<table>
<thead>
<tr>
<th><strong>Table of Contents</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inaugural Lecture</strong></td>
</tr>
<tr>
<td>High Dimensional Clustering, Variable Selection and Social Networks</td>
</tr>
<tr>
<td><em>Jayanta K. Ghosh</em></td>
</tr>
<tr>
<td><strong>Plenary Talk I</strong></td>
</tr>
<tr>
<td>Statistical Aspects of SNP – Gene Expression Studies</td>
</tr>
<tr>
<td><em>Andreas Ziegler</em></td>
</tr>
<tr>
<td><strong>Plenary Talk II</strong></td>
</tr>
<tr>
<td>Statistical Methods for Biomedical Research on Diagnostic Ultrasound</td>
</tr>
<tr>
<td><em>Douglas Simpson</em></td>
</tr>
<tr>
<td><strong>Invited Talks</strong></td>
</tr>
<tr>
<td>Bayesian Methods for Clustering Functional Data</td>
</tr>
<tr>
<td><em>Subhashis Ghoshal</em></td>
</tr>
<tr>
<td>Bayesian Average Error Based Approach to Sample Size Calculations for Hypothesis Testing</td>
</tr>
<tr>
<td><em>Eric Reyes and Sujit K. Ghosh</em></td>
</tr>
<tr>
<td>Bayesian Analysis of Life-data under Competing Risks</td>
</tr>
<tr>
<td><em>Ananda Sen</em></td>
</tr>
<tr>
<td>A Distribution Theory for Predictive-R2 in High-dimensional Models and its Implications for Genome-wide Association Studies</td>
</tr>
<tr>
<td><em>Nilanjan Chatterjee</em></td>
</tr>
<tr>
<td>Comparison of Methods for Unsupervised Learning Methods – an Applied Study Using Proteomic Data From Colon Cancer and Simulations</td>
</tr>
<tr>
<td><em>Nairanjana Dasgupta, Yibing Oliver Chen, Rashmita Bas and Sayed S. Daoud</em></td>
</tr>
<tr>
<td>Incorporating Auxiliary Information for Improved Prediction in High Dimensional Datasets: An Ensemble of Shrinkage Approaches</td>
</tr>
<tr>
<td><em>Bhramar Mukherjee</em></td>
</tr>
<tr>
<td>Unconstrained Bayesian Model Selection on Inverse Correlation Matrices with Application to Sparse Networks</td>
</tr>
<tr>
<td><em>Nitai D. Mukhopadhyay and Sarat C. Dass</em></td>
</tr>
<tr>
<td>Analysis of Sabine River Flow Data Using Semiparametric Spline Modeling</td>
</tr>
<tr>
<td><em>Soutir Bandyopadhyay</em></td>
</tr>
<tr>
<td>Title</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Adaptive Density Regression with Gaussian Process Priors</td>
</tr>
<tr>
<td><em>Surya T. Tokdar</em></td>
</tr>
<tr>
<td>Independent Component Analysis for Dimension Reduction Classification: Hough Transform and CASH Algorithm</td>
</tr>
<tr>
<td><em>Asis K. Chattopadhyay</em></td>
</tr>
<tr>
<td>Predicting the Extent of Uniqueness of a Fingerprint Match</td>
</tr>
<tr>
<td><em>Sarat C. Dass</em></td>
</tr>
<tr>
<td>Statistical Models for Biometric Authentication</td>
</tr>
<tr>
<td><em>Sinjini Mitra</em></td>
</tr>
<tr>
<td>Robust ER-fMRI Designs under a Nonlinear Model</td>
</tr>
<tr>
<td><em>Dibyendu Majumdar</em></td>
</tr>
<tr>
<td>Statistical Planning and Inference in the Data-centric World: Applications and Challenges</td>
</tr>
<tr>
<td><em>Subir Ghosh</em></td>
</tr>
<tr>
<td>Constrained Multi-objective Designs for Functional MRI Experiments via a Modified NSGA-II</td>
</tr>
<tr>
<td><em>Abhyuday Mandal</em></td>
</tr>
<tr>
<td>Bayesian Learning of Material Density Structure and Blurring Function by Inverting Images Taken with Electron Microscopy Techniques</td>
</tr>
<tr>
<td><em>Dalia Chakrabarty</em></td>
</tr>
<tr>
<td>The Discrete Charm of Convex Optimization: from Shape-constrained Density Estimation to Empirical Bayes Decision Rules</td>
</tr>
<tr>
<td><em>Ivan Mizera</em></td>
</tr>
<tr>
<td>The use of Topographic Mapping in Visualisation and Classification: Application to Finding Planets around Other Stars</td>
</tr>
<tr>
<td><em>Somak Raychaudhury</em></td>
</tr>
<tr>
<td>Bootstrapping for Significance in Clustering</td>
</tr>
<tr>
<td><em>Ranjan Maitra</em></td>
</tr>
<tr>
<td>Modeling Climate Characteristics Using Small Area Methodology</td>
</tr>
<tr>
<td><em>Snigdhasnu Chatterjee</em></td>
</tr>
<tr>
<td>Functional Factor Analysis for Periodic Remote Sensing Data</td>
</tr>
<tr>
<td><em>Chong Liu, Surajit Ray, Giles Hooker and Mark Friedl</em></td>
</tr>
<tr>
<td>Title</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Recombination Detection and its use in Inferring the Recent Human Evolutionary History</td>
</tr>
<tr>
<td><em>Marta Melé, Marc Pybus, Asif Javed, Laxmi Parida, Francesc Calafell and Jaume Bertranpetit</em></td>
</tr>
<tr>
<td>High Resolution Mapping of DNA Methylation</td>
</tr>
<tr>
<td><em>Karin S. Dorman, X. Yin and X. Wang</em></td>
</tr>
<tr>
<td>Pathway Association Tests Using Reduced Gene Subsets for Genome-wide Association Studies</td>
</tr>
<tr>
<td><em>Jingyuan Zhao, Simone Gupta, Mark Seielstad, Jianjun Liu and Anbupalam Thalamuthu</em></td>
</tr>
<tr>
<td>An Investigation into Market Evolution and Consumer Dynamics in an Online Art Market: a Case of Modern Indian Art</td>
</tr>
<tr>
<td><em>Mayukh Dass</em></td>
</tr>
<tr>
<td>Scale Equivariant Estimators for Diffusion Models</td>
</tr>
<tr>
<td><em>L. Ramprasath</em></td>
</tr>
<tr>
<td>Correlations in Financial Time-series: An Econophysicist's Perspective</td>
</tr>
<tr>
<td><em>Anirban Chakraborti</em></td>
</tr>
<tr>
<td>Logic Trees for Multilevel Data: Application to Cancer Care</td>
</tr>
<tr>
<td><em>Mousumi Banerjee</em></td>
</tr>
<tr>
<td>Predictive Problems in High-dimensional Data: Some Examples</td>
</tr>
<tr>
<td><em>Madhuchhanda Bhattacharjee</em></td>
</tr>
<tr>
<td>Probability Machines and Risk Modeling</td>
</tr>
<tr>
<td><em>Abhijit Dasgupta and James Malley</em></td>
</tr>
<tr>
<td><strong>Contributed Papers</strong></td>
</tr>
<tr>
<td>A Bayesian Semiparametric Approach for Incorporating Longitudinal Information on Exposure History for Inference in Case-Control Studies</td>
</tr>
<tr>
<td><em>Dhiman Bhadra, Michael J. Daniels, Sungduk Kim, Malay Ghosh and Bhramar Mukherjee</em></td>
</tr>
<tr>
<td>Bayesian Estimation of Reliability Measures for a Family of Lifetime Distributions Using Masked System Lifetime Data</td>
</tr>
<tr>
<td><em>Sanjeev K. Tomer and Ashok K. Singh</em></td>
</tr>
<tr>
<td>Diffusive Limits for Adaptive MCMC</td>
</tr>
<tr>
<td><em>Gopal K. Basak and Arunangshu Biswas</em></td>
</tr>
<tr>
<td>Title</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Using Centered and Non-centered Algorithms for Simulating S&amp;P Share Index Data</td>
</tr>
<tr>
<td>Statistical Methods to Identify Clonal Variations Present in Tumour</td>
</tr>
<tr>
<td>Admixture Time Estimation Using Random Genetic Drift Model</td>
</tr>
<tr>
<td>Estimating Relationship between Random Individuals from their Genome Sequence</td>
</tr>
<tr>
<td>A Deterministic Model for HIV Infection Incorporating Identification Rate in a Heterosexual Population: Mathematical Analysis</td>
</tr>
<tr>
<td>Genome-wide Association Studies for Bivariate Sparse Longitudinal Data</td>
</tr>
<tr>
<td>The Interactions among Candidate Genes May Lead to Severity of Mental Disorders</td>
</tr>
<tr>
<td>Estimating Headrooms of Indian ADRs- An Application of Nonlinear Filtering Technique</td>
</tr>
<tr>
<td>Replacement Analysis Using Fuzzy Linguistic Theory</td>
</tr>
<tr>
<td>Mammogram Segmentation by Near Set Approach and Mass Lesions Classification with Rough Neural Network</td>
</tr>
<tr>
<td>An Application of Data Mining Tools for the Study of Shipping Safety</td>
</tr>
<tr>
<td>On Univariate Process Capability Indices for Unilateral Tolerances</td>
</tr>
<tr>
<td>Functional Data Analytic Methodology for PBPK Models</td>
</tr>
<tr>
<td>Title</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Early Marriage and the Length of First Birth Interval in India</td>
</tr>
<tr>
<td>Fitting a Mixture of Conway-Maxwell-Poisson Distributions to Count Data</td>
</tr>
<tr>
<td>New Stein’s Estimators of Regression Coefficients in Replicated Measurement Error Model</td>
</tr>
<tr>
<td>Copula Based Dependence Structures among Maternal Variables</td>
</tr>
<tr>
<td>Statistics in Industry</td>
</tr>
<tr>
<td>And the Two Twains Shall Meet</td>
</tr>
<tr>
<td>&quot;What's in a Name?&quot;</td>
</tr>
<tr>
<td>Reliability, Availability, and Productivity for Wind Turbines</td>
</tr>
<tr>
<td>Tracking Consumer Segments</td>
</tr>
<tr>
<td>Managing Expectations: Statistics in Risk Management</td>
</tr>
<tr>
<td>Posters</td>
</tr>
<tr>
<td>Multiple Comparison Procedure for Comparing Several Exponential Populations with Control in Terms of Location Parameters Under Heteroscedasticity</td>
</tr>
<tr>
<td>Optimum All Order Neighbor Balanced μ-Resolvable BIB Designs</td>
</tr>
<tr>
<td>Title</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Simultaneous Prediction Intervals for Comparing Several Treatments</td>
</tr>
<tr>
<td>with More than One Control</td>
</tr>
<tr>
<td><em>Sukhbir Singh, Anju Goyal and Vishal Maurya</em></td>
</tr>
<tr>
<td>Some Probability Models for Waiting Time to First Conception</td>
</tr>
<tr>
<td><em>R.C.Yadava, S.K. Singh, Anup Kumar and U Srivastava</em></td>
</tr>
<tr>
<td>On Comparing the Performance of Senior and Young Cricketers in</td>
</tr>
<tr>
<td>Indian Premier League</td>
</tr>
<tr>
<td><em>Hemanta Saikia and Dibyajyoti Bhattacharjee</em></td>
</tr>
<tr>
<td>Almost Unbiased Ratio and Product Type Exponential Estimators</td>
</tr>
<tr>
<td>*Rohini Yadav, Lakshmi N. Upadhyaya, Housila P. Singh and S.</td>
</tr>
<tr>
<td>Chatterjee*</td>
</tr>
<tr>
<td>Bayesian Estimation and Prediction of Generalized Inverted</td>
</tr>
<tr>
<td>Exponential Distribution</td>
</tr>
<tr>
<td><em>Sanku Dey, Tanujit Dey and Debasis Kundu</em></td>
</tr>
<tr>
<td>On the Bayes Estimators of the Parameters of Non-Zero Inflated</td>
</tr>
<tr>
<td>Generalized Power Series Distributions</td>
</tr>
<tr>
<td><em>Peer Bilal Ahmad</em></td>
</tr>
<tr>
<td>Relations for Marginal and Joint Moment Generating Functions of</td>
</tr>
<tr>
<td>Generalized Exponential Distribution Based on Lower Generalized</td>
</tr>
<tr>
<td>Order Statistics and Characterization</td>
</tr>
<tr>
<td><em>R.U. Khan, and Benazir Zia</em></td>
</tr>
<tr>
<td>On Exact Non-Parametric Alternative Tests in Two-Way Layout</td>
</tr>
<tr>
<td><em>Maidapwad S. L. and Sananse S.L</em></td>
</tr>
<tr>
<td>A Test for the Significance of Treatment Effect in a Mixture</td>
</tr>
<tr>
<td>Population</td>
</tr>
<tr>
<td><em>Bikram Karmakar, Kushal Kumar Dey and Kumaresh Dhara</em></td>
</tr>
<tr>
<td>Valedictory Address</td>
</tr>
<tr>
<td><em>Pranab K. Sen</em></td>
</tr>
</tbody>
</table>
At various times in the recent past, 2005 onwards, I have discussed the relevance of Bayesian Nonparametrics (specifically, Dirichlet mixtures of the normal) for handling non-unique clustering in high dimensional problems, for example the problem given to ISI by Professor Madhab Gagdil. The data had about 50,000 superpixels, each based on 1000 pixels of satellite pictures on the Western Ghat area, in four bands. In such problems, $k$-means clustering does not provide unique results, results depend on starting points. Dirichlet mixtures can be shown to be very close to $k$ means clustering, and the NP Bayes approach can assign probabilities to the different clusterings that come out as MCMC outputs. I also discussed a notion of central clustering and a neighborhood of it such that all satisfactory clusterings will lie there. I have developed a metric between clusterings with help from young colleagues, for example Tapas Samanta, Kajal Dihidar etc. Later I had discussions with Sourabh Bhattacharya and Sabyasachi Mukhopadhyay. In Kulis and Jordan (2011), they achieve remarkable progress on these problems. Further major insight and powerful new results and methods have been obtained recently by Bogdan et. al. (2011). They prove a deep theorem connecting $k$ means clustering with Dirichlet mixtures, throwing major new light on the whole subject of cluster analysis. In the second part of the talk I will discuss what is feasible and what is not, in high dimensional classification and testing, as in microarrays. Feasible may mean attaining an oracle (lower bound to risk), non-feasible may mean that the best one can do is no better than trivial procedures based on randomization only. These ideas are based on Ghosh et. al.(2009). I will give reasons why one would expect such limits to what is possible in high dimensional variable selection with moderate sample size in linear models. In particular this may apply to the numerous studies of SNP's. This is partly based on work with Ritabrata Dutta. In the third part of the talk, I show how one may apply these insights to speculate and ask questions about the huge social networks that have been developed in recent years. I will illustrate with theoretical and methodological questions, answers to which may help us understand and describe these networks in more meaningful ways. They may also help us produce such synthetic networks and simple models of social networks, fulfilling an old goal of SRU, ISI, a line of research introduced by Professor Suraj Bandyopadhyay. Please take this as a tribute to ISI where study of these topics was pioneered. I hope to return to these topics in new work with Ritabrata Dutta.
Plenary Talks

