FOREWORD

The main focus of Statistical Computation and Visualization 2008 (SCV2008) is on recent development of concept, technique, and software for statistical computation and data visualization with their applications to scientific studies and industrial projects. We are very fortunate to be able to invite more than 30 worldwide renowned researchers in related fields from more than 10 countries to present their most recent works.

In addition to the *Statistica Sinica* Discussion Paper Session, 28 presentations are grouped into 10 sessions with diversified topics in statistical computation and visualization in this 3-day workshop. I wish to express my gratitude to all participants of SCV2008 and in particular to our administrative colleagues of the Institute for a wonderful job done within a very tight schedule. Special thanks go to Ms. Zi-Xiu Lai for coordinating SCV2008 administrative work and to Mr. Chiun-How Kao for designing/maintaining SCV2008 webpage/database with the creative poster and cover page designs.

I hope you enjoy the programs of the workshop, the environment in Academia Sinica, and the culture and food in Taipei City.

Chun-houh Chen  
Institute of Statistical Science  
Academia Sinica, Taiwan  
December 1, 2008
ABACUS: Computation

HISTOGRAM: Visualization

Paired moon-shape wood DIVINATION BLOCKS: Chinese coin tossing
(flat: Ying, convex: Yang) →
(Ying, Yang): granted
(Yang, Yang): no good
(Ying, Ying): unclear

Pot of numbered bamboo DIVINATION SLIPS: Sampling

You shake the pot till one slip falls out and you toss the pair of blocks to validate that slip. Put the slip back (with replacement) and repeat the sampling process (I suppose bootstrapping is not an option) till you get a slip validated with 3 granted pairs (Ying, Yang) in a row (probability < 1/8 I guess since I don’t think they are fair coins). Obtain and interpret the oracle verse matched with that granted slip. You can practice these block tossing and slip sampling at the 3 most popular temples in Taipei City:
Lung-Shan (Dragon Mountain) Temple (built 1738) (http://www.lungshan.org.tw/)
Dalongdong Baoan Temple (1742) (http://www.paoan.org.tw/ASP/Home/default.asp)
Qingshui Temple, Mengjia (1787) (where the DIVINATION SLIPS pot photo was taken)
Or there is the Fu-Nan Temple right on Academia Road, Section 1.
<table>
<thead>
<tr>
<th>Time</th>
<th>Dec. 1 (Mon.)</th>
<th>Dec. 2 (Tue.)</th>
<th>Dec. 3 (Wed.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td>Registration</td>
<td>Session 3: Analytic and Symbolic Systems</td>
<td>Session 7: Visualization and Application</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chair: Chen-Hai Tsao</td>
<td>Chair: Eunsik Park</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 1: Edwin Diday</td>
<td>Speaker 1: Jürgen Symanzik</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 2: Leland Wilkinson</td>
<td>Speaker 2: Catherine Hurley</td>
</tr>
<tr>
<td>10:00</td>
<td>Opening ceremony</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistica Sinica Discussion Paper Session</td>
<td>Coffee break</td>
<td>Coffee break</td>
</tr>
<tr>
<td></td>
<td>Session Chair: Donald Ylvisaker</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paper Presentation: Stephan Huckemann</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discusant 1: Rabindra Bhattacharya</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discusant 2: John T. Kent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00</td>
<td>Lunch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>Session 1: Statistical Testing</td>
<td>Session 5: Graphic Concept</td>
<td>Session 9: Interactive Graphics and Animation</td>
</tr>
<tr>
<td></td>
<td>Chair: Chen-An Tsai</td>
<td>Chair: Yasumasa Baba</td>
<td>Chair: Han-Ming Wu</td>
</tr>
<tr>
<td></td>
<td>Speaker 1: Jaromí Antoch</td>
<td>Speaker 1: Jae-Chang Lee</td>
<td>Speaker 1: Dae Heung Jang</td>
</tr>
<tr>
<td></td>
<td>Speaker 2: Seung-Ho Kang</td>
<td>Speaker 2: Kwan-Liu Ma</td>
<td>Speaker 2: Antony Unwin</td>
</tr>
<tr>
<td></td>
<td>Speaker 3: Yi Tsong</td>
<td>Speaker 3: Shizuhiko Nishisato</td>
<td>Speaker 3: Simon Urbanek</td>
</tr>
<tr>
<td>15:45</td>
<td>Coffee break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>Session 2: Functional Multivariate Method</td>
<td>Session 6: Genomic Statistics</td>
<td>Session 10: Data/Dimension Reduction with Visualization</td>
</tr>
<tr>
<td></td>
<td>Chair: Naisyin Wang</td>
<td>Chair: Ueng-Cheng Yang</td>
<td>Chair: Yuh-Jye Lee</td>
</tr>
<tr>
<td></td>
<td>Speaker 1: Yutaka Tanaka</td>
<td>Speaker 1: James Chen</td>
<td>Speaker 1: Daniel B. Carr</td>
</tr>
<tr>
<td></td>
<td>Speaker 2: Jane-Ling Wang</td>
<td>Speaker 2: Seiya Imoto</td>
<td>Speaker 2: Yueh-Yun Chi</td>
</tr>
<tr>
<td></td>
<td>Speaker 3: Thomas Hotz</td>
<td>Speaker 3: W. Y. Wendy Lou</td>
<td>Speaker 3: Masahiro Mizuta</td>
</tr>
<tr>
<td>17:45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td>Welcome Party</td>
<td>Banquet</td>
<td></td>
</tr>
<tr>
<td>21:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:00</td>
<td>Registration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:00</td>
<td>Opening ceremony</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td><strong>Statistica Sinica Discussion Paper Session</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Session Chair: Donald Ylvisaker</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paper Presentation: Stephan Huckemann</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intrinsic Shape Analysis: Geodesic PCA for Riemannian Manifolds Modulo Isometric Lie Group Actions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussant 1: Rabindra Bhattacharya</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussant 2: John T. Kent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00</td>
<td><strong>Lunch</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td><strong>Session 1: Statistical Testing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chair: Chen-An Tsai</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Speaker 1: Jaromír Antoch</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nonparametric comparison of ROC curves. Testing equivalence and clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Speaker 2: Seung-Ho Kang</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exact Tests for Testing the Homogeneity of Two Binomial Proportions in Extremely Unbalanced 2 x 2 Contingency Tables</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Speaker 3: Yi Tsong</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiple comoparison issues in thorough QTc Clinical Trials</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:45</td>
<td><strong>Coffee break</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td><strong>Session 2: Functional Multivariate Method</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chair: Naisyin Wang</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Speaker 1: Yutaka Tanaka</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Influence analysis in ordinary, kernel and functional multivariate methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Speaker 2: Jane-Ling Wang</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Covariate Adjusted Functional Principle Component Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Speaker 3: Thomas Hotz</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shape analysis for pre-aligned, star-shaped objects - what PCA can tell about the evolution of tree rings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:45</td>
<td><strong>Welcome Party</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td><strong>Welcome Party</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21:00</td>
<td><strong>Welcome Party</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Dec. 2 (Tue.)

<table>
<thead>
<tr>
<th>Time</th>
<th>Session 3: Analytic and Symbolic Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td>Chair: Chen-Hai Tsao</td>
</tr>
<tr>
<td></td>
<td>Speaker 1: Edwin Diday</td>
</tr>
<tr>
<td></td>
<td>SYMBOLIC DATA ANALYSIS: a new model for DATA MINING and its SYR software</td>
</tr>
<tr>
<td></td>
<td>Speaker 2: Leland Wilkinson</td>
</tr>
<tr>
<td></td>
<td>Identifying and Querying Local and Large-Scale Structures Using $L_\infty$ Norms</td>
</tr>
<tr>
<td>10:10</td>
<td>Coffee break</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Session 4: Modeling and Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30</td>
<td>Chair: Chul Eung Kim</td>
</tr>
<tr>
<td></td>
<td>Speaker 1: John Aston</td>
</tr>
<tr>
<td></td>
<td>Computing and Visualising Regime Variability in Hidden Markov Models</td>
</tr>
<tr>
<td></td>
<td>Speaker 2: Wolfgang Härdle</td>
</tr>
<tr>
<td></td>
<td>Modelling and Forecasting Liquidity Supply Using Semiparametric Factor Dynamics</td>
</tr>
<tr>
<td></td>
<td>Speaker 3: Henry Horng-Shing Lu</td>
</tr>
<tr>
<td></td>
<td>Multidimensional scaling for large genomic data sets</td>
</tr>
<tr>
<td>12:15</td>
<td>Lunch</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Session 5: Graphic Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00</td>
<td>Chair: Yasumasa Baba</td>
</tr>
<tr>
<td></td>
<td>Speaker 1: Jae-Chang Lee</td>
</tr>
<tr>
<td></td>
<td>Some Thoughts on Graphical Methods in Statistics</td>
</tr>
<tr>
<td></td>
<td>Speaker 2: Kwan-Liu Ma</td>
</tr>
<tr>
<td></td>
<td>Visualization in the Right Space</td>
</tr>
<tr>
<td></td>
<td>Speaker 3: Shizuhiko Nishisato</td>
</tr>
<tr>
<td></td>
<td>Radical suggestions on improved graphical display of multidimensional categorical data</td>
</tr>
<tr>
<td>15:45</td>
<td>Coffee break</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Session 6: Genomic Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:00</td>
<td>Chair: Ueng-Cheng Yang</td>
</tr>
<tr>
<td></td>
<td>Speaker 1: James Chen</td>
</tr>
<tr>
<td></td>
<td>Gene Set Analysis in Expression Profiling Studies</td>
</tr>
<tr>
<td></td>
<td>Speaker 2: Seiya Imoto</td>
</tr>
<tr>
<td></td>
<td>Unraveling Dynamics of Biological Networks towards Computational Drug Target Discovery</td>
</tr>
<tr>
<td></td>
<td>Speaker 3: W. Y. Wendy Lou</td>
</tr>
<tr>
<td></td>
<td>Computational Methods for Cluster Detection of Patterns in Biology Sequences</td>
</tr>
<tr>
<td>17:45</td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td>Banquet</td>
</tr>
<tr>
<td>21:00</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Session Title</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------------------------------------</td>
</tr>
<tr>
<td>9:00</td>
<td>Session 7: Visualization and Application</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td>Coffee break</td>
</tr>
<tr>
<td>10:30</td>
<td>Session 8: Data Discretization, Ranking, and Rounding</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>12:15</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:00</td>
<td>Session 9: Interactive Graphics and Animation</td>
</tr>
<tr>
<td>15:45</td>
<td>Coffee break</td>
</tr>
<tr>
<td>16:00</td>
<td>Session 10: Data/Dimension Reduction with Visualization</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>17:45</td>
<td></td>
</tr>
</tbody>
</table>
Statistica Sinica
Discussion Paper Session
Statistica Sinica Discussion Paper Session

Session Chair: Donald Ylvisaker
Paper Presentation: Stephan Huckemann
   Intrinsic Shape Analysis: Geodesic PCA for Riemannian Manifolds Modulo Isometric Lie Group Actions

Discussant 1: Rabindra Bhattacharya

Discussant 2: John T. Kent
Dr. Don Ylvisaker, Professor,

UCLA Statistical Consulting Center
8118 Math Sciences Building, 155404
UCLA
405 Hilgard Ave.
Los Angeles,
USA
E-mail: ndy@stat.ucla.edu

Don Ylvisaker is an Emeritus Professor, Department of Statistics, UCLA.

