International Conference on
Multivariate Statistical Methods in the 21st Century:
The Legacy of Prof. S.N. Roy

December 28–29, 2006

Applied Statistics Unit
Applied Statistics Division
Indian Statistical Institute
203 B.T. Road. Kolkata 700 108. India

Cosponsored by

- Institute of Mathematical Statistics, USA;

Collaborators

Calcutta Mathematical Society; Calcutta Statistical Association; Forum for Interdisciplinary Mathematics; Indian Association for Productivity, Quality and Reliability; Indian Bayesian Society; Indian Society for Probability and Statistics; Indian Statistical Association; International Indian Statistical Association; University of California-Riverside.

Abstracts

(* **indicates speaker**)

STATISTICAL CHALLENGES AND LIMITATIONS IN ANALYZING 1H NMR METABONOMICS DATA

M. Abbas*, M. Srivastava, R.K. Sharma and Raja Roy

We know that disease processes, drugs or toxins cause adjustments of the concentrations and fluxes of endogenous metabolites involved in cellular pathways. This metabolic adjustment is expressed as a finger print of biochemical perturbations, which is characteristic of the site of a toxic insults or disease process. Such perturbations can be captured by using a technology called Metabonomics. Metabonomics is usually conducted on biofluids like urine, blood serum or even tissue extracts. High resolution 1H nuclear magnetic resonance (NMR) spectroscopy has proved to be one of the most powerful technologies for biofluids. NMR-based metabonomics can be used to address a wide range of clinical, environmental and toxicological problems. Multivariate data analysis has significantly contributed in the development of metabonomics. Both unsupervised analysis (e.g. searching for patterns in the NMR spectral profile) and supervised analysis (e.g. modelling classes of pathological dysfunction). Methods such as Principal Component Analysis (PCA), Non-Linear Mapping procedures (NLM) and Hierarchical Cluster Analysis (HCA) from unsupervised area, and Soft Independent Modeling of Class Analogy (SIMCA), Partial Least Squares based-Discriminant Analysis (PLS-DA) methods from supervised area have been attempted successfully. Many multivariate algorithms are based on the assumptions that data are normally distributed. In reality, biological data are extremely complex and may not always comply with that assumption. Therefore alternative approaches such as Neural Network (NN) have been found suited to such non-linear classification problems. In the paper, we present theoretical backgrounds and applications of afore-mentioned techniques in metabonomics and discuss few interesting case studies including two of our own.

IMPROVING THE HANSEN-HURWITZ ESTIMATOR IN PPSWR SAMPLING

Arun Kumar Adhikary

An attempt has been made to improve upon the Hansen-Hurwitz (1943) estimator based on PPSWR sampling scheme through Rao-Blackwellisation. In order to derive the sampling

variance of the improved estimator obtained by Rao-Blackwellisation it is of interest to derive the probability distribution of the number of distinct units in the sample drawn according to PPSWR sampling scheme as the improved estimator is based solely on the distinct units in the sample. It has been possible to write down the exact distribution of the number of distinct units in the sample drawn by PPSWR sampling scheme in a closed form.

Assuming a super-population model the model - expected design - variance of the improved estimator is worked out and it is compared with that of the Hansen-Hurwitz estimator under the same super-population model.

The percentage gain in efficiency in using the improved estimator over the Hansen-Hurwitz estimator is worked out for certain selected values of the model parameters.

VARIANCE ESTIMATION FOR MISSING DATA USING MULTI AUXILIARY INFORMATION

Raghunath Arnab

Any large scale survey may be prone to nonresponse problems. No exact formulation of the nature of nonresponse in surveys is available. So, several methods of handling nonresponse problems are proposed by survey statisticians. It is common practice to insert values for non-respondents. Normally imputed values are treated as true values and estimates of the parameters and their variances are obtained by using standard formulas. In this paper, the problems of estimation of the population total and its variance have been studied in the presence of nonresponse using multi auxiliary information. The proposed methodology is based on the assumption that the set of respondents (comprises response sample) is a Poisson sub-sample from the initial sample. The proposed method is more general than the method presented by Särndal (1992) since one can use auxiliary information in estimating the unknown response probabilities. The relative efficiencies of the proposed estimator are compared by simulation studies.