Statistical Aspects of SNP – Gene Expression Studies

Andreas Ziegler
Institut für Medizinische Biometrie und Statistik
Universität zu Lübeck, Germany

Co-authors: Silke Szymczak and Markus Scheinhardt

Single nucleotide polymorphism (SNP)-gene expression association studies (eQTL analysis) are performed to clarify the molecular function of genetic variants identified in genome-wide association (GWA) studies. The aim of these studies is to detect differences in gene expressions given genotype. Standard statistical approaches such as the analysis of variance (ANOVA) or the Kruskal-Wallis test can lead to false positive associations since model assumptions are often violated in SNP-eQTL studies. In fact, the empirical type I error level can be up to 99% when the nominal error level is 5%. A promising alternative for controlling the type I error are adaptive linear rank methods. Here, characteristics of the distributions are estimated in the first step. The information on skewness and tail length is then used to select an appropriate test statistic in the second. In this presentation, we compare two adaptive tests that were proposed in the literature using extensive Monte Carlo simulations of a wide range of different symmetric and skewed distributions. We propose a novel adaptive test which combines the advantages of both literature-based approaches. We illustrate the application of all three adaptive tests using two examples from different eQTL studies.

Statistical Methods for Biomedical Research on Diagnostic Ultrasound

Douglas Simpson
University of Illinois
Urbana-Champaign, IL, USA
dgs@illinois.edu

Diagnostic ultrasound is among the most widely used imaging techniques in biomedicine. Common uses include prenatal ultrasonic imaging of the fetus, echocardiogram images of the heart, ultrasound imaging of tumors in the breast and prostate. Current research aims to extend the range of applications and increase the power of ultrasonic imaging through quantitative ultrasound technology. Statistical issues and results associated with these efforts will be presented including image segmentation, pattern recognition, tissue characterization and semiparametric functional data analysis.
Invited Papers
Bayesian Methods for Clustering Functional Data

Subhashis Ghoshal
North Carolina State University, Raleigh, NC, USA
subhashis.ghoshal@ncsu.edu

Sophisticated instruments of today now allow taking essentially continuous measurements on subjects, leading to basic observations that are best viewed as functions. Clustering is about finding which functional observations are to be grouped together as similar in some appropriate sense. We expand functional observations in a wavelet basis function, and then use the coefficients in the expansion, to be called features, to represent the data in a multivariate setting. A latent variable approach using a Dirichlet process prior has become a standard technique for tying observations a priori, but in the present context, such a tying pattern may lead to unreasonable priors since two functional observations are unlikely to share all of their features. We design a prior to allow partial feature sharing which is analogous to the Indian buffet process. We propose a similarity index to measure the extent of partial feature sharing between subjects and develop Markov chain Monte-Carlo methods for posterior computation. We also study some asymptotic properties of our proposed procedure. We further extend the proposed method to accommodate covariates and modify suitably to use in functional classification problems.

Bayesian Average Error Based Approach to Sample Size Calculations for Hypothesis Testing

Sujit K. Ghosh
North Carolina State University, Raleigh, NC, USA
sgosh2@ncsu.edu

Co-author: Eric Reyes

Under the classical statistical framework, sample size calculations for a hypothesis test of interest maintain pre-specified Type-I and Type-II error rates. These methods often suffer from several practical limitations. For instance positing suitable values under a given hypothesis becomes more difficult when the null hypothesis is composite (e.g., hypotheses for non-inferiority tests). Additionally, classical methods often rely heavily on asymptotic (normal) approximations when testing two composite hypotheses, which may be questionable in many common situations. We propose a framework for hypothesis testing and sample size determination using Bayesian average errors that does not suffer from the same limitations as those methods developed under the classical framework and provides a general approach to handling simple and complex hypotheses. We consider rejecting the null hypothesis, in favor of the alternative, when a test statistic exceeds a cutoff. We choose the cutoff to minimize a
weighted sum of Bayesian average errors and choose the sample size to bound the total error for the hypothesis test. We apply this methodology to several designs common in medical studies.

**Bayesian Analysis of Life-data under Competing Risks**

Ananda Sen  
University of Michigan, Ann Arbor, MI, USA  
anandas@umich.edu

The focus of the talk is time-to-event analysis under competing risks. Such analyses are relevant for estimating hazards of an event of interest accounting for any competing events that may potentially terminate the observation. Competing risks can also arise in the context of analyzing repairable systems for which the relevant data comprise of successive event times of a recurrent phenomenon along with an event-count indicator. Such data commonly occur in areas such as manufacturing, software debugging, risk analysis, and clinical trials. When observing a recurrent phenomenon such as repeated occurrences of tumor, a terminal event such as death can act as a (possibly dependent) competing risk. Analysis of a recurrent event when there are multiple sources of recurrence (e.g. hospitalization due to different causes) is another somewhat different application in the same general area of competing risks. The talk will focus on Bayesian approaches to handle such data. In addition to the flexibility of allowing complex modeling, the Bayesian framework also provides an integrated mechanism where the cause of failure (or recurrence) may not be fully identified (masked). Parametric and semi-parametric models will be discussed in the context of appropriate applications drawn from an industrial setting as well as a clinical trial.

**A Distribution Theory for Predictive-R2 in High Dimensional Models and its Implications for Genome-wide Association Studies**

Nilanjan Chatterjee  
National Cancer Institute, Bethesda, MD, USA  
chattern@mail.nih.gov

Modern genome-wide association studies have led to the discoveries of thousands of susceptibility loci across a variety of quantitative and qualitative traits. Although the loci discovered so far have limited ability for prediction of any individual trait, heritability calculations indicate that power for predictive models can be potentially enhanced substantially by building polygenic models on larger data sets. In this talk, I will first describe a novel theoretical framework that allow evaluation of the distribution of R2 and other related measures of predictive performance of high dimensional statistical models based on the sample
size of the training dataset, the threshold for variable selection, the number of underlying predictive variables and the distribution of their effect-sizes. I use this theoretical framework together with empirical estimates of heritability and effect-size distribution for susceptibility SNPs for eight different complex traits to obtain estimates of sample-size required and optimal thresholds for SNP selection for building future polygenic models that may have prediction powers close to that of an “ideal” limiting model that could be built with an infinite sample size. The general framework we provide can be useful for planning development of prediction models in other contexts, such as for future studies of rare variants.

**Comparison of Methods for Unsupervised Learning Methods – an Applied Study using Proteomic Data from Colon Cancer and Simulations**

Nairanjana Dasgupta  
Washington State University, Pullman, WA, USA  
dasgupta@wsu.edu

*Co-authors:* Yibing Oliver Chen, Rashmita Basu and Sayed S. Daoud

Recently as analysis of large scale data like genomic and proteomic data gets popular, clustering has become a method of choice to identify patterns. The popularity stems from easy availability of clustering options in most statistical software packages. Biologists often use clustering methods to identify novel classes and groups and call them “class discoveries”. In this talk we will look at agreement between different clustering approaches – classification, hierarchical as well as model-based techniques. We used a proteomic data for colon cancer as an illustration of the methods. Based on the different proteins observed (using the number of peptides for each protein), we used hierarchical (using Jaccard and Euclidean distances), *k*-means, and model based clustering methods to group patients. Our findings suggest little agreement between the different techniques in terms of clustering. The four methods agreed for only 34 out of the 92 proteins. We study it further using simulations and get similar results. This indicates that clustering is an exploratory tool and should be used with caution in the science of “discoveries”.


Incorporating Auxiliary Information for Improved Prediction in High Dimensional Datasets: An Ensemble of Shrinkage Approaches

Bhramar Mukherjee
Michigan State University, East Lansing, MI, USA
bhramar@umich.edu

With advancement in genomic technologies, it is common that two high-dimensional datasets are available, both measuring the same underlying biological phenomenon with different techniques. We consider predicting a continuous outcome Y using X, a set of p markers which is the best measure of the underlying biological process. This same biological process may also be measured by W, coming from prior technology but correlated with X. On a moderately sized sample we have \( (Y,X,W) \), and on larger sample we have \( (Y,W) \). We utilize the data on W to boost prediction of Y by X. When p is large and the subsample containing X is small, this is a \( p>n \) situation. When p is small, this is akin to the classical measurement error problem; however, ours is not the typical goal of calibrating W for use in future studies. We propose to shrink the regression coefficients of Y on X towards different targets that use information derived from W in the larger dataset, comparing these with the classical ridge regression of Y on X, which does not use W. We also unify all of these methods as targeted ridge estimators. Finally, we propose a hybrid estimator which is a linear combination of multiple estimators of the regression coefficients and balances efficiency and robustness in a data-adaptive way to theoretically yield smaller prediction error than any of its constituents. The methods are evaluated via simulation studies. We also apply them to a gene-expression dataset. mRNA expression of 91 genes is measured by quantitative real-time polymerase chain reaction (qRT-PCR) and microarray technology on 47 lung cancer patients with microarray measurements available on an additional 392 patients. The goal is to predict survival time using qRT-PCR. The methods are evaluated on an independent sample of 101 patients. This is joint work with Jeremy Taylor and Philip S. Boonstra.

Unconstrained Bayesian Model Selection on Inverse Correlation Matrices with Application to Sparse Networks

Nitai D. Mukhopadhyay
Virginia Commonwealth University, Richmond, VA, USA
ndmukhopadhy@vcu.edu

Co-author: Sarat C. Dass

Bayesian statistical inference for an inverse correlation matrix is challenging due to non-linear constraints placed on the matrix elements. The aim of this paper is to present a new parametrization for the inverse correlation matrix, in terms of the Cholesky decomposition, that is able to model these constraints explicitly. As a result, the associated computational schemes for inference based on Markov Chain Monte Carlo sampling are greatly simplified and expedited. The Cholesky decomposition is also utilized in the development of a class of
hierarchical correlation selection priors that allow for varying levels of network sparsity. An explicit expression is obtained for the volume of the elicited priors. The Bayesian model selection methodology is developed using a Reversible Jump algorithm and is applied to a dataset consisting of gene expressions to infer network associations.