Contact information: ndy@stat.ucla.edu

805-969-3092 (home phone) 818-585-5362 (cell phone)

Interests: Design of Experiments, Applied Statistics

Current activities: Consulting Editor for *Statistica Sinica*

Chair, Memorials Committee, IMS

Member, Panel on Coverage Evaluation in the 2010 Census

Consultant, California State Lottery

Consultant, Applied Survey Research

Honors: IMS Fellow, ASA Fellow, Distinguished Service Award, ICSA
Stephan Huckemann received his degree in mathematics from the University of Giessen, Germany in 1987. He was a visiting lecturer and scholar at the University of Michigan, Ann Arbor (1987 - 1989) and a postdoctoral research fellow at the ETH Zürich, Switzerland (1989 - 1990). He then worked as a commercial software developer and returned to academia as a contributor to the computer algebra system MuPAD at Sciface Software and the University of Paderborn, Germany (2001 - 2003). Currently he holds a DFG research position at the Institute for Mathematical Stochastics, University of Göttingen, Germany. His research interests include pattern recognition, statistical shape analysis, and fingerprint modelling.
Intrinsic Shape Analysis: Geodesic Principal Component Analysis for Riemannian Manifolds Modulo Lie Group Actions

Stephan Huckemann
Georgia Augusta Universität, Göttingen, Germany

Abstract
A general framework is laid out for principal component analysis (PCA) on quotient spaces that result from an isometric Lie group action on a complete Riemannian manifold. If the quotient is a manifold, geodesics on the quotient can be lifted to horizontal geodesics on the original manifold. Thus, PCA on a manifold quotient can be pulled back to the original manifold. In general, however, the quotient space may no longer carry a manifold structure. Still, horizontal geodesics can be well-defined in the general case. This allows for the concept of generalized geodesics and orthogonal projection on the quotient space as the key ingredients for PCA.

Generalizing a result of Bhattacharya and Patrangenaru (2003), geodesic scores can be defined outside a null set. Building on that, an algorithmic method to perform PCA on quotient spaces based on generalized geodesics is developed. As a typical example where non-manifold quotients appear, this framework is applied to Kendall’s shape spaces. In fact, this work has been motivated by an application occurring in forest biometry where the current method of Euclidean linear approximation is unsuitable for performing PCA. This is illustrated by a data example of individual tree stems whose Kendall shapes fall into regions of high curvature of shape space: PCs obtained by Euclidean approximation fail to reflect between-data distances and thus cannot correctly explain data variation. Similarly, for a classical archeological data set with a large spread in shape space, geodesic PCA allows new insights that have not been available under PCA by Euclidean approximation. We conclude by reporting challenges, outlooks, and possible perspectives of intrinsic shape analysis.

This is joint work with Thomas Hotz and Axel Munk also from the University of Göttingen.
Dr. Rabindra N Bhattacharya, Professor,
Department of Mathematics, The University of Arizona
617 N. Santa Rita Ave. P.O. Box 210089
Tucson, AZ 85721–0089, USA
E-mail: rabi@math.arizona.edu

1. PH.D: University of Chicago (1967), Statistics

2. APPOINTMENTS:
   Regular Appointments:
   - Assistant Professor of Statistics, UC, Berkeley (1967-72)
   - Associate, Full Professor of Mathematics, Univ. of Arizona (1972-1982)
   - Professor of Mathematics, Indiana University (1982-2002)
   - Professor of Mathematics, University of Arizona (2002--)
   Visiting Research Professorships:
   - Indian Statistical Institute (1978-79)
   - University of Goettingen, Germany (Summers of 1992-94)
   - University of Bielefeld, Germany (Summers of 1992-94, 98)
   - Oregon State University (2000-01)

3. HONORS & AWARDS:
   - IMS Fellow (1978);
   - DMV (German Math. Society) Lecturer (1989);
   - Alexander von Humboldt Forschungpreis (1994)- Humboldt Prize;
   - IMS Special Invited (Meadallion) Lecturer (1996);
   - Guggenheim Fellow (2000)

4. SERVICE TO PROFESSION:
   Service to U of A
   - Member, BIO5 Faculty
   - Member, Executive Committee of Statistics Graduate Interdisciplinary Program (GIDP)
   - Member, Statistics GIDP Search and Curriculum Committees
   - Member, Mathematics Personnel Committee
   - Undergraduate Advising for the Department of Mathematics

Some Other Service:
   - Member of IMS Council: 1998-2001
   - Panel Member of CBMS Regional Conference Board: 1990
Discussion: Intrinsic Shape Analysis: Geodesic PCA for Riemannian Manifolds Modulo Isometric Lie Group Actions (Authors: S. Huckemann, T. Hotz., and A. Munk)

Rabi Bhattacharya
Department of Mathematics
The University of Arizona.

Abstract

We discuss a number of potential applications, extensions, etc., of the methodology presented by Huckemann et al., such as feature selection, classification, and some other aspects of statistical inference.
John T Kent earned his Ph.D. degree in statistics from the University of Cambridge (UK) in 1977. He started his professional career as a research fellow at Sidney Sussex College, Cambridge in 1975. In 1977 he moved to Leeds University, UK as a lecturer, becoming a reader in 1983 and a professor in 1989. He is the joint author of a book Multivariate Analysis (1979) with KV Mardia and JM Bibby, and about 75 papers in various aspects of statistics including infinite divisibility, directional data analysis, multivariate analysis, inference, robustness, shape analysis, and image analysis.
Discussion: Intrinsic Shape Analysis: Geodesic PCA for Riemannian Manifolds Modulo Isometric Lie Group Actions

John T Kent
Department of Statistics, University of Leeds, Leeds LS2 9JT, UK

Abstract
Statistical analysis on shape spaces poses special challenges in two contrasting situations: (a) when the data are highly dispersed, and (b) when the data are nearly degenerate (e.g. nearly collinear 3-d shapes). We shall discuss various approaches to tackling these situations and draw out some connections to directional data analysis.
Session 1
Statistical Testing
Session 1: Statistical Testing
Chair: Chen-An Tsai
Speaker 1: Jaromír Antoch
    Nonparametric comparison of ROC curves.
    Testing equivalence and clustering
Speaker 2: Seung-Ho Kang
    Exact Tests for Testing the Homogeneity of Two Binomial Proportions in Extremely Unbalanced 2 x 2 Contingency Tables
Speaker 3: Yi Tsong
    Multiple comparison issues in thorough QTc Clinical Trials
Chen-An Tsai earned his Ph.D. degree in statistics from the National Central University, Taiwan in 1999. He started his professional career as a postdoctoral fellow at the National Center for Toxicological Research, FDA, USA. In 2005, Dr. Tsai was appointed to the position of Associate Research Scientist at the Institute of Statistical Science, Academia Sinica. He is now an Associate Professor at the China Medical University. His research interests include bioinformatics, biostatistics, statistical machine learning, and the development of statistical approaches for medical data. Significance analysis of gene sets in expression profiles is his current contribution to microarray experiment studies.
**Professor RNDr. Jaromir Antoch, CSc.**

Charles University of Prague, Faculty of Mathematics and Physics, Department of probability theory and mathematical statistics, Sokolovska 83, CZ-18675 Praha 8 – Karlin, Czech Republic, Europe

E-mail: jaromir.antoch@mff.cuni.cz

**Jaromir Antoch** earned his Ph.D. degree in statistics from the Charles University of Prague (Czech Republic) in 1982. He spent his entire professional career at the department of probability and mathematical statistics of the faculty of mathematics and physics of the Charles University. Currently he is a full time professor and a head of its statistical group. During the years he spent, among others, several years in France (Universities at Bordeaux, Toulouse and Grenoble) and Italy (Universities at Cagliari, Napoli and Rome).

His research interests include statistical computing, simulations, change point detection, robust and nonparametric statistics, industrial statistics and applications (especially in the industry).

During 1992-1996 he has been a chairman of the Board of Directors of the European Regional Section of the International Association for Statistical Computing (IASC), for the years 2007-2009 he is a President of IASC and a Council Member of the International Statistical Institute. During 2001-2007 he has been President of the Czech Statistical Society.
NONPARAMETRIC COMPARISON OF ROC CURVES
TESTING EQUIVALENCE AND CLUSTERING
Jaromír Antoch, Luboš Prchal, Pascal Sarda

Keywords: ROC curves, binary classification, collocations, clustering.

Receiver operating characteristic (ROC) curves are a popular and widely used tool that can help to summarize the overall performance of diagnostic methods and/or classifiers assigning individuals $g \in \mathcal{G} = \mathcal{G}_0 \cup \mathcal{G}_1$, $\mathcal{G}_0 \cap \mathcal{G}_1 = \emptyset$, into one of the groups $\mathcal{G}_0$ or $\mathcal{G}_1$. Figure 1 provides examples of ROC curves, that visualizes the probability of correctly classified $\mathcal{G}_1$ individuals against the probability of misclassified $\mathcal{G}_0$ individuals for all critical values of the diagnostic method.

In practice, ROC curves are often used to compare several diagnostic methods (classifiers). Typically it is assumed that the method with its ROC curve closest to the point $[0,1]$ is the best one for the particular problem. However, we point out that the optimality criteria can be more complex and may depend on actual situation. Figure 1 displays three plots, each with a pair of ROC curves corresponding to different association measures. It illustrates three typical situations one can meet in practice.

First, everyone would agree that the solid curve in Figure 1a outperforms the dashed one. Figure 1b seems to be the opposite case, because both association measures provide, at least optically, equivalent ROC curves. Finally, concerning the overall performance of the presented ROC curves, the situation in Figure 1c is not at all clear. In all three cases, nevertheless, a natural question arises: Are these ROC curves significantly different?

There exist several methods for testing the difference between two ROC curves, some of them being listed in references. During the lecture the problem of testing equivalence of two ROC curves will be addressed and illustrated on a real data set stemming from the field of computational linguistics. A transformation of corresponding ROC curves, that motivates a test statistic based on a distance of two empirical quantile processes will be suggested, its asymptotic distribution shown and a simulation scheme enabling to find critical values proposed. The procedure is applied to the ROC curves measuring quality of the automatic collocation extraction. It is shown that obtained $p$-values can be used as a distance between the curves and enable ROC curves clustering.