BIVARIATE CIRCULAR DISTRIBUTIONS VIA CONDITIONAL SPECIFICATION

Barry C. Arnold

Conditional specification may be conveniently used to generate flexible classes of bivariate circular distributions, just as it may be used to develop useful models for distributions with support in $\mathbb{R}^2$. Particular attention will be paid to models with conditionals in frequently used families, such as circular normal, cardioid and wrapped Cauchy.

USE OF MULTIVARIATE AUXILIARY INFORMATION THROUGH TWO STAGE DESIGN USING POSTSTRATIFICATION IN ESTIMATING POPULATION RATIO

Shashi Bahl⋆ and Geeta

In the usual procedure of post-stratification in two stage sampling using multi-auxiliary information estimation of population ratio has been discussed. New estimators based on this sampling scheme are proposed. The general properties of proposed estimators are studied. The situations, in which the proposed estimators are better, are also discussed. It has been empirically demonstrated that the suggested scheme not only provides the estimates of characteristics under study but also improves the precision of the estimate.

IDENTIFICATION AND ESTIMATION OF TRUNACATED BIVARIATE NORMAL STOCHASTIC FRONTIER MODEL: A BAYESIAN APPROACH

Debdas Bandyopadhyay and Arabinda Das⋆

Truncated bivariate normal distribution has recently been used to model the joint probability distribution of component errors in stochastic frontier production model. In this paper we show that this model suffers from identification problem and use a Bayesian approach for identification and estimation of the parameters of the model. Monte-Carlo simulations are also provided to compare finite sample properties of the estimation procedure.

RECENT DEVELOPMENTS AND OPEN PROBLEMS IN NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION UNDER SHAPE CONSTRAINTS

Mouli Nath Banerjee

I will discuss some recent developments in estimating a regression function under shape constraints like monotonicity and convexity in fully nonparametric and semiparametric settings. One particular class of models that will be considered in some depth is an extension of the GLM to incorporate shape constraints in one or more covariates. Under a constraint like monotonicity of the regression function (in one or more covariates), the usual square-root n (n being the sample size) consistency of the MLE of the regression function breaks down; furthermore, the MLE is no longer asymptotically Gaussian.

I will discuss inference for the regression function using likelihood ratio based inversion,

which turns out to be an effective strategy, since the likelihood ratio statistic (for testing for the value of the regression function at a point) in a broad class of monotone function models is asymptotically pivotal, though no longer chi-squared; rather it is described in terms of a functional of the slopes of unconstrained and constrained convex minorants of Brownian motion with quadratic drift. I will also talk about some of the open problems in this area.

GAUSSIAN PREDICTIVE PROCESSES MODELS FOR LARGE MULTIVARIATE SPATIAL DATA

Sudipto Banerjee*, Andrew Finley, Huiyan Sang and Alan Gelfand

With accessibility to geocoded locations where scientific data are collected through Geographical Information Systems (GIS), investigators are increasingly turning to spatial process models for carrying out statistical inference. Over the last decade hierarchical models implemented through Markov Chain Monte Carlo (MCMC) methods have become especially popular for spatial modelling, given their flexibility and power to estimate models (and hence answer scientific questions) that would be infeasible with classical methods. However, fitting hierarchical spatial models often involves expensive matrix decompositions whose computational complexity increases exponentially with the number of spatial locations. This renders them infeasible for large spatial data sets. This computational burden is exacerbated in multivariate settings with several spatially dependent response variables.

Our current work proposes to use a predictive process derived from the original spatial process that projects process realizations to a lower-dimensional subspace thereby reducing the computational burden. We discuss attractive theoretical properties of this predictive process as well as its greater modelling flexibility compared to existing methods. In particular, we show how the predictive process seamlessly adapts to settings with nonstationary processes, with richer and more complex space-varying regression models and with multivariate coregionalized models. A computationally feasible template that encompasses these diverse settings will be presented. We will illustrate our methodology with simulated and real data sets.