Analysis of Sabine River Flow Data Using Semiparametric Spline Modeling

Soutir Bandyopadhyay
Lehigh University, Bethlehem, PA, USA
sob210@Lehigh.edu

In this article, a modeling approach for the mean annual flow in different segments of Sabine river, as released in the NHDPlus data in 2007, as a function of five predictor variables is described. Modeling flow is extremely complex and the deterministic flow models are widely used for that purpose. The justification for using these deterministic models comes from the fact that the flow is governed by some explicitly stated physical laws. In contrast, in this article, this complex issue is addressed from a completely statistical point of view. A semiparametric model is proposed to analyze the spatial distribution of the mean annual flow of Sabine river. Semiparametric additive models allow explicit consideration of the linear and nonlinear relations with relevant explanatory variables. We use a conditionally specified Gaussian model for the estimation of the univariate conditional distributions of flow to incorporate auxiliary information and this formulation does not require the target variable to be independent.

Adaptive Density Regression with Gaussian Process Priors

Surya T. Tokdar
Duke University, Durham, NC, USA
tokdar@stat.duke.edu

While density estimation appears as one of the simplest and most widely researched nonparametric models, estimating conditional densities, also known as density regression, has posed a far greater challenge. A major effort was needed to uplift the popular Dirichlet processes to Depended Dirichlet processes suitable for modelling conditional distributions that vary smoothly over the conditioning variable. However, a much simpler framework obtains when one models densities by Gaussian processes. A smooth Gaussian process over the product space of the variable of interest and the conditioning variable, when exponentiated and separately normalized over the response variable at each value of the conditioning variables, produces a conditional density process that is smooth over the product space. In this talk, I will explore theoretical properties of such a process used as a prior distribution in a Bayesian analysis of conditional densities. Our work builds on the impressive recent work by Aad van der Vaart, Harry van Zanten and co-authors in characterizing properties of 're-scaled' Gaussian process models. I will talk about recent extensions of this framework to allow local re-scaling and rotation that are suitable for dimension reduction in multivariate density regression.
Independent Component Analysis for Dimension Reduction
Classification: Hough Transform and CASH Algorithm

Asis Kumar Chattopadhyay
Calcutta University, Kolkata, India
asiswa_stat@yahoo.com

Classification of galaxies has been carried out by using two recently developed methods viz. Independent Component Analysis (ICA) with K-means clustering and Clustering in Arbitrary Subspace based on Hough Transform (CASH) for different data sets. The first two sets are consisting of dwarf galaxies and their globular clusters whose distributions are non Gaussian in nature. The third one is a larger one containing a wider range of galaxies consisting of dwarfs to giants in 56 clusters of galaxies. Morphological classification of galaxies are subjective in nature and as a result, cannot properly explain the formation mechanism and other related issues under the influence of different correlated variables through a proper scientific approach. Hence objective classification by using the above mentioned methods are preferred to overcome the loopholes.

Predicting the Extent of Uniqueness of a Fingerprint Match

Sarat C. Dass
Michigan State University, East Lansing, MI, USA
sdass@stt.msu.edu

It is possible for fingerprints from two different persons to be closely matched with each other. A spurious match such as this should be detected to avoid a false positive identification. Given a match between a fingerprint pair, we would like to quantify the extent of this match statistically. This is possible to do if the variability of the underlying processes are accounted for and modelled adequately. The extent of a match depends on two sources of variability: (1) inter-class variability arising from the spatial configuration of fingerprint features, and (2) intra-class variability arising from image quality, variability due to the fingertip placement on the sensor and elastic distortion of the skin. To adequately capture feature variability, we develop distributions on the feature space based on marked point processes that model clustering tendencies and spatial correlations between neighboring marks. Inference is carried out in a Bayesian MCMC framework. The proposed class of models is fitted to real fingerprint images to demonstrate the flexibility of fit to different kinds of fingerprint feature patterns arising in practice. Evidence of a Paired Impostor Correspondence (EPIC) is developed as a measure of fingerprint uniqueness, and its predictive value is obtained using simulation from the fitted models to quantify the extent of an observed match.
Statistical Models for Biometric Authentication

Sinjini Mitra
California State University, Fullerton, CA, USA
smitra@exchange.fullerton.edu

The modern world has seen a rapid evolution of the technology of biometric authentication, prompted by an increasing urgency to ensure a system's security. The need for efficient authentication systems has sky-rocketed since 9/11 and they are used everywhere from criminal identification in law enforcement to immigration at airports to e-commerce transactions. Based on a person's essentially unique biological traits, these methods are potentially more reliable than traditional identifiers like PINs and ID cards. Although the importance of biometrics is well-understood, there are several challenges that biometric-based authentication methods face today, such as scalability, uniqueness, privacy issues, and so on. In the first part of the talk, I would like to point out some of these issues that are critical to the success of biometric technology in practice, and investigate the role of Statistics as a field in resolving some of these concerns. Although biometrics research is advancing today quite rapidly, it is primarily concentrated within the computer science/engineering community. So the need arises to involve more statistical expertise by educating the bigger statistical community of the challenges, opportunities and needs for addressing many of the big issues relating to biometric authentication.

In the second part of the talk, I will introduce a statistical framework based on Hierarchical Random Effects model to address the `scalability` issue and demonstrate how it can be applied to evaluate performance of a biometric authentication technique in terms of assessing its predictive ability on future large-scale databases consisting of thousands of images with various distortions such as differential lighting conditions, partial occlusions, expressions, etc.

Robust ER-fMRI Designs under a Nonlinear Model

Dibyen Majumdar
University of Illinois at Chicago, Urbana, IL, USA
dibyen@uic.edu

Co-authors: Ming-Hung Kao, Abhyuday Mandal and John Stufken

Most studies on Event-related functional magnetic resonance imaging (ER-fMRI) experimental designs are based on linear models, in which a known shape of the hemodynamic response function (HRF) is assumed. However, the HRF shape is usually uncertain at the design stage. To address this issue, we consider a nonlinear model to accommodate a wide spectrum of feasible HRF shapes, and propose an approach for obtaining maximin efficient designs. Our approach involves a reduction in the parameter space and an efficient search algorithm. We also propose
an efficient way to evaluate the obtained design. The design that we obtain are compared with those widely used by researchers.

**Statistical Planning and Inference in the Data-Centric World:**

**Applications and Challenges**

Subir Ghosh  
University of California, Riverside, USA  
subir.ghosh@ucr.edu

In the age of BIG DATA, the role of statistical planning/design of experiments will be examined in this presentation. With the amazing computational power and technological advances, complex experiments like High-Throughput Quantitative Assays, Multiplex Assays, Viral Assays for Understanding the Emergence of Viral Infections and Enhancing Rapid Detection of Viruses, The Santa Cruz Experiment (: Can a City’s Crime Be Predicted and Prevented?) are performed. New statistical, computational, and mathematical challenges are emerging from such complex experiments.

We present a new approach to factor rotation for functional data. This rotation is achieved by rotating the functional principal components towards a pre-defined space of periodic functions designed to decompose the total variation into components that are nearly periodic and nearly-aperiodic with a pre-defined period. We show that the factor rotation can be obtained by calculation of canonical correlations between appropriate spaces which makes the methodology computationally efficient. Moreover we demonstrate that our proposed rotations provide stable and interpretable results in the presence of highly complex covariance. This work is motivated by the goal of finding interpretable sources of variability in vegetation index obtained from remote sensing instruments and we demonstrate our methodology through an application of factor rotation of this data.

**Constrained Multi-objective Designs for Functional MRI Experiments via a Modified NSGA-II**

Abhyuday Mandal  
University of Georgia, Athens, GA, USA  
amandal@stat.uga.edu

*Co-authors:* Ming-Hung Kao and John Stufken

Functional magnetic resonance imaging (fMRI) is an advanced technology for studying brain functions. Due to the complexity and high cost of fMRI experiments, high quality multi-objective (MO) fMRI designs are in great demand; they help to render precise statistical inference, and are keys to the success of fMRI experiments. Here, we propose an efficient
approach for obtaining MO fMRI designs. In contrast to existing methods, the proposed approach does not require users to specify weights for the different objectives, and can easily handle constraints to fulfil customized requirements. Moreover, the underlying statistical models that we consider are more general. We can thus obtain designs for cases where brief, long or varying stimulus durations are utilized. The usefulness of our approach is illustrated using various experimental settings. We also show the importance of taking the stimulus duration into account at the design stage.

**Bayesian Learning of Material Density Structure and Blurring Function by Inverting Images taken with Electron Microscopy Techniques**

Dalia Chakrabarty
University of Warwick, Coventry, UK
D.Chakrabarty@warwick.ac.uk

The paper will present a new non-destructive and non-invasive, Bayesian inverse methodology that is developed to learn the unknown density function \( \rho : \mathbb{R}^3 \rightarrow \mathbb{R} \) and the unknown blurring function \( \eta : \mathbb{R} \rightarrow \mathbb{R} \), given image data \( I \subseteq \mathbb{R}^2 \). The novelty of the advanced deconvolution methodology lies in its ability to perform the estimation of the unknown functions via multiple, successive inversions of the data, given that the measured radiation in a pixel is the projection of the convolution of \( \rho(x) \) and \( \eta(z) \), followed by its spatial-averaging over a stipulated volume in \( \mathbb{R}^3 \) inside the system, i.e. \( \mathbb{P}(\rho \circ \eta) \). Here \( X \in \mathbb{R}^3 \), is the three dimensional spatial vector and the third component of \( X \) is \( Z \in \mathbb{R} \).

The methodology has been developed to be applied to the non-parametric reconstruction of the non-linear, non-convex, heterogeneous, sparse/dense, multi-modal material density of material samples, typified by substructure over multiple length scales, imaged by bulk microscopy techniques such as Electron Probe Microscopy and Scanning Electron Microscopy, along with non-parametric or semi-parametric reconstruction of \( \eta(z) \). The complexity of the models varies with the resolution of the imaging technique - \( \sim 1 \mu m \) for Electron Probe Microscopy to \( \sim \)few nm for Scanning Electron Microscopy. This influences the number of integrals (\( \geq 2 \)) over the convolution of the unknowns that are required, in order to result in the image data in a given 2-D pixel.

To render the problem well-conditioned, we expand the data space by invoking multiple images at distinct beam energies while invoking geometric priors on \( \rho(x) \) and strong priors on \( \eta(z) \) using information available in existing microscopy literature. For the \( k \)th beam energy, \( k = 1, \ldots, N_{\text{eng}}, N_{\text{eng}} \in \mathbb{Z}_+ \) and the \( i \)th of the stipulated volumes inside the material – over which the spatial averaging is performed - (\( i = 1, \ldots, N_{\text{data}} \in \mathbb{Z}_+ \)), the contribution of the relevant voxels to \( \mathbb{P}(\rho \circ \eta) \) is identified, and in our fully discretised model, the likelihood is defined as a function of the distance \( D^{(i)} \in \mathbb{R} \) between this contribution and the image data in the
corresponding pixel. The posterior of the unknowns is shown, in the small noise limit, to be a product of Dirac measures centred on solutions of $D_i^{(k)} = 0$, implying uniqueness of our estimate of the product of the unknowns. The uniqueness of the estimates of $\rho(x)$ and $\eta(z)$, and quantification of the uncertainties in the same is discussed and examined under varying strengths of the priors on $\eta(z)$, in a series of inversions of simulated image data. The posterior is sampled from using adaptive Metropolis-Hastings. Applications to real data will also be discussed.

**The Discreet Charm of Convex Optimization: From Shape-Constrained Density Estimation to Empirical Bayes Decision Rules**

Ivan Mizera  
University of Alberta, Canada  
ivanmizera@gmail.com

*Co-author: Roger Koenker*

It is believed - more among data analysts with computational background than among classical statisticians - that maintaining convexity of the optimization problems involved in proposed statistical algorithms saves a lot of trouble and energy, and generally gets the job done. While we neither intend nor dare to proclaim this as a universal dogma, we will try to illustrate, on two parables taken from our research, that there may be something to this thesis. In particular, we discuss our take on shape-constrained density estimation and its outreach into empirical Bayes compound decision rules - modern interior-point methods of convex optimization being a sort of underlying theme in the given context.

**The Use of Topographic Mapping in Visualisation and Classification: Application to Finding Planets around Other Stars**

Somak Raychaudhury  
University of Birmingham, UK  
somak@star.sr.bham.ac.uk

*Co-authors: Peter Tino and Nikolaos Gianniotis*

We present a probabilistic generative approach for constructing topographic maps of high-dimensional multivariate astronomical datasets. This formalism defines a low-dimensional manifold of models, resulting from a smooth non-linear mapping from a low-dimensional latent
space to the space of probabilistic models of the observed data. The use of a training set helps to project each vector to a point in the latent space, obtained from the mean of posterior probabilities over the local noise models, given the data. The smooth nature of the mapping between the latent space and the model manifold enables us to analytically calculate magnification factors that reveal local contractions or expansions in the projections, resulting from the non-linearity in the topographic mapping. As an application of this general approach, we apply it to samples of artificially generated and real light curves of eclipsing binary stars, and show that this is an efficient way of automatic searching for transiting extra solar planets in large datasets.