Figure 1. Examples of ROC curves for several linguistic measures.

First, everyone would agree that the solid curve in Figure 1a outperforms the dashed one. Figure 1b seems to be the opposite case, because both association measures provide, at least optically, equivalent ROC curves. Finally, concerning the overall performance of the presented ROC curves, the situation in Figure 1c is not at all clear. In all three cases, nevertheless, a natural question arises: Are these ROC curves significantly different?

There exist several methods for testing the difference between two ROC curves, some of them being listed in references. During the lecture the problem of testing equivalence of two ROC curves will be addressed and illustrated on a real data set stemming from the field of computational linguistics. A transformation of corresponding ROC curves, that motivates a test statistic based on a distance of two empirical quantile processes will be suggested, its asymptotic distribution shown and a simulation scheme enabling to find critical values proposed. The procedure is applied to the ROC curves measuring quality of the automatic collocation extraction. It is shown that obtained $p$-values can be used as a distance between the curves and enable ROC curves clustering.

REFERENCES


Acknowledgement: Work has been supported by grants GAČR 201/06/0186 and MSM 0021620839.
Authors thank Chun-houh Chen and Institute of Statistical Science Academia Sinica for hospitality, support and thorough care.

Addresses: Dept. of Probability and Statistics, Charles University in Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic; {jaromir.antoch,lubos.prchal}@mff.cuni.cz
Institut de Mathématiques de Toulouse, UMR 5219, Université Paul Sabatier, 118 route de Narbonne, F–31062 Toulouse cedex, France; sarda@cict.fr
Seung-Ho Kang earned his Ph.D. degree in statistics from the University of Wisconsin at Madison in 1997. He started his professional career as a postdoctoral fellow at the National Center for Toxicological Research, USA. He became an assistant professor in the medical school in the University of Texas at Houston in 2000. In 2001, Dr. Kang went back to Korea to continue his research and education career. His research interests include the categorical data analysis (especially exact tests) and statistical methods in clinical trials. Dr. Kang has been teaching three Ph.D. students who are statistical reviewers in Korea Food and Drug Administration since 2006.
Tests for the Homogeneity of Two Binomial Proportions in Extremely Unbalanced 2 x 2 Contingency Tables

Seung-Ho Kang
Department of Applied Statistics
Yonsei University, Seoul, Korea

Abstract
Asymptotic tests such as the Pearson chi-square test are unreliable for testing the homogeneity of two binomial probabilities in extremely unbalanced cases. Two exact tests (conditional and unconditional) are available as alternatives and can be implemented easily in StatXact 6.0. In equal sample cases it is well known that the unconditional exact test is more powerful than the conditional exact test. However, in this paper, we show that the opposite result holds in extremely unbalanced cases. The reason is that the peaks of the type I error occur at the extremes of the nuisance parameter when the imbalance among the sample sizes becomes severe. After we show that the conditional exact test is more powerful than the unconditional exact test in extremely unbalanced cases whose sample ratio is greater than 20, we compare the conditional exact test with the Berger and Boos approach (1994, Journal of the American Statistical Association 89, 1012-1016) in which the supremum is taken over a confidence interval for the nuisance parameter. The Berger and Boos approach turns out to be slightly more powerful than the conditional exact test in extremely unbalanced data. A real example is provided.
Dr. Yi Tsong, Deputy Division Director,

Division of Biometrics VI, Office of Biostatistics, Office of Translational Sciences
Center for Drug Evaluation and Research, FDA
10903 New Hampshire Ave, Silver Spring, MD 20993
e-mail: yi.tsong@fda.hhs.gov

Yi Tsong received his Ph.D. in math. Statistics from the University of North Carolina at Chapel Hill in 1979. He did his post-doctoral training in cardiovascular prevention and biostatics at Northwestern Medical School (1978-1980). He worked as senior statistician in pattern recognition at Lockheed Engineering and Management Company (1981-1983) and biostatistical consultant at The University of Texas Medical Branch at Galveston (1984-1987) before joining FDA. He served as team leader of postmarketing risk assessment, statistical reviewer of NDA submission of critical care and pain relief products. He is currently the Deputy Division Director and Acting Team Leader for statistical team of Chemistry and Manufacturing Control. He specializes in postmarketing risk assessment, drug manufacturing process control and quality assurance, active control noninferiority/equivalence tests, adaptive designs and QTc trials. He received 8 CDER and 12 FDA level awards for contributions in postmarketing drug risk assessment, for advisory on CDER postmarketing risk assessment external contracts, medication errors, quality control evaluation, drug compliance, in vitro bioequivalence, drug compliance, drug abuse potential studies, setting quota of scheduled substances, adaptive design and non-inferiority tests, et al. He publishes frequently in numerous professional journals. He served as Treasurer, Board Director and President of International Chinese Statistical Association. He serves also as the Associate Editor of Statistics in Medicine and J. of Biopharmaceutical Statistics.
**Multiple Testing Issues of Thorough QTc Trials**

Yi Tsong*, Jinglin Zhong, Meiyu Shen and Joanne Zhang  
CDER, FDA, Silver Spring, Maryland, USA  

*presenter

**Abstract**

The ICH E14, 2005 defined that drug-induced prolongation of QT interval as evidenced by an upper bound of the 95% confidence interval around the mean effect on QTc of 10 ms. Furthermore it defined that a negative thorough QT/QTc study is one in which the upper bound of the 95% one-sided confidence interval for the largest time-matched mean effect of the drug on the QTc interval excludes 10 ms. It leads to the requirement of showing non-inferiority of the test treatment to placebo at multiple time points. Conventionally, it is carried out by testing multiple hypothesis at 5% type I error rate each. On the other hand, when the study result is negative, ICH E14 recommended to validate the negative result by showing that the study population is sensitive enough to show at least 5 ms prolongation of QTc interval of a carefully selected positive control. The validation test is often carried out by demonstrating that the mean difference between positive control and placebo is greater than 5 ms at at least one of the selected few time points. The multiple comparison nature of the validation test led to the concerns of type I error rate adjustment for multiple comparisons. Boos et al (2007) showed that when the repeated measured responses follow a multivariate normal distribution with equal variance, the conventional test is conservatively biased. Based on the objective for bias correction, Boos et al (2007) proposed three confidence intervals of maximum of multiple sample mean differences with no type I error rate adjustment requirement for multiple comparisons. It has been shown that the proposed approaches can improve the power of the QTc studies with crossover design. However, Boos et al (2007) didn’t show that the proposed approaches control type I error rate. We evaluated the type I error rate of the three proposed procedures under various configurations of the null hypothesis using a simulation study. The results are compared with the conventionally used approaches by regulatory reviewers. Our simulation study shows that the all three unbiased confidence interval approaches proposed by Boos et al (2007) inflate type I error rate in the setting of demonstrating rejecting null hypothesis at all time points.
Session 2

Functional Multivariate Method
Session 2: Functional Multivariate Method
Chair: Naisyin Wang
Speaker 1: Yutaka Tanaka
  Influence analysis in ordinary, kernel and functional multivariate methods
Speaker 2: Jane-Ling Wang
  Covariate Adjusted Functional Principle Component Analysis
Speaker 3: Thomas Hotz
  Shape analysis for pre-aligned, star-shaped objects - what PCA can tell about the evolution of tree rings
Naisyin Wang is Professor of Statistics and Toxicology at Texas A&M University, where she has been on the faculty since 1992. She serves as a Biometrics co-editor 2006-2008. She is a Fellow of the American Statistical Association, a Fellow of Institute of Mathematical Statics and an elected member of International Statistics Institute. Her research interests cover Longitudinal data analysis, measurement error models, semiparametric regression, and biological/genomic applications particularly in nutrition and toxicology studies.
Dr. Yutaka Tanaka

Professor of Statistics
Department of Information Systems and Mathematical Sciences
Nanzan University
27 Seire-cho, Seto, Aichi, 489-0863, Japan
E-mail: ytanaka@ms.nanzan-u.ac.jp

Yutaka Tanaka studied mathematical engineering at University of Tokyo, Japan under the guidance of Professor Sigeiti Moriguti, and earned his Ph.D. in Mathematical Statistics at Kyushu University, Fukuoka, Japan. His Ph.D. thesis was “On some generalized methods of optimal scaling” (supervisor Professor Akio Kudo). His research interests include methods of quantification or optimal scaling, variable selection in multivariate statistical methods, in particular, in the case where no outside variable exists, sensitivity analysis in multivariate methods and development of software related to the above statistical techniques. Before moving to Nanzan University he had been a professor of statistics at Okayama University, Okayama, Japan, for 25 years and had worked for Takeda Chemical Industries as a biostatistician for 17 years. He was presented “Distinguished Achievement Award” in 1993 by Japanese Society of Computational Statistics (JSCS) for his research on sensitivity analysis in multivariate methods. He also received “Award for Software Development” jointly with T. Tarumi, Y. Mori, Y. Odaka, S. Watadani and Y. Yamamoto in 2005 from JSCS for the development of Software SAMMIF (Sensitivity Analysis in Multivariate Methods based on Influence Functions).

Publications:
3. Local influence in principal component analysis: relationship between the local influence and influence function approaches revisited, Comp. Statist. & Data Analysis, 44/1-2(2003), 143-160;
4. On local influence in canonical correlation analysis, Comm.Statist.-Theory & Methods, 31/12(2002), 2325-2347;
8. Influential observations in principal factor analysis, Psychometrika 54/3(1989), 475-485;
Influence analysis in ordinary, kernel and functional multivariate methods

Yutaka Tanaka
Nanzan University, Nagoya, Japan

Abstract

In recent years there have been developed kernel and functional multivariate methods as extensions to ordinary multivariate methods to analyze nonlinear relations and to deal with continuous functions instead of multidimensional observations. In this presentation we discuss influence or sensitivity analysis in these three types of multivariate methods.

The basic idea is to derive influence functions or differential coefficients in perturbation parameters and approximate the influence of a subset of observations as the sum of the influences of the observations which belong to the subset on the basis of the additive property of the influence (Tanaka, 1994), that is, the generalized influence function for a subset is defined as a function of the sum of the influence functions for observations in the subset. Therefore, we may search for subsets of the observations which have large norms of the influence functions to detect influential subsets of observations, where the inverse of the covariance matrix of the estimator for parameter vector is used as the metric to compute the norm. We have developed a software SAMMIF for sensitivity analysis in ordinary multivariate methods (Mori et al., 1998).

The same technique can be used to develop methods of influence analysis in kernel multivariate methods, say, in kernel principal component analysis. In kernel methods observations are mapped to a higher dimensional space. Our idea is to approximate the higher dimensional space by lower dimensional space by using principal component analysis (PCA) or other dimension reduction methods and then utilize the influence analysis in the corresponding ordinary multivariate method. We will show an example of kernel PCA of the gene network data (Tanaka and Yamanishi, 2007).