FRAMEWORK FOR NATIONAL ECONOMIC STATISTICS : MICRO-MICRO LINKAGE MATTERS

R. B. Barman

The ever growing importance of economic statistics as basis for supporting sound economic policy for growth and stability in the globally integrated environment has been a

major driving force in the globally integrated environment has been a major driving force in the improvement of official statistical system in many countries. This process is helped by proliferation of information and communication technology which facilitates capture of raw data as byproduct of transaction processing and their use for business decision using analytical tools. However, the main framework for collection and compilation of national economic statistics continues to be the system of national account designed to support macroeconomic analysis. This framework is of limited use for deeper understanding of the multi dimensional distributional characteristics of the economy and the dynamic impulses contributing to changes in such distribution. In my view, the transmission channels appear hazy in the absence of micro linkages. The development of framework for micro-macro linkages in national economic statistics is no doubt challenging, but it is still possible to create building blocks for such a framework. Considering that this will help shed much better light on the evolution of various dimensions of socio-economic development , helping in informed decision at different levels, I will discuss how this idea can be taken forward for improving the framework for national economic statistics.

TWO SAMPLE ADAPTIVE SEQUENTIAL TRIALS

Subir K. Bhandari

Two sample adaptive sequential rule is considered for general simple and multiple hypothesis testing context and exact optimal rules are attempted for. This paper is relevant in the context of searching for rules that tries on unconstraint minimisation of Average Number of Wrong Treatments in the context of Bernoulli clinical trials.

TIME SERIES ANALYSIS OF CATEGORICAL DATA USING AUTO-MUTUAL INFORMATION: A STUDY OF AR(2) MODEL

Atanu Biswas⋆ and Apratim Guha

There is little attention in the modeling of time series data of categorical nature. In this paper, we present a framework based on the Pegram's operator (1980) that was originally proposed only to construct discrete AR(p) processes. We extend the Pegram's operator to accommodate categorical processes with ARMA representations. The concept of correlation is not always suitable for categorical data. As a sensible alternative, we use the concept of mutual information, and introduce auto-mutual information to define the time series process of categorical data. Some inferencial aspects are discussed. The inferential procedure is easy for AR(1) model and becomes complicated for more complicated models. In the present paper we discuss the inferential aspects of AR(2) models with the help of a real data set.

TESTING FOR UNIT ROOTS IN PANELS WITH A FACTOR STRUCTURE

Jorg Breitung and Samarjit Das⋆

This paper considers various tests of the unit root hypothesis in panels where the cross section dependence is due to common dynamic factors. Three situations are atudied. First , the common factors and idiosyncratic components may both be nonstationary. In this case test statistics based on test statistics are invalid Second, if the common component is I(1) and the iodiosyncratic component is stationary (the case of cross-unit cointegration), then both the OLS and the GLS statistics fail. Finally, if the iodiosyncractic components are I(1) but the common factors are stationary , then the OLS based test statistics are asymptotically valid in this situation. A Monte Carlo study conducted to verify the asymptotic results.

