a large set of amino acid sequences from HIV protease.

"Group Testing with Blockers and Synergism" by Minge Xie at Rutgers University

This paper develops models and estimation procedures to obtain quantitative information from data in group testing practice, by studying several group testing procedures under violations of the standard assumption that tested items act independently of one another (Dorfman 1943): (1) when there are blockers, objects placed in a pool with a positive cause the pool to test negative; and (2) when there are combination effects, synergism or additive effects cause a pool containing no single active compounds to test active. The investigation is focused on, but not limited to, the square array pooling method (Phatarfod and Sudbury 1994).

4.10 Computational Biology

"Genomic Level Inferences on Protein Function: A Bayesian Approach" by Charles E. Lawrence at VA - Wadsworth Center

Complete genome sequences offer new opportunities to predict the functions of proteins and to assess their significance. A summary is presented of the PROBE Bayesian multiple sequence alignment and databases search algorithm (Neuwald et. al.1997

and Liu et. al.1999) and the Bayesian classification algorithm (Qu et. al 1998).

"Computing With Trees" by Susan P. Holmes at Stanford University with Persi Diaconis

The paper presents a natural coordinate system for trees using a correspondence, that produces a distance between phylogenetic trees and a way of enumerating all trees in a minimal step order, with the set of perfect matchings in the complete graph. Coding of trees by matrices instead of pointers simplifies use of higher level languages such as matlab instead of C, thus enabling researchers to use methods without considering the programs as black boxes.

"Gene Mapping and Fourier Transforms on Groups" by Augustine Kong at University of Chicago

This paper describes a recent advance in linkage computations using Fourier transforms of measures on discrete groups (Kruglyak and Lander 1998). That is an elegant example of how mathematics and applications work together.

5. ACKNOWLEDGMENTS

The author gratefully acknowledges contributions of both Interface '99 Speakers and Interface '99 Co Chairs.

6. REFERENCES

[1] Weisberg, S. ed. Proceedings of Inerface '98: 30th Symposium on the Interface of Computing Science and Statistics. Interface Foundation of North America, 1999.

[2] Goodman, A. and Elder, J. IV. Knowledge Discovery and the Interface of Computing and Statistics. Proceedings: The Fourth International Conference on Knowledge Discovery and Data Mining, 1998.

[3] *G*oodman, A.. Report from Interface '98: Knowledge Discovery and the Interface of Computing and Statistics. SIGKDD Explorations 1, 1 (June 1999).

[4] Elder, J. IV. The Interface '98 Conference – a Resource for KDD. SIGKDD Explorations 1, 1 (June 1999).

[5] Berk, K. and Pourahmadi, M., ed. Proceedings of Interface '99: 31st Symposium on the Interface of Computing Science and Statistics. Interface Foundation of North America, 2000.

[6] Goodman, A. and McCornack, R.. Recommended Procedure for Constructing a SPEC Program Evaluation Sheet. Statistical Program Evaluation Committee, 1962. *Available on Request.*

[7] Eisner, W. and Goodman, A. Determination of Dominant Error Sources in an Inertial Navigation System by Iterative Weighted Least Squares. American Institute of Aeronautics and Astronautics Journal 2, 4 (April 1964).

[8] Chernoff, H.. The Use of Faces to Represent Points in N-Dimensional Space Graphically. Stanford University, Department of Statistics Technical Report 71 (December 1971).

About the authors:

Arnold Goodman: Before co founding the UCI Center for Statistical Consulting, Arnold Goodman used his Ph.D. in statistics from Stanford University to solve problems in information technology for Rockwell, McDonnell Douglas, Atlantic Richfield and County of Los Angeles. He founded the Interface of Computer Science and Statistics, and is a Fellow of the American Statistical Association.