Methods of influence analysis can be developed in functional data analysis, e.g., functional PCA (Yamanishi and Tanaka, 2005). In functional data analysis a smoothing parameter lambda is introduced in estimating weight functions, and the lambda is estimated using cross-validation. We derived influence functions for weight functions of functional PCA with and without taking into account the influence through the lambda.

Influence of variables on the result of analysis will also be discussed.

References.
Dr. Jane-Ling Wang, Professor

Department of Statistics, University of California
Davis, CA 95616 USA
E-mail: wang@wal.d.ucdavis.edu

Jane-Ling Wang earned her Ph.D. degree in Statistics from the University of California, Berkeley in 1982. She has been Assistant, Associate and Full Professor at the University of California, Davis. Her research interests include functional/longitudinal data analysis, survival analysis, and non/semi-parametric modeling.
Covariate Adjusted Functional Principal Components Analysis for Longitudinal Data

Jane-Ling Wang
University of California, Davis

Abstract
Classical multivariate principal component analysis has been extended to functional data and termed Functional principal component analysis (FPCA). Much progress has been made but most existing FPCA approaches do not accommodate covariate information, and it is the goal of this talk to develop alternative approaches to incorporate covariate information in FPCA, especially for irregular or sparse functional data. Two approaches are studied: the first incorporates covariate effects only through the mean response function, but the second approach adjusts the covariate effects for both the mean and covariance functions of the response. Both new approaches can accommodate measurement errors and allow data to be sampled at regular or irregular time grids. Asymptotic results are developed and numerical support provided through simulations and a data example. A comparison of the two approaches will also be discussed.
Dr. Thomas Hotz, Research Fellow

Institute for Mathematical Stochastics, University of Goettingen
Goldschmidtstraße 7, 37077 Goettingen, Germany
E-mail: hotz@math.uni-goettingen.de

Thomas Hotz received his degree in mathematics from the University of Heidelberg, Germany in 2002. He then worked as a research associate at the Department of Epidemiology and Public Health, University of Leicester, UK. Afterwards he held a position as Associate Statistician at the United Nations Statistics Division, New York, before he returned to Germany. In 2007 he received a PhD from the Institute for Mathematical Stochastics, University of Goettingen, Germany where he currently works as a postdoctoral research fellow. His research interests comprise mathematical and statistical methods in image analysis, as well as for the analysis of shapes; he combines both when studying fingerprints.
Shape Analysis for pre-aligned, star-shaped objects
- what PCA can tell about the evolution of tree rings

Thomas Hotz
Institute for Mathematical Stochastics, University of Goettingen, Germany.

Abstract

We discuss different shape spaces for analysing pre-aligned, star-shaped objects where landmarks have been placed at fixed angles. In particular, we look for Euclidean shape spaces which render standard methodology of multivariate analysis applicable. The application that initiated this research lies in forestry where one is interested in the temporal evolution of tree rings. We compare the results of principal components analysis applied to a data set of stem disks which we performed in the different shape spaces. We demonstrate that essential features of a tree’s growth are contained in the first two principal components which capture major events in the tree's lifetime.
Session 3

Analytic and Symbolic Systems
Session 3: Analytic and Symbolic Systems
Chair: Chen-Hai Tsao
Speaker 1: Edwin Diday
   SYMBOLIC DATA ANALYSIS: a new model for DATA MINING and its SYR software
Speaker 2: Leland Wilkinson
   Identifying and Querying Local and Large-Scale Structures Using L∞ Norms
Dr. C. Andy Tsao, Professor

Department of Applied Mathematics, National Dong Hwa University
Shoufeng Township, Hualien, Taiwan 97401
E-mail: chtsao@mail.ndhu.edu.tw

C. Andy Tsao Ph.D. in Statistics, Cornell University, USA 1994. His research interests include statistical machine learning, robust Bayesian analysis, statistical evidences and statistical decision theory.
Dr. Edwin Diday Professor.

CEREMADE, Université Paris Dauphine
Place du Maréchal De Lattre De Tassigny
Paris 75016 France.

E-mail: diday@ceremade.dauphine.fr

E. Diday is Professor of Exceptional Class at the University of Paris Dauphine in Computer science and Mathematics. He was at INRIA head of project until 1994 and the Scientific Manager of two EUROSTAT project SODAS and ASSO (17 teams from 9 European countries) until 2003. He was involved in three other European consortium. He is in the editorial board of the book series: "Studies in classification, data analysis and knowledge organisation" and several International Journals. He is author or editor of 14 books and of more than 50 refereed papers. More than 50 doctorate dissertations have been obtained by students under his direction. He is past president of the Francophone Society of Classification. He is member of the International Statistical Institute. His most recent contribution concerns Maximal and Stochastic Galois Lattices, Symbolic Data Analysis and Spatial Classification. He is awarded laureate of the Montyson Price given by the French Academy of Sciences.

Recent books
http://books.elsevier.com/elsevier/?isbn=0080446124

Recent International Journal papers
SYMBOLIC DATA ANALYSIS: A new model for DATA MINING and its visual SYR software.

E. Diday
CEREMADE PARIS-DAUPHINE University Paris (France)

Abstract

The usual Data mining model is based on two parts: the first concerns the units (called here "individuals"), the second, contains their description by several standard variables including a class variable. The Symbolic Data Analysis model needs two more parts: the first concerns units called "concepts" and the second concerns their "description". The concepts are characterized by a set of properties called "intent" and by an "extent" defined by the set of individuals which satisfy these properties. These concepts are described by "symbolic data" which are standard categorical or numerical data and moreover interval, histograms, sequences of values, etc. These new kind of data allows keeping the internal variation of the extent of each concept. Then, new knowledge can be extracted from this model by new tools of Data Mining extended to concepts considered as new units. Among these tools, Spatial Classification allows a graphical visualisation of the given concepts on a grid and at different level of generalisation organised by a spatial hierarchy or pyramid (allowing overlapping clusters). The SYR software has been developed by SYROKKO company after the academic SODAS software developed by two EUROPEAN projects until 2003. The first aim of SYR is to extract, from a data file (.txt, .csv, ACCESS database) of several millions of units a reduced number of units which are "concepts" summarizing the initial data. Then SYR can create handle (select, cut, move rows or columns ...) and visualise a symbolic data file thanks to user-friendly graphical output and produce new knowledge by Symbolic Data Analysis tools.

References

**Dr. Leland Wilkinson**

Executive VP  
SYSTAT Software Inc.  
225 West Washington St.  
Chicago, IL, 60606, USA  
E-mail: leland.wilkinson@systat.com

Leland Wilkinson is Executive VP of SYSTAT Software Inc., Adjunct Professor of Statistics at Northwestern University, and Adjunct Professor of Computer Science at the University of Illinois Chicago. He received his Ph.D. from Yale University in 1975. Wilkinson wrote the SYSTAT statistical package and founded SYSTAT Inc. in 1984. He sold the company to SPSS in 1994 and worked there for ten years on research and development of visualization systems. SPSS eventually sold SYSTAT to Cranes Software International and Wilkinson rejoined SYSTAT in 2008. He is a Fellow of the American Statistical Association and an elected member of the International Statistical Institute. In addition to journal articles and the original SYSTAT computer program and manuals, Wilkinson is the author (with Grant Blank and Chris Gruber) of *Desktop Data Analysis with SYSTAT* and *The Grammar of Graphics*. 
Identifying and Querying Local and Large-Scale Structures Using $L^\infty$ Norms in Visual Analytics

SYSTAT Software and University of Illinois at Chicago

Abstract

We introduce a new visual analytic framework based on the $L^\infty$ norm. This framework involves a three-operator algebra on hyper-rectangles. We have developed a visual analytic system that generates set-wise rules from simple gestures in an exploratory visual GUI. Logging these rules allows us to apply our analysis to a new sample or batch of data so that we can assess the validity and reliability of our visual-analytic model. The basic idea is to design an analytic system around rectangular (weighted hypercube) description regions. The composition of these regions (using three operators) can be used to define local and large-scale structures and provides the basis for a formal description of structures suitable for visual analytics.
Session 4

Modeling and Applications
Session 4: Modeling and Applications
Chair: Chul Eung Kim
Speaker 1: John Aston
   Computing and Visualising Regime Variability in Hidden Markov Models
Speaker 2: Wolfgang Härdle
   Modelling and Forecasting Liquidity Supply Using Semiparametric Factor Dynamics
Speaker 3: Henry Horng-Shing Lu
   Multidimensional scaling for large genomic data sets
Chul Eung Kim, PhD, Professor

Yonsei University, Department of Applied Statistics
Suhdaemungu Shinchondong 134, Seoul 120-749, South Korea
E-mail: cekim@yonsei.ac.kr.

Ph.D. in statistics from Carnegie Mellon University, Pittsburgh, USA in 1989. From 1990 until now, assistant, associate, and full professor of statistics at Yonsei University. Currently department head.


Dr. John Aston,
Assistant Professor / Assistant Research Fellow,

Centre for Research in Statistical Methodology,
Department of Statistics, University of Warwick,
Coventry CV4 7AL, United Kingdom
E-mail: j.a.d.aston@warwick.ac.uk
and
Institute of Statistical Science, Academia Sinica
128 Academia Rd. Sec.2, Nankang, Taipei, Taiwan 11529
E-mail: jaston@stat.sinica.edu.tw

John Aston earned his Ph.D. degree in statistics at Imperial College, University of London in 2002 and was then a National Institute of Statistical Sciences (NISS) junior fellow at the US Census Bureau from 2002-2004. He started his career as an assistant research fellow at the Institute of Statistical Science, Academia Sinica in 2004. From 2008, he also works at the Centre for Research in Statistical Methodology at Warwick University. His research interests include the analysis of higher-order Markovian and hidden Markovian sequences and waiting time distributions with application to human brain mapping, genetics and econometrics; Change-point analysis using hidden Markov models; Wavelet and other methods for time series and image processing especially in human brain mapping; Non-Gaussian models for the seasonal adjustment of time series; Contingency table analysis; Functional data analysis with particular application in Linguistics; Software development for econometrics in the languages Ox and Matlab; Human brain mapping software development. He is an associate editor for Statistica Sinica and has recently organised the Statistica Sinica special issue on Statistical Challenges and Advances in Brain Science.
Computing and Visualising Regime Variability in Hidden Markov Models

John Aston
CRiSM, Dept of Statistics, University of Warwick, Coventry, UK
and
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.