**Bootstrapping for Significance in Clustering**

**Ranjan Maitra**  
Iowa State University, Ames, IA, USA  
maitra@iastate.edu

A bootstrap approach is proposed for assessing significance in the clustering of multi-dimensional datasets. The developed procedure compares two models and declares the more complicated model a better candidate if there is significant evidence in its favour. The performance of the procedure is illustrated on two well-known classification datasets and comprehensively evaluated in terms of its ability to estimate the number of components via extensive simulation studies, with excellent results. Finally, the methodology is applied to the problem of \( k \)-means colour quantization of several standard images in the literature, and demonstrated to be a viable approach for determining the minimal and optimal numbers of colours needed to display an image without significant loss in resolution.

**Modeling Climate Characteristics Using Small Area Methodology**

**Snigdhansu Chatterjee**  
University of Minnesota, Minneapolis, MN, USA  
chatterjee@stat.umn.edu

Data on climatological variables like temperature, precipitation, sea-level pressure anomalies and concentration of various greenhouse gases are useful in understanding the physical processes governing the climate of this planet, and the various natural and anthropomorphic forcings that act upon such physical processes. Climate datasets are high dimensional, dependent across time, location and other co-variables, and the physical regimes for which data is available could be unstable and non-stationary, and may have one or more change points or regime shifts. We discuss some statistical techniques that might be useful for understanding and interpreting such data. In particular, we present some details on a small area model that has several characteristics that could be useful in understanding and interpreting climate
characteristics. Examples from our ongoing research on both small area techniques and climate statistics are presented.

**Functional Factor Analysis for Periodic Remote Sensing Data**

**Surajit Ray**
Boston University, Boston, MA, USA
sray@math.bu.edu

*Co-authors:* Chong Liu, Giles Hooker and Mark Friedl

We present a new approach to factor rotation for functional data. This rotation is achieved by rotating the functional principal components towards a pre-defined space of periodic functions designed to decompose the total variation into components that are nearly periodic and nearly-aperiodic with a pre-defined period. We show that the factor rotation can be obtained by calculation of canonical correlations between appropriate spaces which makes the methodology computationally efficient. Moreover we demonstrate that our proposed rotations provide stable and interpretable results in the presence of highly complex covariance. This work is motivated by the goal of finding interpretable sources of variability in vegetation index obtained from remote sensing instruments and we demonstrate our methodology through an application of factor rotation of this data.

**Recombination Detection and its Use in Inferring the Recent Human Evolutionary History**

**Jaume Bertranpetit**
IBE, Institute of Evolutionary Biology, Spain

*Co-authors:* Marta Mele, Marc Pybus, Asif Javed, Laxmi Parida and Francesc Calafell

Recombination is one of the main forces shaping genome diversity, but the information it generates is often overlooked. A recombination event creates a junction between two parental sequences that may be transmitted to the subsequent generations. Just like mutations, these junctions carry evidence of the shared past of the sequences and could potentially be used as genetic markers. The algorithm uses a model-based approach to detecting recombinations in a set of haplotypes. The underlying combinatorial algorithm, uses the parsimony principle. We have developed a program, IriS, which detects past recombination events from extant
sequences and specifies the breakpoint location and which are the recombinants sequences. We have validated and calibrated IRiS for the human genome using coalescent simulations replicating standard human demographic history and a variable recombination rate model, and we have fine-tuned IRiS parameters to simultaneously optimize for false discovery rate, sensitivity, and accuracy in placing the recombination events in the sequence. Thus in the same way that SNP data along a chromosome constitutes a haplotype, recombination data makes a rectype that can be inferred for specific chromosomes and pooled for populations, describing in a new approach the recombinational landscape.

Then, we applied IRiS to data collected over 30 populations in the Old World corresponding to 1240 males which were genotyped in 1250 SNPs on the X chromosome with the aim of using the patterns of recombination to make inferences on the history of these populations. Specifically, based on the number of recombinations detected in each population we were able to infer their effective population size. We have found that Sub-Saharan African populations have an $N_e$ that is approximately 4 times greater than those of non-African populations and that outside of Africa, South Asian populations had the largest $N_e$. We also observe that the patterns of recombinational diversity of these populations correlate with distance out of Africa if that distance is measured along a path crossing South Arabia. No such correlation is found through a Sinai route, suggesting that anatomically modern humans first left Africa through the Bab-el-Mandeb strait rather than through present Egypt.

The present work opens a new paradigm in recombination studies as individual events may be detected through a computational approach and may be used in very different genetic and genomic approaches.

**High Resolution Mapping of DNA Methylation**

Karin S. Dorman  
Iowa State University, Ames, IA, USA  
kdorman@iastate.edu

*Co-authors: X. Yin and X. Wang*

DNA methylation is a critical epigenetic modification that is actively involved in the regulation of gene expression. Aberrant methylation plays a role in several human diseases, including cancer, and can be associated with environmental stresses, such as cold exposure in maize. Methylated DNA immunoprecipitation hybridized to microarrays (MeDIP-chip) provides high-throughput methylation profiling at a genome-wide or whole-genome level, but requires extensive follow up computational and statistical analysis for accurate and high-resolution estimation of methylation levels. We propose a Hidden Markov Model to analyze MeDIP-chip data, which
incorporates genomic information of CpG site locations to draw inferences on the hidden CpG methylation states. To handle the computational challenge of CpG-dense genomic regions, such as CpG islands, we group nearby CpG sites into methylatable "components," whose combined methylation state is estimated. Our simulation results suggest the model is reasonably efficient at deconvoluting the DNA methylation profile from MeDIP-chip output, and can achieve much greater resolution than standard methods. We find that standard MeDIP protocols may not use the technology to best advantage; our simulation results suggest non-standard experimental conditions would optimize resolution. Finally, we analyze MeDIP data from a portion of maize chromosome 9 to demonstrate our method on real data.

Pathway Association Tests Using Reduced Gene Subsets for Genome-wide Association Studies

Anbupalam Thalamuthu
Genome Institute of Singapore, Singapore
anbupalamt@gis.a-star.edu.sg

Co-authors: Jingyuan Zhao, Simone Gupta, Mark Seielstad and Jianjun Liu

Single Nucleotide Polymorphism (SNP) analysis only captures a small proportion of associated genetic variants in the Genome-Wide Association Studies (GWAS) partly due to small marginal effects. The pathway-based approaches are alternative ways to analyze the GWAS on complex diseases and hold promise to reveal the mechanisms of complex diseases.

We propose a pathway analysis based on genes, instead of SNPs, as the basic unit, which is expected to increase effect sizes of susceptible units and reduce the degrees of freedom of the joint test statistic. In any particular sample, only a subset of genes within a pathway is expected to be associated, so the joint analysis of all genes will result in a loss of power. To increase the power, we carry out the pathway-based analysis through jointly testing a reduced gene subset, which is selected by using the Least Absolute Shrinkage and Selection Operator (LASSO) penalized logistic regression combined with model selection. The proposed approach can also be used to investigate the joint effect of several genes in a candidate gene study.

Simulation with variable spectrum of number of genes and effect sizes demonstrate that the proposed approach enjoys a higher power than several approaches to detect associated pathways. Also we applied the proposed method to a published GWAS of psoriasis and identify 6 biologically plausible pathways from the whole genome after adjustment for multiple testing. Majority of the pathways reported here for psoriasis are consistent with previous studies. Also a number of important genes for psoriasis are selected by the proposed pathway approach. This approach is easily extendable to meta-pathway analysis involving multiple sample cohorts and data integration of several related data sources.
An Investigation into Market Evolution and Consumer Dynamics in an Online Art Market: A Case of Modern Indian Art

Mayukh Dass
Texas Tech University, Lubbock, TX, USA
mayukh.dass@ttu.edu

We investigate how a market evolves from its inception through its growth, maturity and decline phases. Specifically, we look at the online auction market of modern Indian art, since its inception in December 2000. Using market data from 2000-2008, we identified the various stages of market evolution through a novel method and investigated how the consumers’ (bidders) activity varied as the market evolved. We found the bidding activity as measured by the numbers of bidders, bids, bids per bidder, proxy bids, and the speed with which the pre-auction estimates were crossed to differ across the market cycle. We then examined how the effects of various drivers of price formation during auctions, including products (type and medium of art), brands (emerging or established artists), and market characteristics, vary with market evolution. Results show that brand characteristics play a significant effect on price dynamics, with established artists influential at the growth stage and emerging artists at the decline stage. Available market information (pre-auction estimates) had a significant impact on price formation across all phases of market evolution.

Scale Equivariant Estimators for Diffusion Models

L. Ramprasath
Institute for Financial Management and Research, Chennai, India
rprasath@ifmr.ac.in

This paper considers the definition and calculation of scale equivariant estimators of parameters in diffusion models, based on discretely sampled data. These estimators, for example, ensure that a single time series of asset prices, expressed in different units of currency do not result instructurally different models. We present examples of some popular estimators, proposed for various financial and biological applications, which are not equivariant according to this criterion. In some cases, even standard estimators like the MLE are not scale equivariant and this in turn leads to easy manipulation of key parameter estimates. We finally present a class of estimators based on optimal estimating functions, and prove that members of this class are always equivariant. The results are illustrated using the Gompertzian diffusion model for tumor growth, which is same as the Black-Karasinski interest rate model in finance. The examples mentioned here force us to rethink about the estimators we typically use, for estimating the parameters of a diffusion model.
Correlations in Financial Time-series: An Econophysicist's Perspective

Anirban Chakraborti
Ecole Centrale, Paris, France
anirban.chakraborti@ecp.fr

We present a perspective on correlation studies in the price time-series of stocks, using (i) the minimum spanning tree approach and (ii) the multidimensional scaling tools. The first is a method for finding a hierarchical arrangement of stocks by using correlations of asset returns. With an appropriate metric, based on the correlation matrix, a fully connected graph is defined in which the nodes are companies, or stocks, and the “distances” between them are obtained from the corresponding correlation coefficients. The minimum spanning tree is generated from the graph by selecting the most important correlations and it is used to identify clusters of companies. The second is a set of data analysis techniques that display the structure of distance-like data as a geometrical picture, where again the “distances” between them are obtained from the corresponding correlation coefficients. We then use clustering algorithms to identify clusters of companies. Our general aim is to study the dynamical evolution of the correlations in the market. The topics will be discussed in a pedagogical manner, with illustrations of how the complex economic system (financial market) enriches the list of existing dynamical systems that physicist have been studying for long.

Logic Trees for Multilevel Data: Application to Cancer Care

Mousumi Banerjee
University of Michigan, Ann Arbor, MI, USA
mousumib@umich.edu

Logic regression is an adaptive classification and regression procedure primarily aimed towards interaction detection. Given a set of binary predictors X1, X2, ..., Xk, logic regression searches for Boolean (logical) combinations of the original predictors to create new, better predictors of outcome. For binary outcomes, the goal is to try to fit regression models of the form \( \text{logit}[P(Y=1)] = b_0 + b_1 L_1 + ... + b_p L_p \), where \( P(Y=1) \) is the probability that the binary outcome is 1, and \( L_j \) is any Boolean expression of the predictors, such as "if \( X_1, X_2, X_3 \), and \( X_4 \) are true," or "\( X_5 \) or \( X_6 \) but not \( X_7 \) are true". The \( L_j \) and \( b_j \) are estimated simultaneously using a stochastic optimization algorithm. Boolean combinations of the covariates, called logic trees, are represented graphically as a set of and-or rules. In this paper we consider an extension of logic regression to multilevel data. We apply our methodology to the linked Surveillance, Epidemiology, and End Results-Medicare (SEER-Medicare) registry to model kidney cancer
treatment patterns in the population. By specifically taking into account the clustering of patients within physician practices and hospital environments, the logic regression approach allows us to identify interactions between patient, physician, and hospital factors associated with the adoption of various treatment strategies. Results from our study can inform efforts towards dismantling barriers to quality cancer care in the population.