ON SOME PROBABILISTIC ALGORITHMS FOR COMPUTING TUKEY'S HALF SPACE DEPTH

Biman Chakraborty⋆ and Probal Chaudhuri

The halfspace depth of a d-dimensional point $\boldsymbol{\theta}$ relative to a data set $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ is defined as the smallest number of observations in any closed halfspace with boundary through $\boldsymbol{\theta}$. Let H be the hyperplane passing through $\boldsymbol{\theta}$ and $d-1$ data points $\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_{d-1}}$. Then the multivariate Hodges's sign test (Hodges 1955) statistic is obtained by maximizing the absolute difference between the number of data points that fall on opposite sides of the plane H. Chaudhuri and Sengupta (1993) showed that this is equivalent to finding halfspace depth at $\boldsymbol{\theta}$ and the problem of computing depth of $\boldsymbol{\theta}$ reduces to a combinatorial optimization problem over $d-1$ subsets of data points. For bivariate data, the halfspace depth of a point can be computed in $O(n \log n)$ time (Matoušek 1991, Rousseeuw and Ruts 1996). For $d = 3$, an exact algorithm is available to compute it in $O(n^2 \log n)$ time (Rousseeuw and Struyf, 1998). An analogous algorithm can be constructed to compute the exact depth of a point in dimension $d > 3$ with time complexity $O(n^{d-1} \log n)$. It appears that there is virtually no hope for solving such a problem exactly for high dimensional data. In this work, we develop and study some probabilistic algorithms (Chakraborty and Chaudhuri 2003) to approximate the halfspace depth of a point for large high dimensional data using techniques based on Markov chains. Our algorithms were motivated by well known probabilistic combinatorial optimization techniques like genetic algorithms (Goldberg 1989), simulated annealing (Kirkpatrick et al. 1983, Geman and Geman 1984) and Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970). We establish convergence of our algorithms to the true halfspace depth as the number of iterations in the Markov chain increases. We also demonstrate the performance of our algorithm when applied to real and simulated high dimensional data sets.

CANCER CLASSIFICATION AND GENE SELECTION WITH ADAPTIVE BAYESIAN NEAREST NEIGHBOR: AN INTEGRATED APPROACH

Sounak Chakraborty

Researchers in many disciplines face the formidable task of analyzing massive amounts of high-dimensional and highly-structured data. Driven by the complexity of these new problems a new field of study known as "data mining" has emerged. A term that we liberally interpret to include model selection, classification or supervised learning and clustering or unsupervised learning. While most of the well established methods of clustering and classification are algorithmic, the Bayesian approach enables us to develop a probabilistic model. The Bayesian theory gives us a framework to incorporate some prior information in term of the expert's opinion and hence develop much more sophisticated learning criteria. Although the Bayesian methods are often computationally intractable or challenging but they can provide a standard of optimal decision making against which other methods can be measured.

Nearest neighbor methods are among one the most popular methods used in classification. But they are plagued with their sensitivity to the size of the neighborhood k, shape of the metric, poor performance in higher dimension and inability to give a probabilistic interpretation. In this paper we formulate an adaptive Bayesian classification method for the analysis of the glioma and prostate cancer with microarray data based on nearest neighbor setup. It is shown that our proposed Bayesian classification model with adaptive metric can produce much accurate classification scheme compared to several existing classical methods. We have also proposed a Bayesian variable selection scheme for selecting the differentially expressed genes integrated with our model. This integrated approach improves classifier design by yielding simultaneous gene selection and class prediction.

SMOOTH ESTIMATION OF MULTIVARIATE DISTRIBUTION AND DENSITY FUNCTIONS

Yogendra P. Chaubey

Chaubey and Sen (1996) introduced a new method for smooth estimation of survival and density functions based on Hille's lemma. This method was developed with a view to estimate a distribution function concentrated on the non-negative half of real line but it has proved to be useful in various other situations. Chaubey and Sen (1998) adapted it for smooth estimation of hazard and cumulative hazard functions and investigated it further [see Chaubey and Sen (1999)] for smooth estimation of mean residual life. Above results have

been also extended to the case of random censoring where the role of the empirical distribtion function has been replaced by the Kaplan-Meier product limit estimator [see Chaubey and Sen (1998)]. Further generalizations of Hille's lemma yield attractive smoothing techniques such as Bernstein polynomial estimator [Babu, Canty and Chaubey (2000)] of a density on the interval $[0, 1]$, multivariate Bernstein polynomial estimator of a density on $[0, 1]^d$ [Babu and Chaubey (2006)], smooth estimation of multivariate survival and density functions [Chaubey and Sen (2002)] and smooth regression for non-negative data [Chaubey, Sen and Zhou (2002)]. In this talk we will highlight some of these recent developments with a view to non-parametric multivariate density estimation based on a multivariate generalization of Hille's theorem along with some applications.