Abstract

Finite state Hidden Markov models (HMMs) are used in many applications to analyse data with unobserved components or regimes. While the estimated parameters in an HMM give much information about the data in question, they also imply many other distributions associated with the regimes which are not immediately apparent. These include distributions associated with the number of regimes, the length of regimes and the variability of the positions in time where these regime changes take place. Some asymptotic results can be easily formulated based on the Markovian theory underlying the model and are well known. However, less well known is the fact that the exact finite sample distributions for these regimes can also be computationally found using techniques such as finite Markov chain imbedding.

In this talk, the computation of these exact distributions will be discussed and it will be shown that it is possible to characterise these distributions efficiently, in particular in respect to the time needed to calculate approximate distributions using sampling methods. It will also be shown that these distributions can be used to visualise the variability that is implicit in the fitted HMM, and can be used to assess whether the amount of variability is reasonable given the application. The techniques will be demonstrated with reference to genetic data and econometric data.
Wolfgang HÄRDLE did 1982 his Dr. rer. nat. in Mathematics at Universität Heidelberg and 1988 his Habilitation at Universität Bonn. He is currently chair professor of statistics at the Dept. Economics and Business Administration, Humboldt-Universität zu Berlin. He is also director of CASE – Center for Applied Statistics & Economics and director of the Collaborative Research Center “Economic Risk”.

His research focuses on dimension reduction techniques, computational statistics and quantitative finance. He has published 34 books and more than 200 papers in top statistical, econometrics and finance journals. He is one of the “Highly cited Scientist” according to the Institute or Scientific Information.

He has professional experience in financial engineering, structured product design and credit risk analysis. He cooperates with Bundesbank, Sal. Oppenheim and Deutsche Bank.
Modelling and Forecasting Liquidity Supply Using Semiparametric Factor Dynamics

Wolfgang Härdle
Humboldt University, Berlin 10178, Germany

Abstract

More and more exchanges do not only use electronic limit order books as organisational mean. There is also a recent trend towards giving traders access to the order book information. The limit order book is the major source of information when it comes to liquidity supply and is hence extremely important for liquidity risk mitigation.

As the two step functions depicting the order book - the bid and the ask curve - are defined by a high number of points whose location changes over time, the order book happens to be both a high dimensional and dynamic object. As a consequence statistical modelisation is difficult and we are currently unaware of attempts to overcome this problem in a directly data driven fashion. We propose the use of a Dynamic Semiparametric Factor Model (DSFM) that combines nonparametric flexibility with a low dimensional factor model. We use the obtained factor functions and their loadings to describe and understand order book dynamics as well as for forecasting purposes.

Besides dimension reduction, we obtain economically interpretable factor functions and an easy and parametric model for factor dynamics. The so obtained order book forecasts can directly be applied to liquidity risk mitigation.
Henry Horng-Shing Lu received his Ph.D. and M.S. degrees in Statistics from Cornell University, NY, USA, in 1994 and 1990, respectively, and his B.S. degree in electric engineering from National Taiwan University, Taiwan, ROC, in 1986. He is a Professor in the Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan, ROC. He has been a visiting scholar at UCLA, Harvard University and University of Chicago. His research interests include statistics, medical images, and bioinformatics. He and collaborators have more than 30 journal papers published or accepted, including Journal of the American Statistical Association, Journal of Multivariate Analysis, Statistica Sinica, Journal of Computational and Graphical Statistics, IEEE Transactions on Reliability/Image Processing/Medical Imaging, Pattern Recognition, Ultrasound in Medicine and Biology, Trends in Genetics, Proceedings of the National Academy of Sciences of the United States of America, Journal of Computational Biology, Bioinformatics, BMC Bioinformatics and so forth.
Multidimensional scaling for large genomic data sets

Henry Horng-Shing Lu
Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan.

Abstract
We developed a new rapid metric multidimensional scaling (MDS) method with a low computational complexity. Computer simulation showed that the new method of split-and-combine MDS (SC-MDS) is fast, accurate and efficient. In the empirical studies evaluated in this report using microarray data on the yeast cell cycle, we found that the performance of K-means in the reduced dimensional space is better than that of K-means in the original space. When we remove the back-ground noise by projecting expression profiles to the reduced dimensional MDS space, the clustering result reveals clearly the pattern of two cell cycles. The biological validation also confirms the advantage of SC-MDS. Hence, the proposed SC-MDS is useful for biological investigation using whole genome data. This is a joint work with Dr. Jengnan Tzeng and Wen-Hsiung Li.
Session 5

Graphic Concept
Session 5: Graphic Concept
Chair: Yasumasa Baba
Speaker 1: Jae-Chang Lee
    Some Thoughts on Graphical Methods in Statistics
Speaker 2: Kwan-Liu Ma
    Visualization in the Right Space
Speaker 3: Shizuhiko Nishisato
    Radical suggestions on improved graphical display of multidimensional categorical data
Dr. Yasumasa Baba, Professor Emeritus
The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan
Tel: +81-(0)3-3446-1501
E-mail: baba@ism.ac.jp

Yasumasa Baba earned his Master of Science from Tohoku University in 1970 and Doctor of Science from Kyushu University. He has been with The Institute of Statistical Mathematics, Tokyo, Japan for about 35 years. His research interests focus on Multidimensional Analysis and Research Survey. Dr. Baba is now Professor Emeritus and Adjunct Professor, Center for Engineering and Technical Support at ISM and he is currently teaching statistical courses at several universities.

Education:
Bachelor of Science, Department of Physics, Faculty of Science, Tohoku Univ., 1968
Master of Science, Graduate School of Science, Tohoku Univ., 1970
Doctor of Science, Graduate School of Science, Tohoku Univ., 1974

Degree:

Employment:
1974-1985 Researcher, Fourth Division, ISM
1985-1997 Associate Professor, Department of Statistical Methodology, ISM
1997-2005 Professor, Center for Information on Statistical Sciences, ISM
2000-2002 Director of Center for Information on Statistical Sciences, ISM
2005-2008 Professor, Department of Data Science, ISM
2008-present Adjunct Professor, Center for Engineering and Technical Support, ISM
2008-present Professor (by Special Appointment), Transdisciplinary Research Integration Center, Research Organization of Information and Systems

Research Fields: Multidimensional Analysis

Research Themes: Dynamical analysis of multidimensional data
Dr. Jae Chang LEE, Professor Emeritus,

Department of Statistics, Korea University
5-1 Anam-dong, Sungbuk-ku, Seoul, Korea 136-701
E-mail: jaeclee@korea.ac.kr

Jae Chang LEE received his Ph.D. in statistics from the Ohio State University, Columbus, Ohio, USA in 1972. He served as chief in the Statistical Engineering Lab of Korea Standard Research Institute as a founding member. Since 1979 he was a professor of statistics at Korea University until he retired from the position in 2007. His research interests include nonparametric inference, multivariate analysis and combining data from different sources.

He is an elected member of Korean Academy of Science and Technology. He was President of the Korean Statistical Society, President of IASC and Vice President of ISI. He serves currently as a Co-Editor of CSDA and a member of Scientific Advisory Board of the Institute of Statistical Mathematics in Japan. He is a visiting Distinguished Graduate Professor at Chuo University in Tokyo now.
Some Thoughts on Graphical Methods in Statistics

Jae Chang LEE
Korea University, Seoul, Korea

Abstract
Various methods of statistical graphs and graphic representations of multivariate data are reviewed. A few desirable characteristics of graphic presentations in statistics are examined and classified. Some useful properties of a good statistical graph will be discussed for practical evaluation: simplicity, helping to retain in memory, maximal information, detection capability, invariance, continuity, and full utilization of computer capability.
Dr. Kwan-Liu Ma, Professor of Computer Science

University of California at Davis
One Shields Avenue, Davis, CA 95616, USA
E-mail: ma@cs.ucdavis.edu

Kwan-Liu Ma is a professor of computer science at the University of California-Davis. He leads the VIDI (Visualization and Interface Design Innovation) research group, and directs the DOE SciDAC Institute for Ultra-Scale Visualization. His research spans the fields of visualization, high-performance computing, and user interface design. Professor Ma received his PhD in computer science from the University of Utah in 1993. During 1993-1999, he was with ICASE/NASA LaRC as a research scientist. In 1999, he joined UC Davis. In the following year, Professor Ma received the Presidential Early Career Award for Scientists and Engineers (PECASE) for his work in parallel visualization. In 2001, he received the Schlumberger Foundation Technical Award for his work in large data visualization. In 2007, he received the College of Engineering's Outstanding Mid-Career Research Faculty Award.

Professor Ma is presently leading a team of 20 researchers working on projects in large data visualization, information visualization, visual interface design, artistically inspired illustrations, and volume visualization. He actively serves the research community by playing leading roles in premier visualization meetings. In 2008, he serves as a paper chair of the IEEE Visualization Conference, Eurographics Symposium on Parallel Graphics and Visualization, IEEE Pacific Visualization Symposium, and Workshop on Visualization for Security (VizSEC). Professor Ma also serves on the editorial boards of the IEEE Computer Graphics and Applications and the IEEE Transactions on Visualization and Graphics.
Visualization in the Right Space

Kwan-Liu Ma
University of California, Davis

Abstract
For every type of visualization there is a most convenient space, and a change into the right space inevitably makes relationships clearer. I will present several case studies demonstrating that being in the right space leads to more effective visualization.
Dr. Shizuhiko Nishisato, Professor Emeritus
University of Toronto, Canada
Visiting Professor, Kwansei Gakuin University, Japan
E-mail: snishisato@oise.utoronto.ca

Career: McGill University, University of Toronto, Ontario Institute for Studies in Education
Date of Birth: June 9, 1935 (Sapporo, Japan)
Citizenship: Canadian
Home Address: 9 St. George's Road, Toronto, Ontario M9A 3S9, Canada [Telephone 1-416-239-7503]

Education: Ph.D. (University of North Carolina at Chapel Hill, U.S.A), M.A. (1961) & BA (Hokkaido University, Japan)

Visiting professorships: Kwansei Gakuin University (Japan), Doshisha University (Japan), Universidad de Murcia (Spain), Institute of Statistical Mathematics (Japan), Universitäät Karlsruhe (TH) (Germany)

Academic Societies
Past President of the Psychometric Society, Past Editor of Psychometrika, Past Member of the Board of Trustees (Psychometric Society), Past Member of the Board of Directors (Classification Society of North America), Member of the Editorial Board for Special Series on "Studies in Classification, Data Analysis and Knowledge Organization" Springer-Verlag (Germany).

Awards
Fellow of the American Statistical Association, the Distinguished Alumnus Award (the University of North Carolina Psychology Alumni Association.), The Hayashi Award (Behaviormetric Society of Japan), The 1994 Ontario Volunteer of the Year Service Award by the Province of Ontario, Canada.