Predictive Problems in High-dimensional Data: Some Examples

Madhuchhanda Bhattacharjee
University of Pune, Pune, India

An emerging and important area of research is to relate high-dimensional genetic or genomic data to various (clinical) phenotypes of patients. For dichotomous/categorical responses this is typically known as classification problem and has been looked into by experts of many different fields, with statistics playing only a minor role therein. Often the underlying response or phenotype is continuous and these are recoded into categories for ease of interpretation. However many situations require dealing with the actual phenotype in its continuous form and we review certain aspects of modeling and prediction of such response variables. Along with the use of whole repertoire of regression models, several applications of survival models have also been proposed. As is seen elsewhere as well, predominantly models are formed in classical statistical framework compared to Bayesian setup. We investigate applications of some these models along with new (Bayesian) models proposed by us with applications to real data. Apart from the regular issues that are discussed when the data has large number of predictors compared to number of samples, there are aspects that are easily overlooked pertaining to prediction in such situations. Since the objective is to develop an accurate predictor which can be adopted in practice (in real life) it becomes essential to identify proper method of validation (and needless to say carry out validation accordingly). Often mistakes occur, firstly because of lack of understanding of difference between model fitting and deriving a predictive model, secondly inadequate validation system. These render such a model either useless or disappointing in real life prediction. We conclude with some discussion on this topic including few of our observations on this matter.
Probability Machines and Risk Modeling

Abhijit Dasgupta
National Institute of Arthritis and Musculoskeletal and Skin Disorders
and Center for Information Technology
National Institutes of Health, Bethesda, MD, USA

Co-author: James Malley

Machine learning tools are increasingly being used for data mining, data modeling and predictive analysis in biomedical fields, alongside many other fields. The recent literature has shown some ML tools to have nice statistical properties like consistency. We explore some ML tools like Random Forest and k-nearest neighbors, which have been shown very broadly to be consistent, for the purpose of probability estimation in a regression context. Probability estimation is central to biomedical research, for predicting risk of incidence of various diseases, mortality and morbidities. Probability regression directly applies to assessing the probability of outcomes of interest, or changes in such probabilities, based on different patient characteristics and interventions. It can also be used for computing propensity scores for adjustment and matching in observational studies. Currently these questions are typically answered using logistic regression and related methods. The ML-based methods we describe allow the probability predictions to occur in the natural scale, and the nature of these algorithms automatically ensures boundedness of the estimates without the imposition of explicit constraints. Additionally, these methods provide robust nonparametric modeling of the regression function with minimal assumptions about the form of the relationships. Thus this class of methods potentially forms a turnkey solution to risk estimation and risk modeling with minimal assumptions about functional forms of the relationships between outcomes and predictors. We will describe some examples from the literature as well as from our own practice.
Contributed papers
In a typical case-control study, exposure information is collected at a single time-point for the cases and controls. However, case-control studies are often embedded in existing cohort studies containing a wealth of longitudinal exposure history on the participants. Recent medical studies have indicated that incorporating past exposure history, or a constructed summary measure of cumulative exposure derived from the past exposure history, when available, may lead to more precise and clinically meaningful estimates of the disease risk. In this paper, we propose a flexible Bayesian semiparametric approach to model the longitudinal exposure profiles of the cases and controls and then use measures of cumulative exposure based on a weighted integral of this trajectory in the final disease risk model. The estimation is done via a joint likelihood. In the construction of the cumulative exposure summary, we introduce an influence function, a smooth function of time to characterize the association pattern of the exposure profile on the disease status with different time windows potentially having differential influence/weights. This enables us to analyze how the present disease status of a subject is influenced by his/her past exposure history conditional on the current ones. The joint likelihood formulation allows us to properly account for uncertainties associated with both stages of the estimation process in an integrated manner. Analysis is carried out in a hierarchical Bayesian framework using Reversible jump Markov chain Monte Carlo (RJMCMC) algorithms. The proposed methodology is motivated by, and applied to a case-control study of prostate cancer where longitudinal biomarker information is available for the cases and controls.
Bayesian Estimation of Reliability Measures for a Family of Lifetime Distributions Using Masked System Lifetime Data

Sanjeev K. Tomer
Banaras Hindu University, Varanasi, India
sktomer73@gmail.com

Co-author: Ashok K. Singh

We consider the Bayesian estimation of reliability measures of individual components using the lifetime data obtained from multi-component series systems. The lifetime of each component of the system is assumed to follow a ‘family of lifetime distributions’. The observed data is suppose to have time to failure of systems as well as their cause of failure in which the cause of failure may be masked for some of the observations. Using such data, we obtain Bayes estimates of component reliabilities and hazard rates under competing risk model. We present a simulation study and perform computations using Gibbs sampler. We also give an ad-hoc technique to generate sample observations from a complicated posterior distribution.

Diffusive Limits for Adaptive MCMC

Arunangshu Biswas
Presidency College, Kolkata, India
arunb12002@gmail.com

Co-author: Gopal K. Basak

Adaptive Markov Chain Monte Carlo (AMCMC) are a class of algorithms that have been recently proposed in MCMC literature. The main advantage of AMCMC is that, the tuning parameter, which determines how fast the simulation converges to the desired distribution $\psi(\cdot)$, is a function of the previous sample paths. This destroys the Markovian character of the chain. However it can be shown that, under some conditions that the adaptive chain converges to the target distribution $\psi$. In this paper we use a diffusion approximation technique on a discrete time AMCMC. The resulting diffusion gives some idea of the dynamics of the chain. Although this is a degenerate one, it satisfies Hormander's hypoellipticity condition and hence it has positive density on its support. Next we show it has a unique invariant distribution whose marginal distribution is $\psi$. Finally we perform some simulation to compare the Adaptive with the Non-Adaptive or Standard MCMC.
Using Centered and Non-Centered Algorithms for Simulating S&P Share Index Data

Mahsa Ashouri
Isfahan University of Technology, Iran
ashouri.mahsa@gmail.com

In this article, centered and non-centered algorithms are used for simulating S&P share index data and these two methods are compared by drawing estimate of autocorrelation function. By this plot, we can identify which algorithm is better for this simulation. Centered and non-centered algorithms are two expand algorithms of Markov chain Monte Carlo (MCMC) methods for Bayesian inference for volatility processes. In these two algorithms two different ways of parameterizing are used.

Statistical Methods to Identify Clonal Variations Present in Tumour

Sayar Karmakar
Indian Statistical Institute, Kolkata, India
sayarkarmakar@gmail.com

Co-authors: M. Mukherjee, A. Basu and A. Ghosh

Cancer is an important cause of death worldwide. It is a class of diseases characterised by uncontrolled growth of abnormal cells, which arise through mutation of DNA sequences in some parts of human body. There exist some driver mutations of DNA sequences which, we assume, arise early in the original cancer cells providing it a selective advantage to form distinct groups of identical cells with common parent, called clones. The main objective of this study is to identify the driver mutations and determine the number of clones present in the tumour.

Recently developed sequencing technologies enable us to examine a large proportion of the genome-sequence. Each locus \(j\) is probed a random number of times \(n_j\), called the depth at that locus, and one can also count the number of times \(x_j\) alternation to ‘some’ reference is observed. Note that a cancerous tumour is a collection of some normal cells and groups of almost identically mutated cells called clones. If the genome sequence of a normal cell is considered as the reference, we can get an alteration at locus \(j\) when a clonal cell with a mutation at that locus is investigated. If we assume that a particular locus can be affected in at most one clone, which is a fairly reasonable assumption in practice, \(x_j\) can be viewed as a
binomial random variable with parameters $n_i$ and $p_i$, where $p_i$ is the proportion of that type of clonal cells in the tumour. Therefore, in order to find the number of clones, one has to identify how many different values of $p_i$’s are there. One option is to estimate these $p_i$’s by their sample analogs $x_i/n_i$ and then perform cluster analysis. The popular k-means algorithm can be used for this purpose, where $k$, the number of clones, can be chosen using the gap statistic. However, this method does not take the variation in the values of the $n_i$’s into consideration. So, instead of using this method, in this article, we adopted a likelihood based approach, and chose the value of $k$ using the Bayesian information criterion (BIC). We used the idea of hierarchical clustering to generate a nested sequence of models indexed by different values of $k$ and chose the one that minimizes BIC. We used several simulated and real data sets to compare the performance of this estimator with that based on gap statistic and established its consistency under appropriate regularity conditions.

Situations get more complicated when some of the loci are affected by more than one clone. In this case, the number of clusters obtained by the earlier method overestimates the actual number of clones, and it needs to be modified. We proposed a novel method for such modification, and the resulting estimator performed well on simulated and real data sets. This analysis is expected to provide a better understanding of the mechanism of the disease. It explains specific pathways which would provide distinct selection advantage to the tumour cells. Identification of these pathways might lead to better therapeutics for the disease.

**Admixture Time Estimation Using Random Genetic Drift Model**

**Shrijita Bhattacharya**  
Indian Statistical Institute, Kolkata, India  
shrijitab@gmail.com

*Co-authors:* A. Basu and A. Ghosh

Whereas the vast majority of DNA sequence–level variation evolves neutrally and is maintained by a balance between the mutation process and random drift, some parts of the human genome have been subjected to natural selection. Admixed populations are unique in the sense that they represent the sudden confluence of geographically diverged genomes with novel environmental challenges. Thus, genomes from each ancestral population are presented with new challenges. This kind of selection pressure may be quite different from that faced by stationary populations, in which the local environmental changes may occur gradually, allowing for rare advantageous alleles to increase in frequency. Admixed populations therefore offer special opportunities for studying recent selection.
The objective of this study is to device a method for estimating the time elapsed since the admixture event; from the allele frequency distribution of present generation. M. Kimura (1955, *Proceedings of the National Academy of Sciences; volume 41, pages144–150*) derived the exact distribution of allele frequency at a marker locus under random genetic drift with a continuous model. Using the moments of that distribution, we have developed a novel technique of determining the time of admixture based on genotype data comprising both independent and correlated marker sets. When proportion of admixture is known, this technique can be used in a straightforward manner. But, it is usually not known, so one needs to estimate it from the data. We have used a method based on iterative reweighted least squares for this estimation.

We carried out extensive simulation using the Wright-Fisher Model (*random mating, non-overlapping generations*) to study the performance of our proposed estimator for widely varying choices of time of admixture (T), proportion of admixture (α), Fst vales (F) (which measures the separation between two ancestral populations; see S. Wright (1969) *Evolution and the Genetics of Populations Volume 2: the Theory of Gene Frequencies, pages 294–295*), size of the admixed population (N) and number of markers (M). For varying choices of T, α and F, there was no visible difference in the performance of the estimator, but its performance improved significantly with increasing values of N. Increase in M attributed to a remarkable improvement in the efficiency of the estimator, and the improvement was much more than that obtained by increasing the value of N at the same rate.

Determining the time of admixture becomes even more challenging when some of the markers have been subjected to natural selection. In such cases, one option is to identify the markers under natural selection and carry out the analysis using the remaining ones. But identification of these markers is itself a challenging problem. Here, we have proposed a novel method for identifying these markers, and having identified these markers under selection, we have proposed a revised estimator for the time of admixture.

We carried out simulation studies considering a certain proportion (p) of markers under natural selection. The performance of our estimator was studied for different choices of p and also for different choices of associated survival probabilities. The proposed estimator worked well even for moderately large values of M and N. Interestingly, when there was no marker under natural selection, it could nearly match the performance of the estimator discussed earlier.
Estimating Relationship between Random Individuals from their Genome Sequence

Soudeep Deb
Indian Statistical Institute, Kolkata, India
soudeep.deb@gmail.com

The motive of this study is to find some suitable method to predict the relationship between two random individuals when we have their genome sequence data. Two individuals are genetically related at a given locus (i.e. identical by descent or IBD) if they share alleles that are inherited from a common ancestor. Two individuals can also share the same polymorphic marker identical-by-state (IBS), in which alleles at a locus are shared between individuals but are not necessarily inherited from a common ancestor. IBS and IBD methods have been used for linkage and association studies, and to define the relatedness between individuals. Two individuals may share 0, 1 or 2 alleles at a marker. And, on the other hand, we call it IBD 0, 1 or 2 based on how they inherited the alleles. Cotterman defined three coefficients to denote the probabilities of two persons sharing 0, 1 or 2 alleles IBD at a particular marker. IBD probabilities can determine the relationship between two persons.

In this study, we describe two methods to estimate IBD probabilities from the genome sequence data. IBD probabilities depend upon the heterozygosity. At first, we consider the case when we know that the heterozygosity remains constant for all the markers. Next, we consider a general case where heterozygosity follows some distribution known to us and here, we used that distribution to estimate IBD probabilities in a more general case.

We applied all our theoretical methods to a simulated dataset to show that the methods are, more or less, appropriate. Later, the methods were applied to a real-life dataset to find the relationship between individuals. In this study, we also provide some scope of future research related to the topic.

A Deterministic Model for HIV Infection Incorporating Identification Rate in a heterosexual Population: Mathematical Analysis

Manindra Kumar Srivastava
SMS Engineering College, Lucknow, India

Co-author: Purnima Singh

In this paper, we developed a model to study the spread of HIV infection, which can cause Acquired Immunodeficiency Syndrome (AIDS), through Vertical and horizontal transmissions and introduced the concept of identification program by considering identification rate in heterosexual population and qualitatively and numerically analyzed. We obtained equilibrium
points of the model at two states (infection-free and Endemic). We have investigated the criteria for existence of endemic equilibrium point of the model. We determined local and global dynamics of these steady states of the system using stability theory and computer simulation and concluded.