Major Books

Other Publications: Some 130 papers, many of which are on dual scaling (quantification of categorical data)
Radical suggestions on improved graphical display of multidimensional categorical data

Shizuhiko Nishisato
University of Toronto, Canada

Abstract
In quantification theory of categorical data, it is almost an established procedure to show the configuration of multidimensional categorical data by symmetric scaling, often called a French plot. This talk identifies a number of theoretical problems associated with symmetric scaling, and offers a new look at the problem of graphical display. The proposed procedure is based on total information in data without distorting relations between data points, with implications for discarding the traditional notion of the principle of parsimony (the idea of projection) and overlaying the row space onto the column space. It is an easily understandable, common-sense procedure, the only obstacle being on how to convince researchers to discard a time-honored traditional widely-used current practice.
Session 6

Genomic Statistics
Session 6: Genomic Statistics
Chair: Ueng-Cheng Yang
Speaker 1: James Chen
   Gene Set Analysis in Expression Profiling Studies
Speaker 2: Seiya Imoto
   Unraveling Dynamics of Biological Networks towards Computational Drug Target Discovery
Speaker 3: W. Y. Wendy Lou
   Computational Methods for Cluster Detection of Patterns in Biology Sequences
Dr. Ueng-Cheng Yang, Associate Professor

Institute of Biomedical Informatics, National Yang-Ming University
No. 155, Sec. 2, Linong St., Beitou District, Taipei City 112, Taiwan
E-mail: yang@ym.edu.tw

Ueng-Cheng Yang earned his Ph.D. degree from Department of Molecular Biology, Princeton University. He started his professional career as a postdoctoral fellow in Department of Chemistry, Yale University. Dr. Yang’s research interests include genomics, gene expression, proteomics data analysis, disease gene discovery, and pathway analysis. He is now Director of Center for Systems and Synthetic Biology at National Yang-Ming University.
Dr. James J. Chen, Mathematical Statistician

National Center for Toxicological Research, U.S. Food and Drug Administration
3900 NCTR Rd., HFT-20, Jefferson, AR, 72079, USA
E-mail: JamesJ.Chen@fda.hhs.gov

James J. Chen is Senior Biomedical Research Service, U. S. Food and Drug Administration. Dr. Chen received a Ph.D. degree in Statistics from Iowa State University (1978). He is Adjunct Professor in the Department of Biostatistics, University of Arkansas for Medical Science and Adjunct Professor in the Department of Public Health, China Medical University, Taiwan. Dr. Chen is an associate editor of Journal of Biopharmaceutical Statistics, Journal of the Chinese Statistical Association, BMC Research Notes, and member of Editor Board of the Open Bioinformatics Journal. Dr. Chen is an elected Fellow of the American Statistical Association (1996).

Dr. Chen has over 200 scientific publications in peer-reviewed journals and numerous invited subject review articles. Dr. Chen has served on the FDA, EPA, and interagency committees and workshops that directed at developing scientific and regulatory issues and guidelines, and has provided consultations to FDA and EPA scientists on the statistical analysis of toxicological data and on risk assessment procedures. Dr. Chen has collaborated with many scientists on a broad spectrum of research problems. His current research interests are 1) Statistical methods for biomarker identification, and 2) statistical modeling for quantitative risk assessment.
Gene Set Analysis in Expression Profiling Studies

James J. Chen
National Center for Toxicological Research, U.S. FDA
Jefferson, AR 72079, U.S.A.

Abstract

Biological phenomena often occur through the interactions of multiple genes, via signaling pathways, networks, or other functional relationships. Gene set analysis (GSA) is a statistical approach to determine whether some functionally predefined classes of genes express differently (enrichment and/or deletion) in different phenotypes. Most GSA methods use a univariate gene-by-gene analysis or assume independence without accounting for functional relationships among genes in the gene set. We propose using global and MANOVA (multivariate analysis of variance) tests for the one-sided and two-sided hypothesis, respectively. The one-sided test means that the changes of gene expressions in the gene class are in one direction: either up or down. The two-sided test means that changes of gene expressions in the gene class can be both up- and down-regulation. The global test and MANOVA tests are compared to several existing GSA methods. The proposed tests are shown to perform well in terms of the control of Type I error and power. Several publicly available microarray datasets under two and three experimental conditions are analyzed.
Dr. Seiya Imoto, Associate Professor,

Human Genome Center, Institute of Medical Science,
University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
E-mail: imoto@ims.u-tokyo.ac.jp

Seiya Imoto is currently an Associate Professor of Human Genome Center, Institute of Medical Science, University of Tokyo. He received BS, MS, and Ph.D. in Mathematics from Kyushu University in 1996, 1998 and 2001, respectively. His current research interests cover Bayesian analysis of high-dimensional heterogeneous data, DNA microarray gene expression data analysis, network analysis for life sciences, data assimilation for pathway simulation model construction and computational drug target discovery.
Unraveling Dynamics of Biological Networks towards Computational Drug Target Discovery

Seiya Imoto
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639 Japan

Abstract
In recent years, much attention has been focused on predicting mode-of-action of a drug as a gene interaction pathway by using genome-scale transcriptome data. Based on the estimated gene interaction pathways, it is believed that increasing therapeutic efficacy of a drug would be achieved by finding signaling pathways that control drug’s mode-of-action. Strategic methodology, however, has not been well-established in order to find such signaling pathways, because of the lack of comprehensive protein profiling technology. Here, we present a computational method for finding them by integrating genome-scale transcriptome expression profile data and proteome protein-protein interaction data. The method can identify the drug-affected gene interaction pathways and their dynamic behavior, and find signaling pathways strongly affecting them. The method is applied to elucidate the mode-of-action of an anti-hyperlipidemia drug, fenofibrate, in human endothelial cells. We demonstrate the method by finding a set of signaling pathways that have the potential to increase the efficacy of fenofibrate.

References


Dr. W.Y. Wendy Lou, Professor

Dalla Lana School of Public Health, University of Toronto
155 College Street, Toronto, Ontario, Canada M5S 3M7
E-mail: wendy.lou@utoronto.ca

W.Y. Wendy Lou is a Professor in the Dalla Lana School of Public Health and the Department of Statistics at the University of Toronto. She holds the Canada Research Chair in Statistical Methods for Health Care. Her work focuses on the development of statistical methodology for the study of chronic conditions and quality improvement, as well as on biomedical applications of the distribution theory of runs and patterns. After completing her Ph.D. at the University of Toronto, she held a faculty position (Assistant, Associate Professor) in the Department of Biomathematical Sciences at the Mount Sinai School of Medicine of New York University, prior to returning to Toronto. Her current professional services include being an Associate Editor for *Methodology and Computing in Applied Probability*, the Publications Officer of the Statistical Society of Canada (SSC), and the Program Chair of the 2009 SSC Annual Meeting in Vancouver.
Computational Methods for Cluster Detection of Patterns

in Biological Sequences

W.Y. Wendy Lou
University of Toronto, Toronto, ON, Canada M5S 3M7.

Abstract
The detection of clusters of patterns, either identical or non-identical, in a long biological sequence is often of interest in establishing genetic signals. The probability of occurrence of a cluster (event) for a given sequence, assuming some underlying distribution for the bases (outcomes) that make up the biological sequence, can be used as a tool for this purpose. In this talk, a motivating example involving the detection of palindrome clusters in DNA sequences will be first presented, followed by methods we propose to achieve the computational objectives. The total number of palindrome clusters, defined through a general concept of runs and patterns, is finite Markov chain imbeddable in the sense of Fu and Koutras (1994). Numerical examples will be given to illustrate our results.
Session 7
Visualization and Application
Session 7: Visualization and Application
Chair: Eunsik Park
Speaker 1: Jürgen Symanzik
    Statistical Graphics & Visual Data Mining in the Medical Field
Speaker 2: Catherine Hurley
    Comparison sequences for visualization: applications and algorithms
Dr. Eunsik Park, Assistant Professor

Department of Statistics, Chonnam National University
300 Yong-bong-ro, Buk-gu, Gwangju 500-757, Korea
E-mail: espark02@chonnam.ac.kr

Eunsik Park earned her Ph.D. degree in statistics from Seoul National University, Korea in August 1995. Prior to the current appointment, she was a research statistician at the University of North Carolina, Chapel Hill in 1998-2000 and at the International Vaccine Institute in 2000-2004. Her research interests include longitudinal data analysis, clinical trials, generalized linear models, and data mining.
Jürgen Symanzik earned his Ph.D. degree in Statistics and Computer Science from Iowa State University in Ames, Iowa (USA), in 1996. After two years as a Postdoctoral Researcher at George Mason University in Fairfax, Virginia (USA), he joined Utah State University in Logan, Utah (USA), as an Assistant Professor in 1999. Since 2005, Dr. Symanzik is an Associate Professor (with tenure) at the same institution. His research interests include all forms of statistical graphics and visualization, e.g., static/dynamic/interactive, on the Web, in a geographical framework, and in combination with virtual reality. In addition, he is interested in visual data mining and the incorporation of new technologies in his teaching duties.

Dr. Symanzik is the current North American Editor for Computational Statistics since 2005 and he was an Associate Editor for this journal from 1998 to 2005. He was a co-chair of the 2003 Annual Meeting of the Interface Foundation of North America (IFNA). He held multiple offices (at the regional and the national level) within the American Statistical Association and he was a member of the IFNA Board of Directors (from 2002 to 2007). In 2007, Dr. Symanzik became an elected member of the International Statistical Institute (ISI).
**Statistical Graphics & Visual Data Mining in the Medical Field**

Jürgen Symanzik  
Utah State University, Department of Mathematics and Statistics,  
3900 Old Main Hill, Logan, UT 84322-3900, USA

**Abstract**

In this talk, we will demonstrate how statistical graphics and visual data mining can be applied to a variety of biostatistical research projects. First, we will see how various graphical methods can be used for the simultaneous exploration of cgh and gene expression arrays. Next, we will demonstrate how graphics can be used to further assess numerical results in the context of expert raters' agreement for diagnosing carpal tunnel syndrome. Finally, we will present exploratory graphical displays for various components of functional actigraphy data. Actigraphy is an emerging technology for measuring a patient's overall activity level nearly continuously over time.
Dr. Catherine Hurley, Senior Lecturer,
Department of Mathematics,
National University of Ireland Maynooth,
Maynooth, Co. Kildare, Ireland.
E-mail: catherine.hurley@nuim.ie

Catherine Hurley earned her PhD degree in statistics at the University of Washington in 1987. After a two-year post-doctoral position at the University of Waterloo, Canada, she moved to George Washington University, U.S.A. in 1989 as an assistant professor. In 1995 she took up her current position as a lecturer in the mathematics department of what was then S. Patrick’s College, now renamed as National University of Ireland Maynooth. Her research interests are statistical software, graphics and data visualization.
Comparison sequences for visualization: applications and algorithms

Catherine B. Hurley
National University of Ireland Maynooth.

Abstract
Visualization methods are important in the exploration, analysis and presentation of data. A carefully chosen graphic creates a visual impression of the overall behaviour of a dataset or a model fit. At a detail level, comparisons are important, that is, comparisons between variables, cases, groups, clusters or models. We present algorithms which construct sequences of all pairwise comparisons, and visit more interesting comparisons first. These sequences are in fact edge traversals of weighted graphs, based on eulerian tours and hamiltonian decompositions. Applications range from a new display for pairwise comparison of treatment groups, to a guided parallel coordinate display and on to a road map for dynamic scatterplots.
This is joint work with R.W. Oldford.
Session 8

Data Discretization, Ranking, and Rounding
Session 8: Data Discretization, Ranking, and Rounding
Chair: Chuan-kai Yang
Speaker 1: Moon Yul Huh
  Discretization and mutual information estimation
Speaker 2: Hans Schneeweiss
  Probabilistic rounding
Speaker 3: Philip L.H. Yu
  Decision tree modeling for ranking data
Chuan-kai Yang received his Ph.D. degree in Computer Science from Stony Brook University, New York, USA, in 2002, and his M.S. and B.S. degree in Computer Science and in Mathematics from National Taiwan University in 1993 and 1991, respectively. He is currently an associate professor of the Information Management department, National Taiwan University of Science and Technology. His research interests include computer graphics, scientific visualization, multimedia systems, and computational geometry.
Moon Yul HUH earned his Ph.D. degree in statistics from Southern Methodist University, Dalla, Texas (USA) in 1978. He started his professional career as an research scientist at Korea Institute of Science and Technology in 1978. In 1981, Dr. Huh moved to Sungkyunkwan University, Seoul, Korea as an associate professor, and now is professor. His research interests include data/information visualization, and data mining. Professor Huh served as the President of Korea Statistical Society during 1998-1999, and is the president-elect of ARS which starts from 2009.
Discretization and Mutual Information Estimation

Moon Yul HUH
Sungkyunkwan University, Seoul, Korea.

Abstract
I am going to talk about the properties of discretization in estimating mutual information (MI) and suggest a discretization method LU-Chi. The method applies discretization only to a subset of the whole data set, and the Chi-square homogeneity test of a contingency table is used to determine stopping further splitting an interval. The efficiency of the method will be demonstrated using the Monte-Carlo experiments and the real data sets. Also shown will be the efficiency of applying the Laplace adjustment to the cell frequencies of the discretized data sets when the sample size is not large (less than 300).
Hans Scheeweiss earned his Ph.D. in mathematics at the University of Frankfurt (Germany) in 1956 and his second Ph.D. (Habilitation) in statistics and econometrics at the University of Saarbrucken (Germany) in 1964. He became professor at the University of Saarbrucken and full professor at the University of Munich since 1973. He is professor emeritus since 2001. He held visiting positions, among others, at the Institute of Advanced Studies in Vienna (Austria), at the University of Waterloo (Canada), at the Shanghai University of Finance and Economics and at the Academia Sinica at Taipei. His research interests include measurement error models, structural models, factor analysis, and decision theory.
Probabilistic Rounding

Schneeweiss, H. and Komlos, J.
Department of Statistics, Ludwig-Maximilians-University, Akademiestr. 1/I, 80799 Munich, Germany

Abstract
Using rounded data in estimating moments and regression parameters gives rise to biased estimates. With simple rounding, the bias can be approximately eliminated by using Sheppard's correction. In some applications, notably in historical surveys, data are often rounded asymmetrically such that some rounded values are preferred over other ones, e.g., even values over odd values or integers over half integers.

Probabilistic rounding is a model that can handle this case. It is based on a rounding profile function and comprises deterministic asymmetric rounding and mixture rounding as special cases. A variant of Sheppard's correction can be derived, which depends on the profile function and generalizes the common Sheppard's correction.
Dr. Philip L.H. Yu, Associate Professor,
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam Road, Hong Kong, China
E-mail: plhyu@hku.hk

Philip Yu earned his Ph.D. degree in statistics from The University of Hong Kong in 1993. He is now an associate professor at the Department of Statistics and Actuarial Science of The University of Hong Kong. His research interests include data mining, analysis of ranking data and risk management. Applying advanced statistical methodology and high performance computing technology in .NET framework, his recently developed software on asset allocation and risk management won the “Best Web Services Applications for Smart Client” in 2005.

Current Professional Activities:
1. Associate Editor, Computational Statistics and Data Analysis
2. Associate Editor, Journal of Probability and Statistics
3. Vice President, Hong Kong Statistical Society
4. Scientific Secretary, Asian Regional Section, International Association of Statistical Computing
5. Fellow member, International Statistical Institute
6. Member, Technical Committee of Computational Finance and Economics, IEEE Computational Intelligence Society
Decision tree modeling for ranking data

Philip L.H. Yu
The University of Hong Kong, Hong Kong, China.

Abstract

Ranking data arises from many applications in marketing, psychology and politics. We establish new decision tree models for the analysis of ranking data by adopting the concept of classification and regression tree. We modify the existing splitting criteria, Gini and entropy, which can precisely measure the impurity of a set of ranking data. Two types of impurity measures for ranking data are introduced, namely n-wise and top-k measures. Minimal cost-complexity pruning is used to find the optimum-sized tree. In model assessment, the area under the ROC curve (AUC) is applied to evaluate the tree performance. The proposed methodology is implemented to analyze a partial ranking dataset of Inglehart's items collected in the 1993 International Social Science Programme survey. Change in importance of item values with country, age and level of education are identified. (Joint work with W.M. Wan and Paul H. Lee)
Session 9

Interactive Graphics and Animation
Session 9: Interactive Graphics and Animation

Chair: Han-Ming Wu

Speaker 1: Dae Heung Jang
   Using Maple animation for statistics education

Speaker 2: Antony Unwin
   Statistical Objects for Interactive Graphics:
   What can you do with which and how?

Speaker 3: Simon Urbanek
   Web Based Interactive Graphics
**Dr. Han-Ming Wu, Assistant Professor**

Department of Mathematics, Tamkang University  
151 Ying-chuan Road Tamsui  
Taipei County 25137, Taiwan, R.O.C.  
E-mail: hmwu@mail.tku.edu.tw

Han-Ming Wu received his Ph.D. degree in Statistics from the National Chiao Tung University (Taiwan) in 2003. He was a Postdoctoral Fellow in the Institute of Statistical Science, Academia Sinica, Taipei, Taiwan during 2003-2008. Currently he is a full time Assistant Professor in the Department of Mathematics, Tamkang University (Taiwan). His research interests include information visualization, bioinformatics, medical image processing, and statistical computing.
Dr. Dae-Heung Jang, Professor,

Pukyong National University
599-1 Dacecon-dong, Nam-gu, Busan, KOREA
E-mail: dhjang@pknu.ac.kr

Dae-Heung Jang earned his Ph.D. degree in statistics from Seoul National University, Seoul (KOREA) in 1991. He started his professional career as a full-time lecturer of statistics at Pukyong National University in 1983. His research interests include experimental designs(response surface methodology), statistical quality control, statistical graphics, and statistics education.
Using Maple animation for statistics education

Dae-Heung Jang
Pukyong National University, 599-1, Daeyeon-dong, Nam-gu, Busan, KOREA

Abstract

If we, statistics teachers, have incorporated animations into demonstrations in our statistics course teaching, we can provide students vivid moving images to tie to statistical modeling and concepts. Also, animations can enhance the student’s enjoyment of statistical course.
Curriculum Vitae


Degrees: M.A. (Cantab.), M.Sc. (London), Ph. D. (Dublin)

Employment: Trinity College, Dublin University
Senior Lecturer 1988-1993
Augsburg University
Professor for Computer-Oriented Statistics and Data Analysis 1993 -

Societies: Fellow, Royal Statistical Society
Fellow, American Statistical Association (elected 2007)
Member, International Statistical Institute
Member, International Association for Statistical Computing
Member, Deutsche Statistische Gesellschaft

European Section Chairman of IASC (1998-2000)

Journals
Joint Editor of
Software editor of Journal of Statistical Software (2005 - )
Associate Editor of Journal of Computational and Graphical Statistics (1997-2006)

Books

Statistical Objects for Interactive Graphics:
What can you do with which and how?

Antony Unwin¹

¹ Dept of Computer Oriented Statistics and Data Analysis, University of Augsburg, 86135 Augsburg, GERMANY.
E-mail: unwin@math.uni-augsburg.de

Keywords: Interactive Graphics, Statistical Objects

1 Abstract

Interactive statistical graphics have been around for many years and there are several excellent software packages offering a range of interactive features. Surprisingly, there is little theory lying behind this work and much of it remains ad hoc. One theoretical approach is to concentrate on the statistical objects in a display and consider how they can and should be interacted with. Defining what is or is not a statistical object is usually simple (a bar in a barchart is, the boundary of the bar is not). Determining possible interactions is not so simple (what can you do with the objects making up a boxplot?). These issues and the links between theory and practice in interactive graphics are discussed in this paper.

References


Dr. Simon Urbanek, Senior Member of Technical Staff

Department of Statistics, AT&T Labs – Research
E-mail: urbanek@research.att.com

Simon Urbanek is Senior Member of Technical Staff in the Department of Statistics at AT&T Labs – Research, specializing in visualization, exploratory model analysis, ensemble models, data streams and data mining. He received his Ph. D. in Statistics from the University of Augsburg, Germany in 2004. He joined AT&T Labs – Research, USA in 2004. Simon has been working on visualization and analysis of large datasets and has developed among others interactive visualization software iPlots and Klimt as well as numerous R packages. Simon is author of the book "Exploratory Model Analysis" and co-author of the book "Interactive Graphics for Data Analysis: Principles and Examples". He is a member of the R-core development team.

e-mail: urbanek@research.att.com

Web: http://research.att.com/~urbanek/
Web Based Interactive Graphics

Simon Urbanek 1

1 AT&T Labs - Statistics Research, Florham Park, NJ, USA.
E-mail: urbanek@research.att.com

Keywords: data analysis, web based methods, R, data mining, interactive graphics

1 Abstract

Interactive graphics are a very valuable and established tool for exploring data, complementing analyses and interpreting results. Good interactive software comes usually as stand-alone program that is used by skilled analysts. However, the benefits drawn from an interactive approach could be applied beyond the scope of a single data analysis. This would require a more open and readily available framework that would allow fully interactive visualization ranging from analysis of reports to data mining drill-down. The Web 2.0 infrastructure allows the creation of highly interactive content. With the proper infrastructure it is possible to leverage this technology to create highly interactive visualizations that can be used for a variety of tasks ranging from data analysis, collaborative work, monitoring, education to interactive report generation. In this talk we will review various approaches used, highlight the requirements for interactive graphics on the Web and present practical examples along with a framework that enable the creation of interactive graphics and analytic Web applications.
Session 10