**Genome-wide Association Studies for Bivariate Sparse Longitudinal Data**

*Kiranmoy Das*
Temple University, Philadelphia, PA, USA
tue40440@temple.edu

**Co-authors:** Jiahan Li, Guifang Fu, Zhong Wang and Rongling Wu

Longitudinal measurements with bivariate response have been analyzed by several authors using two separate models for each response. However for most of the biological or medical experiments, the two responses are highly correlated and hence a separate model for each response might not be a desirable way to analyze such data. A single model considering a bivariate response provides a more powerful inference as the correlation between the responses is modeled appropriately. In this article, we propose a dynamic statistical model to detect the genes controlling human blood pressure (systolic and diastolic). By modeling the mean function with orthogonal Legendre polynomials and the covariance matrix with a stationary parametric structure, we incorporate the statistical ideas in functional GWAS to detect the SNPs which have significant control on human blood pressure. Traditional false discovery rate (FDR) is used for multiple comparisons. We analyze the data from Framingham Heart Study to detect such SNPs by appropriately considering gender-gene interaction. We detect 8 SNPs for males and 7 for females which are most significant in controlling blood pressure. The genotype specific mean curves and additive and dominant effects over time are shown for each significant SNP for both genders. Simulation studies are performed to examine the statistical properties of our model. The current model will be extremely useful in detecting genes controlling different traits and diseases for humans or non-human subjects.
The Interactions among Candidate Genes may Lead to Severity of Mental Disorders

Basant K. Tiwary
Pondicherry University, Puducherry, India
basant68@email.com

A disease phenotype is a product of interactions among various genes in a complex network. The molecular complexity of mental disorders can be unveiled using the network medicine approach on the candidate genes implicated in these disorders. There is a considerable overlap in the manifestation of symptoms in some mental disorders such as unipolar disorder, bipolar disorder and schizophrenia. Four candidate genes namely NRG1, DISC1, BDNF and COMT are reported to be associated with mental disorders. I constructed a gene coexpression network of candidate genes and other coexpressing genes based on mutual information approach in healthy condition and under three mental disorders as well using microarray expression data from postmortem brain available at the gene expression omnibus (GEO). There is no alteration in the overall topology of the networks in healthy and mentally ill brains. There is a significant difference in the degree distribution of nodes between normal and bipolar disorder network and bipolar disorder network and schizophrenia network. In contrast, the four candidate genes have similar node properties in normal and mental disorder networks. However, there is a differential direct connectivity among candidate genes in various mental disorders and between normal and mental disorders. All candidate genes are directly connected with each other in schizophrenia except one pair (NRG1-BDNF) indicating a strong role of inter-gene interactions in the manifestation of severe symptoms in this disease. Moreover, there is a change in the status of hub genes in normal network and disease networks. DISC1 and NRG1 are key hub genes in the unipolar disorder network and the bipolar disorder network but have lost the role of hub genes in schizophrenia network despite their significant association with schizophrenia.

Estimating Headrooms of Indian ADRs- An Application of Nonlinear Filtering Technique

Kaushik Bhattacharjee
IBS Hyderabad, IFHE University, India
kabonline09@gmail.com

14 Indian companies have their ADRs listed on the stock exchanges in the US as level III ADRs, out of which 11 are cross-listed. Empirical literature has documented that although the cross-listed ADRs of other countries do not always follow the fundamental ‘Law of One Price’,
resulting in existence of premiums, yet after taking into account the transaction cost, the scope of arbitrage profit is almost nil (Gagnon and Karolyi, 2010). However, the case of Indian ADRs are exceptions and almost a puzzle since almost all of them traded at a very high premium (as measured by price differential of open & closing prices between the two markets after adjusting for exchange rates & DR ratio) which are so much beyond the transaction cost band, so that it can not be arbitrag ed away, posing the issue as a straightforward violation of Law of One Price (Chakrabarty, 2003; Hansda and Ray, 2003). In the realm of Empirical Finance this is a remarkable issue, especially because after the introduction of two-way fungibility in 2002, the ADRs market and underlying Indian market are supposed to be not perfectly segmented and rather be integrated as conversion and re-conversion did start to take place between ADRs and the domestic stocks.

Researchers (Mazumdar, 2007; Bhattacharjee, 2010) have identified that it is the capital control restriction (through a complex mechanism called “Headroom”) which is perhaps responsible for the existence of this prolonged and sustained high level of premiums of Indian ADRs. However “Headroom” data (as a time series) are not available in public domain. Now considering the fact that Arbitrage transaction do happen between these markets and practitioners have identified and documented it long ago (Knauk, 2007) if it is possible to estimate/forecast available “Headroom” from premium time series, it would be significant contribution to the literature as well as from practitioners point of view.

Given this background setting the issue of estimation can be now described as follows: suppose $Y_t = f(X_t, Z_t)$. Here only $Y_t$ is observable but $X_t$ and $Z_t$ are not observable however both are random variables for a given $t$ and hence $\{X_t\}$ and $\{Z_t\}$ can be considered as two unobservable time series. Therefore the problem is to disentangle the effect of $X_t$ and $Z_t$ from $Y_t$. Apparently, there is no standard methodology in finance literature which could be applied in a straightforward manner. Also it is very difficult to get suitable proxies for either premiums (i.e. $X_t$) or new information ($Z_t$).

However, since we can construct an algorithm containing a transition equation linking two consecutive unobservable states of the Headroom’s and already have a measurement equation relating the observed premium to this hidden state, we may proceed in the following manner: firstly we estimate the hidden state a priori by using all the information prior to that time-step. Then using this predicted value together with the new observation, we obtain a conditional a posteriori estimation of the state.

Because of the fact that presence of common nonlinearity in high frequency finance data (BDS test already confirms this for the daily premiums), we propose to employ extended Kalman Filter (EKF) as well as the Unscented Kalman Filter (UKF). We also plan to tackle the issue of Non-Normality by employing the Particle Filtering (PF) algorithm.
In the modern competitive industrial and commercial climate, economic replacement of plant, equipment, machinery can be essential. Replacement of equipment is depending on its proper maintenance. When the maintenance is in proper time it affects the replacement time of equipment. Conventionally, equipment replacement analysis has been mostly done in deterministic and crisp manner. The concept of Fuzzy Multiple Criteria Decision Making (FMCDM) theory may provide a suitable tool for talking this problem.

In this chapter, using Fuzzy MCDM evaluation methodology the most efficient (informative) maintenance approach i.e. strategies, policies, for replacement are purposed. Finally, the proposed evaluation methodology suggested for assessing and selecting the most efficient (informative) maintenance approach is used on typical data from sugar cane industry and conclusions are drawn.

Mammogram Segmentation by Near Set Approach and Mass Lesions Classification with Rough Neural Network

Vibha Bafna Bora
Visvesvaraya National Institute of Technology, Nagpur, India
vibhabora@rediffmail.com

Co-authors: A.G. Kothari and A.G. Keskar

The mammography is the most effective procedure for an early diagnosis of the breast cancer. Mammographic screening has been shown to be effective in reducing mortality rates by 30%–70% . In this paper, a novel algorithm to detect suspicious lesions in mammograms is developed using near set approach. Near set theory provides a method to establish resemblance between objects contained in a disjoint set. Objects that have, in some degree, affinities are considered perceptually near each other, i.e., objects with similar descriptions. The probe functions are defined in terms of digital images such as: gray level, color, texture, contour, spatial orientation, and length of line segments along a bounded region. Objects in visual field are always presented with respect to the selected probe functions. Moreover, it is the probe functions that are used to measure characteristics of visual objects and similarities among perceptual objects, making it possible to determine if two objects are associated with the same
pattern. Thus, the near set approach to automatic image segmentation is based on the partition of an image into equivalence classes with perceptual tolerance relation. The algorithm has been verified on 54 mammograms from the Mammographic Image Analysis Society (MIAS) database. Results of segmentation are compared with Otsu method & Watershed method of segmentation. The segmentation stage is one of the most important since it affects the accuracy of the Feature Extraction & Classification. In the following classification step, feature extraction plays a fundamental role: some features give geometrical information, other ones provide shape parameters & texture. Once the features are computed for each region of interest (ROI), they are used as inputs to a supervised neural network with momentum. The advantage of the neural network is the high accuracy of classification, but it require more time for training compared to other methods. Here the ability of rough sets is utilized to perform dimensionality reduction to eliminate redundant features from a mammogram. Results indicate that rough sets were able to reduce the number of attributes without sacrificing classification accuracy. The experimental results show that the RNN performs better than purely using neural network in terms of time, and it can get classifying accuracy which is higher than using neural network only.

An Application of Data Mining Tools for the Study of Shipping Safety

Aby Subhash
National Institute of Technology, Calicut, India
abysubhash@gmail.com

Co-author: Kelangath S.

Shipping or commercial sea borne transport has always been the backbone of world trade for centuries. According to International Maritime Organization, More than 90% of global trade is carried out by the sea. The increased demand and dependence on shipping has led to some serious flaws in the management of shipping activities which has resulted in both unregulated and substandard employment practices, resulting in negligence of safety and has led to serious accidents over the years. Accidents occurring in ships could be sufficiently severe enough to cause major structural damage, loss of life or property and may cause environmental pollution. Hence most of the research going in this field aims to assess damages and their associated probability levels, and to minimize the consequences of the accidents and suggest practical ways of improving and developing damage resistant designs.

This paper illustrates the different methods available in data mining techniques and case study based on the available data to demonstrate the use and application of such analysis as an aid in decision making especially considering the impact of ship accidents on our environment. Database of accidents from 1980 – 2009 was analyzed to determine the probability of realization of different variables in the model given different evidences. We have used data reduction techniques like Classification Tree Analysis and Principal Component Analysis (PCA).
Classification Tree was used to analyze the significant variables present in the database and the relationship between these variables. PCA helped to reduce the complexity of the data by decreasing the number of variables that need to be considered by identifying the principal components of the data that can be used both to provide a convenient summary of the data and to simplify subsequent analyses. The results of the classification tree analysis and the principal component analysis reported in the present work can be used by decision makers in companies and international organizations to build knowledge-based expert systems and augment their information in the field of safety policy and management.

Univariate Process Capability Indices for Unilateral Tolerances

Moutushi Chatterjee
SQC & OR Unit, Indian Statistical Institute, Kolkata, India
tushi.stats@gmail.com

Co-author: Ashis K. Chakraborty

Development of quality management practices have gradually strengthen the idea that anything produced must have come through a process where some input is converted into some useful output for which there are customers. In order to measure whether such a process is capable to meet the specifications (often fixed by the customer or higher authority) or not, a very popularly known measure of process capability is used, namely, Cpk. This is, probably, due to the fact that it relates the voice of a customer to that of the producer. Technically, a process capability index (PCI) determines, through its mostly single valued measurement, the capability of a process to produce items within the preassigned specification limits viz., ‘Upper Specification Limit (USL)’ and ‘Lower Specification Limit (LSL)’.

Functional Data Analytic Methodology for PBPK Models

Siddhartha Mandal
University of North Carolina
Chapel Hill, NC, USA

Deterministic and stochastic differential equations are commonly used to describe a wide variety of biological and physiological phenomena. For example, they are used in physiologically based pharmacokinetic (PBPK) models for describing absorption, distribution, metabolism and excretion (ADME) of a chemical in animals. Parameters of PBPK models are important for understanding the mechanism of action of a chemical and are often estimated using iterative
non-linear least squares methodology. However, one of the challenges with the existing methodology is that one cannot readily obtain the uncertainty estimates associated with the parameter estimates. Secondly, the existing methodology does not account for variability between and within animals. Using functional data analytic methodology, in this work we develop a general framework for drawing inferences on parameters in models described by a system of differential equations. The proposed methodology takes into account variability between and within animals. The performance of the proposed methodology is evaluated using a simulation study mimicking a real data set and the methodology is illustrated using a data obtained from a benzene inhalation study.

**Early Marriage and the Length of First Birth Interval in India**

Shilpi Tanti  
Banaras Hindu University, Varanasi, India  
shilpi.tanti@gmail.com

*Co-author: K.K. Singh*

Marriage is one of the proximate determinants of fertility. It affects fertility through frequent and regular exposure to sexual relations and the age of entering into marital union. In Indian context, fertility often takes place after marriage and there is an inverse relationship between age at first marriage and fertility. It’s due to the fact that age at first marriage determines the length of exposure to the risk of having pregnant and the actual commencement of the process of childbearing. The relationship between age at first marriage and marital fertility is, however, based on the assumption that contraceptive practices are non-existent. In India marriage is universal and in rural areas marriages occur early, generally before 18 years. The length of the first birth interval is one of the strongest and most persistent factors affecting fertility in noncontracepting populations, with longer intervals usually associated with lower fertility. In this study, an attempt has been made to find the relation between timing of marriage and the length of first birth interval. There are also some other socio-demographic variables which influences the length of first birth interval. The data from NFHS-3 is used to analyse the fact. The study is restricted for lower age-group of females, i.e., females of age 18 years and below. Kaplan Meier’s plot and Cox’s proportional hazards model are applied to analyse the data besides the descriptive statistics.
Fitting a Mixture of Conway-Maxwell-Poisson Distributions to Count Data

Pragya Sur
Indian Statistical Institute, Kolkata, India
pragya1386@gmail.com

Co-authors: Paromita Dubey, Smarajit Bose and Galit Shmueli

The Poisson distribution is the most popular choice for modeling count data. However, nowadays many situations arise in real life where non-Poisson counts are commonly observed, for example, in visits to websites, e-commerce transactions, etc. A major problem is that the observed data shows over-dispersion or under-dispersion. The Conway-Maxwell-Poisson (CMP) distribution is an ideal modification of the Poisson distribution obtained by introducing a new parameter to allow for such over-dispersion or under-dispersion. Even then there are situations where the CMP distribution fails to capture the true nature of the data. Bimodality in the data is one such instance. We show that a mixture of two CMPs successfully captures such structure of the data. In this paper we shall concentrate on methods by which the parameters of the mixture of two CMP distributions may be estimated. We would also present experimental results to evaluate the performance of these methods on some real life data sets.

New Stein’s Estimators of Regression Coefficients in Replicated Measurement Error Model

Sukhbir Singh
Panjab University, Chandigarh, India
sukhbir83@gmail.com

Co-authors: Kanchan Jain and Suresh Sharma

A replicated ultrastructural measurement error (RUME) regression model is considered where both predictor and response variables are observed with error. Three estimators of regression coefficients for this model are available in literature (bj; j = 1, 2, 3), which are consistent but unbiased only asymptotically [Shalabh, Journal of Multivariate Analysis, 86 (2003) 227-241]. The problem of obtaining consistent Stein’s estimators of regression coefficients which has lower quadratic risk as compared tobj is dealt with. Using minimal mean square error (MMSE) approach to generate Stein’s estimator, new estimator bsj is proposed under general situation where the available estimator bj either under or over estimates the regression coefficient vector β. Conditions are studied under which bsj has lower quadratic risk than bj. It is observed
that when $h_j$ over estimates $\beta$, the new estimator $h_j$ not only has lesser quadratic risk but also has less bias as compared to $h_j$. This is an interesting observation as the Stein’s estimators are known to have larger bias. The asymptotic properties of the proposed estimators are studied without imposing any distributional assumption on any random component of the model. It is observed that the proposed estimators have asymptotic Normal distribution. Monte Carlo simulations study is performed to assess the effect of sample size, replicates and non-normality on the estimators. The methods are illustrated using real economic data.

Copula Based Dependence Structures among Maternal Variables

**Ranjana Gogoi**
Gargaon College, Assam, India
ranjanagogoi@yahoo.co.in

*Co-author: Navajyoti Tamuli*

Birth weight is a reliable index of intrauterine growth and also one of the major factors that determine child survival, future physical growth and mental development. Consequently the study of dependence structures among maternal variables is an important aspect of research in order to establish effective maternal and child health care programmes. The variables which are often found to be important are maternal age, birth weight, maternal weight and parity etc. Many reports from developing countries which established a relationship between such type of demographic variables using only bivariate statistical analysis such as- simple correlation, simple linear regression analysis, student t-test, and one way analysis of variance. Nevertheless, in most cases these measures are not the perfect tool to capture the dependence because they have serious limitations. The best technique to model random variables together is to fit their joint distribution as we can evaluate their marginal behavior and dependence structure. If marginal distributions are non elliptical, then it is a tough task to construct a joint distribution from these marginals. But if we use copula, we can easily construct joint distribution of the variables under study. The present work performs a better selection of marginal distribution functions for maternal characteristics and demonstrates how the concept of copula may be used for establishing joint distribution function with mixed marginal distributions.
Statistics in Industry
And the Two Twains Shall Meet

N. R. Srinivasa Raghavan
Citicorp, India

With the ever increasing penetration of systems and technologies to capture data comes the overwhelming mission of making sense out of this data. Today, most companies capture several peta-bytes of data: mostly internally generated, and some externally acquired. Industry especially needs methods sometimes mundane, sometimes advanced, and sometimes very innovative, to help decision makers respect the diktat:"data driven decisions". In this talk, I will focus on interesting applications of statistics of most genre's that marry the right appetite for modelling with the right approach to market the model! Specific issues pertaining to the manufacturing, service operations, supply chain management, market research and digital analytics will be covered during the presentation. From a modelling perspective, I will cover the methods popular in traditional statistics as well as from machine learning. I also intend to cover the challenges and issues faced by a typical industrial statistical analyst during his/her journey in the applied research domain.

"What's in a Name?"

K Vijay Kumar
Citibank N.A. India Consumer Group

Determining whether two sets of information do, in fact, relate to the same individual is a critical need in many banking applications - from marketing, fraud and AML monitoring to risk assessment. In the absence of a unifying ID, organizations resort to creating their own unique identifiers to tag an individual or entity. In any ID creation process, one key part is the evaluation of phonetic equivalents for name matching purposes. An equally important component is the match-off against a known database of variants, and appropriate weighting of the possibilities to determine matches with the highest degree of confidence. Further checks, such as of address, phone number, email, etc., will often be needed. Many of the algorithms developed for such purposes cater to the demands and circumstances of the western world, and each of the components listed above need to be fine-tuned and customized to Indian needs."What's in a name" discusses the lacunae of existing methodologies in the Indian
context, explores ideas for developing algorithms that could be successfully used in the Indian context, and walk through some sample applications of such algorithms in the Financial Services industry.

Reliability, Availability, and Productivity for Wind Turbines

Abhinanda Sarkar
GE Global Research, India

A wind turbine is an energy-producing device that converts an uncertain amount of mechanical energy into electrical power feeding into grids or other transmission and distribution networks that demand reliable and readily available power. This dichotomy in translating an essentially random input (the wind) into a measurable level of power production that inspires a degree of confidence can be resolved by consolidated analysis of wind speed distribution, power curves of turbines, and their downtime, all working in concert. Risk measures inherited from finance can provide a platform.

Tracking Consumer Segments

Arabinda Ghosh
Zenith Optimedia, India

Identifying & targeting consumers is the basis of marketing for any product category. There are multiple different approaches available to identify the right consumers. For example:
- Demographic way of identification e.g., Age, Gender, Province, Socio Class (using Education, Occupation & other affluence parameters like ownership of durables)
- Psychographic way of identification. Consumers are asked to rate the psychographic attributes in agreement scale. Ratings are then used to combine consumers into different clusters.
- Need based segmentation approach. This approach is more in the context of a product category. The consumers are asked to state their needs in the context of the product category in agreement scale. The rating on the need statements are used to combine consumers into clusters.
- Brand Values segmentation is one of the latest in segmenting consumers. In this approach the statements are bit like need based statements. However, consumers are not asked to give their opinion in terms of their degree of agreement with the statements but they are asked to associate the statements with different brands in the category.

Different approaches are evolved to solve different marketing issues. Brand values segmentation approach proved to have very high correlation with brands that consumers use
which is not the case in other approaches. So it provides good understanding of the drivers of choice for different brands & hence the USPs for different brands, if any.
The brand values segmentation study is fully integrated into the marketing planning model of a multinational company in last decade. While consumer segmentation study is done separately to extract the underlying consumer segments in the market - it became very critical to track the consumer segments over time to understand the performance segments in terms of size & the performance of different brands. The traditional discriminant analysis approach for segment membership prediction was not working as size of the consumer segments were varying a lot between original study & the tracking studies.
Several approaches including discriminant analysis were attempted to solve the issue but none of them yield very consistent solution. Finally a combination of discriminant analysis & logistic regression approach was attempted. Discriminant analysis was used to identify the most discriminating attributes across segments. Logistic regression was used to develop the prediction equation for each segment. The constant term of the Logistic regression equations were calibrated to take care of the design difference between the original segmentation study & the tracking studies. The calibration is done once - but the outcome found to be very consistent across waves of the tracking studies. The approach was implemented successfully across regions in at least 60 countries.

Managing Expectations: Statistics in Risk Management
Karra Sanghamitra
Morgan Stanley International Plc, India
karra.sanghamitra@gmail.com

In financial risk management Value at risk (VaR) is the most widely used statistic for measuring 'risk' in a given portfolio of assets. In this talk, I'll first introduce VaR and the various methods used for computing this metric for a single financial asset. Within this context some relevant questions to look at are (1) Does Normality work? (2) How much or how little data do we need to assess risk? (3) Volatility Clustering and GARCH (4) Quantiles and the use of EVT.

VaR in the context of a portfolio is well defined but computing the statistic for a portfolio creates a number of issues some of which are discussed here. (5) Multi Dimensionality and constraints to dimensionality reduction methods (6) Sub ‐ additive properties of VaR (7) Data sparsity issues.

The financial crisis of 2008 is often dubbed to be a failure of risk management practices across the world. I'll briefly talk about some of the new regulations that are being introduced which resulted in a plethora of empirical research to develop new risk management models to be used in conjunction with VaR.
Posters

Saman Khawaja
Aligarh Muslim University, Aligarh, India
samankhawaja@gmail.com

Co-authors: Shazia Ghufran and M. J. Ahsan

In stratified sampling usually the estimation of several population characteristics is of interest. When more than one characteristics are under study the use of individual optimum allocations is not advisable for one reason or the other. In such situations, some compromise criterion is needed to work out an allocation which is optimum for all characteristics in some sense. In this paper, for estimating $p$-population means the problem of determining the optimum allocation has been studied in presence of non-response. The problem is formulated as a Multiobjective Integer Nonlinear Programming Problem (MINLPP) and a solution procedure is developed using Goal Programming Technique. A numerical example is also worked out to illustrate the computational details.

Multiple Comparisons Procedure for Comparing Several Exponential Populations with Control in Terms of Location Parameters under Heteroscedasticity

Vishal Maurya
Reserve Bank of India, Mumbai, India
vmaurya@rbi.org.in

Co-authors: Amar Nath Gill and Parminder Singh

In this paper, new design-oriented two-stage two-sided simultaneous confidence intervals, for comparing several exponential populations with control population in terms of location parameters under heteroscedasticity, are proposed. When the experimenter has a prior expectation that the location parameter of $k$ exponential populations are not less than the location parameter of control population, one-sided simultaneous confidence intervals provide more inferential sensitivity than two-sided simultaneous confidence intervals. But the two-sided simultaneous confidence intervals have advantages over the one-sided simultaneous
confidence intervals as they provide both lower and upper bounds for the parameters of interest. The proposed design-oriented two-stage two-sided simultaneous confidence intervals provide the benefits of both the two-stage one-sided and two-sided simultaneous confidence intervals of Lam and Ng [Lam, K., Ng, C.K., 1990. Two-stage procedures for comparing several exponential populations with a control when the scale parameters are unknown and unequal. Sequential Analysis 9 (2) 151-164]. When the additional sample at the second stage may not be available due to the experimental budget shortage or other factors in an experiment, one-stage two-sided confidence intervals are proposed, which combine the advantages of one-stage one-sided and two-sided simultaneous confidence intervals proposed by Wu, S.-F., et al.[Shu-Fei Wua, Ying-Po Lin , Yuh-Ru Yu, (2010), One-stage multiple comparisons with the control for exponential location parameters under heteroscedasticity. Computational Statistics & Data Analysis, Volume 54 Issue 5, May, 2010]. The critical constants are obtained using the techniques’ given in [Lam, K., 1987. Subset selection of normal populations under heteroscedasticity. In: Proceedings of the Second International Advanced Seminar/Workshop on Inference Procedures Associated with Statistical Ranking and Selection, Sydney, Australia, August 1987; Lam, K., 1988. An improved two-stage selection procedure. Communications in Statistics - Simulation and Computation 17 (3) 995-1006]. These critical constant are compared with the critical constants obtained by Bonferroni inequality techniques and found that critical constant obtained by Lam (1987, 1989) are less conservative than critical constants computed from Bonferroni inequality technique. Implementation of the proposed simultaneous confidence intervals is demonstrated by a numerical example.

Optimum All Order Neighbor Balanced $\mu$-Resolvable BIB Designs

**Anurup Majumder**

Bidhan Chandra Krishi Viswavidyalaya, India
anurup.majumder@rediffmail.com

**Co-author:** Aatish Kumar Sahu

Keifer and Wynn (1981) considered the effects of correlation on the efficiency of BIB designs. The covariance structure considered by them assume

\[
\text{Cov}(y_{jr}, y_{j'r'}) = \begin{cases} 
\sigma^2, & \text{if } j = j' \text{ and } r = r' \\
\rho \sigma^2, & \text{if } j = j' \text{ and } |r - r'| = 1 \\
0, & \text{otherwise,}
\end{cases}
\]

where, \(y_{jr}\) be the observation of the \(r^{th}\) plot in \(j^{th}\) block, i.e. all observations have the same variance, no correlation exists between blocks, and neighboring plots in the same block have the same amount of correlation.

Let, \(g(j, r)\) = the treatment number of the \(r^{th}\) plot in the \(j^{th}\) block.
$A_i$ = the set of blocks in which treatment $i$ occurs.
$e_{ii'}$ is the number of blocks in $(A_i, A_{i'})$, in which $i$ occurs at an end plus the number
where $i'$ occurs at an end, and $N_{ii'}$ is the number of times $i$ and $i'$ are adjacent in a block.