Data/Dimension Reduction with Visualization
Session 10: Data/Dimension Reduction with Visualization

Chair: Yuh-Jye Lee

Speaker 1: Daniel B. Carr
Visualizing Global Cluster-Compressed Multivariable and Multi-altitude Atmospheric Data

Speaker 2: Yueh-Yun Chi
Object Oriented Analysis for Image Segmentation

Speaker 3: Masahiro Mizuta
Multidimensional scaling - when the dissimilarity data are distributions
Yuh-Jye Lee earned his Ph.D. degree in computer sciences from the University of Wisconsin-Madison in 2001. He started his professional career as an assistant professor of at National Chung Cheng University in 2001. From 2002 to 2007, he was an assistant professor at National Taiwan University of Science and Technology. Now, He is an associate professor in the Dept. of Computer Science and Information Engineering, National Taiwan University of Science and Technology. His research interests are in machine learning, data mining, optimization, information security and operations research. Dr. Lee developed new algorithms for large data mining problems such as linear and nonlinear regression (classification) problem, clustering and semi-supervised learning. These algorithms have been applied to solve many real world problems such as intrusion detection system (IDS), bankruptcy prognosis, microarray gene expression analysis and breast cancer diagnosis and prognosis.
Dr. Daniel B. Carr, Professor
Department of Statistics (MS4A7), George Mason University,
Fairfax, Virginia, 22030, US
E-mail: dcarr@gmu.edu

Daniel Carr earned his Ph.D. degree in statistics from the University of Wisconsin, Madison (USA) in 1976. He received an M.S. in statistics and M.Ed. in counseling from Oregon State University and Idaho State University, respectively, in 1972. This occurred while teaching high school mathematics for four years. (In 1970 he introduced a computer programming class.) He earned a BA in mathematics and psychology from Whitman College in 1968.

Work Experience
1976-1990 Research Scientist: Pacific Northwest National Laboratory (Battelle Northwest)
1990-Present Associate and Full Professor, George Mason University

His research interests involve statistical graphics and both scientific and information visualization. This scope of these interests includes algorithms that support tasks such as communication, exploration, modeling, evaluation, searching and prioritization. The scope also includes graphics design based on expert opinion, studies of human perception and cognition, and usability studies.

His areas of application have emphasized human health and the environment while also addressing a wide range of topics including genomics and proteomics. A long term goal has been to advance the use of statistical graphics by the US Federal Statistical Agencies. There has been some success. (See http://statecancerprofiles.cancer.gov/micromaps/)

His visualization software development includes Explor4, ExplorN, Link Micromaps, CCmaps, DPnets and GLISTEN and other designs implemented in Splus or R. His class on scientific and statistical visualization has reached over 400 Ph.D. students. He has more than hundred publications, mostly related to visualization.

He is a Fellow of the American Statistical Association. He was chair elect when the Statistical Graphics section was formed and has served in some capacity almost every year since.

His historical hobbies have included playing table tennis, singing, and raising children.
**Visualizing Global Cluster-Compressed Multivariable and Multi-altitude Atmospheric Data**

Daniel Carr\(^1\) and Amy Braverman\(^2\)

\(^1\)Department of Statistics, George Mason University, Fairfax, VA 22030, US
\(^2\)Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109-8099 US

**Abstract**

This talk addresses the challenge visualizing global cluster-compressed atmospheric data derived from the Atmospheric Infrared Sounder (AIRS). The AIRS data includes radiance values for 2378 spectral channels collected as the Aqua satellite orbits the earth. NASA processing converts this data to geophysical parameters. The overviews pursued here are related to NASA Level III\(^*\) products that consist of cluster-compressed summaries for 5 x 5 degree grid cells covering the earth. Each typical grid cell has many summary clusters and each individual cluster has 35 variables including temperature and water vapor at 11 altitudes and cloud fraction at 10 altitudes. This talk shows four approaches to visualizing these summaries. Two approaches use dynamic graphic packages (GLISTEN and CCmaps) that do not scale to the whole earth but are of potential interest for localized studies. Older dynamic software called CystalVision does scale and could be refined to help provide insights. The fourth approach, implemented using R, clusters grid cells of the earth based on their multiple cluster-compressed summaries. The resulting four maps for the Northern Hemisphere winter months of years 2002 to 2005 show interesting patterns. For example one cluster of grid cells involves ocean cells west of South America and Africa. The atmospheric similarities are likely related to the upwelling of cold water from Antarctica. Assessment of grid cell similarities over both space and time (years) would involve much more processing but provide a more coherent basis for visualizing clusters over time. Overviews suggestive of emergent patterns can motivate more detailed study.
Dr. Yueh-Yun Chi, Assistant Professor,

Division of Biostatistics
Department of Epidemiology and Health Policy Research
University of Florida
1329 SW 16th Street, Gainesville, FL 32610
E-mail: yychi@biostat.ufl.edu

Yueh-Yun Chi earned her Ph.D. degree in Biostatistics from University of North Carolina in 2005. She continued her biostatistical training as a postdoctoral fellow at University of Washington in Seattle. In 2007, Dr. Chi joined University of Florida as an assistant professor.

She has been collaborating actively with neurosurgeons and medical imaging scientists. Her continuing interest in medical imaging and metabolomics has led to the development of her methodological research.
Object Oriented Analysis for Image Segmentation

Yueh-Yun Chi
Department of Epidemiology and Health Policy Research, University of Florida,
Gainesville, FL

Abstract
High throughput technologies in medical imaging, genetics, and chemical analysis generate an ever increasing number of variables for each independent sampling unit. Such high dimension, low sample size (HDLSS) data present a grand challenge to statisticians: how do we find good estimates and make credible inference? We demonstrate the process of identifying valid and principled data reductions in comparing human and computer accuracy in segmenting kidneys in CT scans. First, we reduced the three dimensional array of distances for each image comparison to a histogram (the object) to be modeled separately. Second, we used nonparametric kernel density estimation to explore distributional patterns and assess multi-modality. Third, systematic search for parametric distributions and truncated variations led to choosing a cube root transformation to an approximate Gaussian. Fourth, representing each histogram by an individually estimated distribution eliminated the HDLSS problem by reducing on average 26,000 distances per histogram to 2 parameter estimates. In the fifth and final step we used classical statistical methods to demonstrate that the two human observers disagreed significantly less with each other than with the computer segmentation. Nevertheless, the size of all disagreements was clinically unimportant relative to the size of a kidney. The hierarchal modeling approach to object oriented data created response variables deemed sufficient by both the scientists and statisticians. We believe the same strategy will succeed in many other arenas.
Masahiro Mizuta earned his Doctor degree in Engineering (Information Engineering) from Hokkaido University, Sapporo (Japan) in 1986. He started his professional career as an Assistant Professor of Behavioral Sciences at Hokkaido University in 1984. In 1988, he became an Associate Professor of Faculty of Engineering at Hokkaido University. He has been a Full Professor of Information Initiative Center of Hokkaido University since 1 April 1999. He was a Vice President of Japanese Society of Applied Statistics (2004-2005) and is a Vice President of Japanese Society of Computational Statistics (2007-2008). He is a chairperson of SPC of IASC2008. His main field of research is Computational Statistics, including graphical representation of high dimensional data, algebraic curve fitting to data, dimension reduction methods and functional data analysis.
Multidimensional Scaling - when the dissimilarity data are distributions

Masahiro Mizuta
Hokkaido University, Sapporo 060-0811, Japan

Abstract

Multidimensional Scaling (MDS) is a powerful tool for analyzing dissimilarity data. In conventional MDS, the input dissimilarity data can be assumed (non-negative) real values. In this talk, we focus on dissimilarity data which are represented as distributions. We assume that the distributions are non-central chi-square. A method which constructs a configuration of concepts with normal distributions is proposed.

The dissimilarity between objects $i$ and $j$ is assumed as non-central chi-square $\chi^2(p, \delta_0 / \gamma_0)$ multiplied by a scalar (say $\gamma_0$), i.e. $s_{ij} = \gamma_0 \chi^2(p, \delta_0 / \gamma_0)$. The purpose of the proposed method is to construct a configuration: $x_i \sim N(\mu_i, \alpha_i^2 I_p)$, $i = 1, 2, \ldots, n$.

Examples for the proposed method are shown with actual data.
Academia Sinica Map

1. Main Entrance
2. Institute of Biomedical Science
3. Waste Water Treatment
4. Institute of Cellular and Organismic Biology
5. Exhibition Hall of Animal Specimen
6. Research Center for Biodiversity
7. Institute of Molecular Biology
8. Institute of Biological Chemistry
9. Life Science Library
10. National Laboratory Animal Center
11. Greenhouse
12. Administration Office and Computing Center
13. Institute of Taiwan History (to be moved)
14. Institute of animal sample
15. Institute of animal sample
16. Institute of Plant and Microbial Biology
17. Institute of Mathematics
18. Ts’ui Yen-ch’ing Memorial Hall
19. Center for Survey Research
20. Institute of Statistical Science
21. Post Office
22. Ecological Pond
23. Genomic Research Center
24. Academic Activity Center (Book shop, restaurants and guest rooms)
25. Institute of Chinese Literature and Philosophy
26. Institute of Earth Science
27. Gymnasium
28. Building for Humanities and Social Sciences
29. Union Library for Humanities and Social Sciences (1-2 floors)
30. Research Center for Applied Sciences (11th floor, South)
31. Research Center for Environmental Changes (11th floor, South)
32. Institute of Sociology (5-7th floors, South)
33. Institute of Linguistics (5-7th floors, South)
34. Institute of Taiwan History (7-8th floors, North)
35. Institute of Political Science, Preparatory Office (5-6th floors, North)
36. Institute of Linguistics (Institute of Law, Preparatory Office (9-10th floors, North)
37. Research Center for Environmental Changes (Laboratory)
38. Institute of Chemistry
39. Agriculture Biotechnology Research Center
40. Research Center for Humanities and Social Sciences
41. Institute of Information Science
42. Institute of Physics
43. Research Center for Applied Sciences (Laboratory)
44. Hu shih Memorial Hall
45. Institute of Modern History
46. Institute of European and American Studies
47. Institute of History and Philology
48. Institute of Historical Relics
49. Fu Ssu-nien Library
50. Institute of Economics
51. Institute of Ethnology
52. Lingnan Fine Arts Museum
53. Archive Building, Institute of Modern History
54. Museum of Taiwan Archaeology
55. Institute of Atomic and Molecular Sciences, Institute of Astronomy and Astrophysics (Preparatory Office) located in the campus of National Taiwan University.

Address of Academia Sinica:
128, Sec. 2, Academia Road,
Nunkang, Taipei,
Taiwan, 115
Website: www.sinica.edu.tw

N S W E

Main Entrance
Metro Network