Kiefer and Wynn (1981) showed that under the above model, a BIB design in which all
The quantities of $e_{ii'} + kN_{ii'}$ (i $\neq i'$) are equal possess strong optimality properties in the set of BIB designs. Since, $kN_{ii'}$ are usually much bigger than the $e_{ii'}$, Kiefer and Wynn (1981) suggested that it would be useful to look for designs with all $N_{ii'}$ equal. They called a BIB design with all $N_{ii'}$ equal as an Equi-Neighbored BIB design (EBIBD).


Kiefer and Wynn (1981) developed symmetric EBIB design with $k = (v + 1)/2$ and minimum $b$, where $v$ is prime. They also developed EBIB design with more than one initial block. Cheng (1983) constructed EBIB designs and some of them have minimum number of blocks, but the value of $k$ is restricted to either 3 or $(v-1)$. Jackroux (1998) constructed a series of EBIB designs with $k = 3$, and some of the designs given by him have smaller number of blocks than Cheng (1983) for given $v$ and $k$.

In the present study, the work of Kiefer and Wynn (1981) has been further extended to $NN^t$ optimum covariance structure, where, $t=1,2,\ldots,k-1$, where $k$ is the number of plots per block. The extension work has done assuming the similar model of Kiefer and Wynn (1981) and employing the least squares estimators. The optimum designs with $NN^t$ structure require the same number of blocks as $NN_1$ or $NN_2$ designs. The designs are simultaneously optimum for all lower order covariance structures i.e. $NN^t'$, $t' < t$.

A list of all ($NN_t$) order neighbor balanced (AONBD) $\mu$-resolvable BIB designs are presented in Table of the paper with $3 \leq k \leq 10$ and $7 \leq v \leq 19$ which are weakly universal optimum and require minimum number of blocks.

Simultaneous Prediction Intervals for Comparing Several Treatments with More than One Control

Anju Goyal
Panjab University, Chandigarh, India
anju.statistics@gmail.com

Co-authors: Sukhbir Singh and Vishal Maurya

In this paper, one-sided and two-sided simultaneous prediction intervals for comparing $k_1$ treatments with $k_2$ controls are proposed. Hung et al. (2002) [Cheung SH, Wu KH and Lim SP (2002) Simultaneous Prediction Intervals for Multiple Comparisons with a Standard, Statistical Papers, 43, 337-347.] gave a procedure for constructing simultaneous prediction intervals for
multiple comparisons with a standard. For \( k2=1 \), this generalization reduces to Hung’s special case. These simultaneous prediction intervals are constructed by using the past random samples from \( k1 + k2 \) normal populations having equal variances. The proposed prediction intervals can be used to predict the differences between the future means of independent random samples from each of \( k1 \) treatment populations compared to \( k2 \) control populations with a predetermined joint probability. Methods on sample size determination are also given. The use of the new procedure is demonstrated with a numerical example.

Some Probability Models for Waiting Time to First Conception

R.C. Yadava
Banaras Hindu University, Varanasi, India

Co-authors: S.K. Singh, Anup Kumar and U. Srivastava

This article is concerned with development of some stochastic models better suited to describe the observed distribution of waiting time to first conception among different groups of females. These groups are formed according to age at marriage, as age at marriage plays crucial role in shaping the distribution of duration of first birth. For different groups, different assumptions for conception rate (analogous to hazard rate in life testing problems) is considered, hence the expression of probability models for different groups have been obtained. Bayes estimates of the parameters involved in the probability models are obtained. These models have been applied to National family Health Survey data, and the estimates of conception rate for different groups also have been obtained.

On Comparing the Performance of Senior and Young Cricketers in Indian Premier League

Hemanta Saikia
Jawaharlal Nehru School of Management, Assam University, Silchar, India
h.saikia456@gmail.com

Co-author: Dibyojyoti Bhattacharjee

Most of the cricketing fraternity opines that Tewnty20 cricket is a game of young cricketers only. This believe got the stump of conviction when Younis Khan, the former captain of Pakistan cricket team, had announced his retirement at the age of 32 from Twenty20 cricket, after team
Pakistan in 2009, won the Twenty20 world-cup in England. While announcing his retirement Khan commented that, he doesn’t consider himself fit enough for this format of the game of cricket. Indeed, it is true that a few young players like P Valthaty, R Ashwin, S Marsh, S Randiv, S Hasan, etc. have managed to drag attention of selectors and spectators towards themselves through their remarkable performance in IPL. But in all the four seasons of IPL played so far, the performances of senior cricketers are noteworthy. Instead, their performances are comparable to that of the youngsters. The senior players are making a huge impact in the IPL largely due to their skills. The performances of skilled players in Twenty20 cricket were a lesson for youngsters that there is no substitute for skill if anyone is to succeed in any format of the game of cricket. Therefore, the present study tries to examine whether the performances of senior and young cricketers in IPL are different. In cricket, since it is difficult to generalize a relation between aging and performance as cricketers with different expertise peak at different ages so the category of cricketers (i.e. senior and young) is determined on the basis of their age. To measure the performance of cricketers a model is developed utilizing the three prime abilities of the game viz. batting, bowling and wicket-keeping. Various cricketing factors are considered related to the performance of cricketers under the above mentioned abilities. All these factors are normalized and accordingly adjusted by using appropriate weights on the basis of their relative importance. The performance measures are obtained for each cricketer separately in all the four seasons of IPL and a comparison is made between the senior and young cricketers using discriminant analysis. As an outcome of the discriminant analysis, the performances of senior and young cricketers in IPL are not found to be significant. Hence, the study concludes that the performance of cricketers in IPL cannot be discriminated on the ground of age.

Almost Unbiased Ratio and Product Type Exponential Estimators

Rohini Yadav
Indian School of Mines, Dhanbad, India
rohiniyadav.ism@gmail.com

Co-authors: Lakshmi N. Upadhyaya, Housila P. Singh and S. Chatterjee

This paper considers the problem of estimating the population mean \( \bar{Y} \) of the study variate \( y \) using information on auxiliary variate \( x \). We have suggested a generalized version of Bahl and Tuteja (1991) estimator and its properties are studied. It is found that asymptotic optimum estimator (AOE) in the proposed generalized version of Bahl and Tuteja (1991) estimator is biased. In some applications, biasedness of an estimator is disadvantageous. So applying the procedure of Singh and Singh (1993), we derived an almost unbiased version of AOE. A numerical illustration is given in the support of the present study.
Bayesian Estimation and Prediction of Generalized Inverted Exponential Distribution

Sanku Dey
St. Anthony’s College, Shillong, India
sanku_dey2k2003@yahoo.co.in

Co-authors: Tanujit Dey and Debasis Kundu

A generalized version of inverted exponential distribution (IED) is introduced in this paper. This lifetime distribution is capable of modeling various shapes of failure rates, and hence various shapes of ageing criteria. The model can be considered as another useful two parameter generalization of the IED. The gamma prior density is assumed for one parameter case, whereas the joint prior density of the two parameter case is composed of the both gamma densities. Under the above priors, we use Gibbs sampling technique to generate samples from the posterior density function. Based on the generated samples, we can compute the Bayes estimates of the unknown parameters and also can construct highest posterior density credible intervals. We also compute the approximate Bayes estimates using Lindley's approximation method. Monte Carlo simulations are performed to compare the performances of the Bayes estimators with the maximum likelihood estimators. One data analysis is performed for illustrative purposes. We further consider two-sample Bayes prediction of the future observation based on the observed sample and also construct a predictive intervals with a given coverage probability.

On the Bayes Estimators of the Parameters of Non-zero Inflated Generalized Power Series Distributions

Peer Bilal Ahmad
Islamic University of Science & Technology, Awantipora, India
peerbilal@yahoo.co.in

In this paper, we study the class of non–zero inflated generalized power series distributions (NZIGPSD) where the inflation occurs at any of the support points. This class includes among others the negative binomial, Poisson and logarithmic series distributions. We derive the Bayes estimators of functions of parameters of the non-zero inflated generalized power series distribution under squared error loss function and weighted square error loss functions when the parameter of inflation is known. The results obtained for non- zero inflated GPSD are then applied to its particular cases like negative binomial, Poisson and logarithmic series distributions. These estimators are better than the classical minimum variance unbiased estimators (MVUE) in the sense that they increase the range of the estimation and also have simpler form. Finally, an example is provided to illustrate the results and a goodness of fit test is done using the maximum likelihood and Bayes estimators.
Relations for Marginal and Joint Moment Generating Functions of Generalized Exponential Distribution Based on Lower Generalized Order Statistics and Characterization

Benazir Zia
Aligarh Muslim University, India
benazir.stats@gmail.com

Co-author: R. U. Khan

In this study we give exact expressions and some recurrence relations for marginal and joint moment generating functions of lower generalized order statistics from generalized exponential distribution. The results for order statistics and lower record values are deduced from the relations derived. Further a characterization of the distribution by considering recurrence relations for marginal moment generating functions of the lower generalized order statistics is presented.

On Exact Nonparametric Alternative Tests in Two-Way Layout

Sainath L. Maidapwad
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India
maidapwadsl.stat07@gmail.com

Co-author: S.L. Sananse

In experimental design, in many situations, we come across with the non-normally distributed data affecting the assumptions for analysis of variance (ANOVA). In the present research paper an attempt has been made to study some of the existing most promising non-parametric test alternatives to the two-way Layout such as Friedman test, Ehrenbergs test, Puri&Sen test and Rank transform tests by Conover & Iman and applied these test on real life data set. Further, an attempt has been made to study the non-parametric tests in case of one and multiple observation per cell, interaction present and absent. We determine that achieving the highest possible asymptotic relative efficiency for existing non-parametric tests. In this study, simulation results are presented for three underlying populations: Gaussian, Uniform and Exponential.
A Test for the Significance of Treatment Effect in a Mixture Population

Kushal Kumar Dey
Indian Statistical Institute, Kolkata, India
kshldey@gmail.com

Co-authors: Bikram Karmakar and Kumaresh Dhara

In case-control experiments, one part of the population is introduced to some medication (case) and the other part is left un-medicated (control). Standard parametric and nonparametric methods can be used to test the effect of the medication. However, if the underlying population has several clusters, which may be created due to one or more unknown factors, most of these standard tests yield unsatisfactory performance. This problem becomes more transparent when these hidden factors (or clusters) have substantially higher effect on response than the medication under consideration, and the proportions of individuals under medication vary widely in different clusters. In such cases, usual parametric and nonparametric two-sample tests often fail to eliminate these cluster effects and hence mislead our inference. In this article, we develop an algorithm to deal with this issue. Our proposed method aims to eliminate the effect of the hidden factors in order to estimate the true effect of medication on the population. Here we first use some efficient clustering algorithms to find the hidden clusters (if any) present in the population. In order to eliminate the differences due to these clusters, we first choose one of these clusters, called pivot cluster, and then project the observations of other clusters on it using a suitable transformation. Standard parametric or nonparametric tests are used on these transformed observations for the case and control. If we consider the response to be normally distributed in each of these clusters, one can use the sample means and sample variance-covariance matrices to determine this transformation. It is computationally very simple and the resulting test procedure does not depend on the choice of the pivot cluster. Instead of sample means and sample variance-covariance matrices, one can use other robust estimators as well. When the distribution of the response variable is not normal in each of these clusters, a transformation based on geometric quantiles can be used. When the clusters are not well separated, it becomes difficult to separate them using any clustering algorithm. In such cases, we use probabilistic framework to overcome this problem. Here we consider the results for different cluster formations and aggregate them using a Bayesian model averaging technique, which takes the posterior probabilities of different cluster configurations into consideration.

We carried out an extensive simulation study involving various types of univariate and multivariate distributions with different number of hidden clusters and varying proportions of observations in those clusters. In many cases, the standard two sample tests failed to detect significant medication effect, and in many other cases, they gave false alarm. But in most of these cases, our proposed method led to correct decision. It had better performance than these existing methods both in terms of level and power properties. We also analyzed some real life data sets to demonstrate the utility of the proposed method.
Valedictory Lecture

Prof. Pranab K. Sen
Cary C. Boshamer Professor in Biostatistics
University of North Carolina, Chapel Hill, NC, USA

The genesis of contemporary issues and applications lies in the past accomplishments and their evolution to suit the future need more effectively. In the Indian subcontinent, we were fortunate to have a blended gestation of Statistics wherein the mathematical foundations were integrated to numerous applied areas. With more new and interdisciplinary fields in evolution, the contemporary issues in Statistics are far more complex and conventional statistical tools are becoming obsolete with a greater frequency. In some novel fields of application, the role and scope of statistical reasoning and rationality are thoroughly appraised with emphasis on the role of Indian Statistical Institute and the Mahalanobis vision of “Unity in Diversity” leading to “Statistics: A Key Technology”.