ESTIMATION OF THE MINIMUM MEAN OF NORMAL POPULATIONS UNDER THE TREE ORDER RESTRICTION

Sanjay Chaudhuri

We consider $s + 1$ univariate normal populations with common variance σ^2 and means $\mu_i, i = 1, 2, ..., K, s$, constrained by the tree order restriction $\mu_0 \leq \mu_i, i = 1, 2, ..., K, s$. The maximum likelihood estimator of μ_0 is known to diverge to $-\infty$ almost surely, $s \to \infty$ as if all means μ_i are bounded from above and all sample sizes $n_i{}^{(s)}$ remain finite. However, this is not true if the means are unbounded or more importantly if $n_0{}^{(s)}$ increases with s, which is the case of most practical interest. In such cases it can be shown that the m.l.e. of μ_0 is consistent, or at least is bounded from below. The consistency of a modified version of an estimator due to Cohen and Sackrowitz (2002) will also be discussed.

SURFACE SHAPE ASYMMETRY ANALYSIS IN COMPUTATIONAL NEUROANATOMY

Moo K. Chung

There is a lack of a systematic investigation of structural brain asymmetry analysis in literature. Most of previous approaches start with mirroring 3D magnetic resonance images of the brain. The structural correspondence across the hemispheres is then established by registering the original image to the mirrored image. A simple difference of an anatomical index between these two images is used as an asymmetry index. We present a radically different asymmetry analysis that utilizes the recently developed weighted spherical harmonic (SPHARM) representation of the cortex. The weighted-SPHARM is a diffusion smoothing technique on a unit sphere expressed as a weighted linear combination of the spherical harmonics. Due to the angular symmetry presented in the spherical harmonics, the weighted-

SPHARM provides a natural mathematical correspondence across the hemispheres without the time-consuming image registration.

MEASURING SPATIAL COST OF LIVING INDICES WITHOUT USING PRICE DATA

Dipankar Coondoo* and Amita Majumder

In this paper we propose a method of estimating multivariate price index numbers from cross-section consumer expenditure data on different items using Engel curve analysis. The novelty of the procedure is that it does not require item-specific price/unit value data. Also, no explicit specification of functional from of the coefficients of the Engel curve is needed. This simple method is likely to be particularly useful in studies of regional comparisons of poverty and inequality, optimal commodity taxes and tax reforms especially when only grouped consumer expenditure data are available. To illustrate the method, we use published data of the 50th round (1993-94) and 55th round (1999-2000) consumer expenditure surveys of India's National Sample Survey Organization (NSSO). We calculate the spatial consumer price index numbers for fifteen major states of India, with all-India taken as base, separately for the rural and the urban sector.

QUALITY INDEX AND MAHALANOBIS D^2 STATISTIC

Ratan Dasgupta

In conformity with 'Bhattacharyya affinity' between two probability distributions, definition of Mahalanobis distance is extended to the case when the dispersion matrices of two populations are not necessarily equal. We show that the Mahalanobis distance is unperturbed by inclusion of highly correlated additional variables. A 'quality index' is proposed to compare individuals, industrial products, computer software and objects with multiple characteristics, in general. As an illustration, a group of students is ranked, on the basis of their scores in three different types of tests for entrance examination.

A MULTIVARIATE EXAMINATION OF BERG VARIABLES IN SPINAL CORD INJURY PATIENTS

Somnath Datta

We present a principal component analysis for the temporal changes of fourteen Berg variables obtained from a population of spinal cord injury patients. A number of intereting findings emerges which validates the use of Berg as an instrument of recovery for SCI patients.

PROTEOMICS BASED APPROACHES FOR EARLY DETECTION OF FETAL ALCOHOL SYNDROME

Susmita Datta

Alcohol consumption during pregnancy has far reaching consequences. Fetal Alcohol Syndrome (FAS) is one of them. Earlier detection of FAS is one of the most important sought after birth defect related research. We consider multivariate statistical analysis of high throughput proteomic profile of mouse amniotic fluid using MALDI-TOF mass spectrometer to detect biomarker of FAS. At an early stage of the research we find a broad based protein to be one of the most important biomarker which is consistent with earlier epidemiological study.

INVESTMENT DECISIONS AND PRICING OF FINANCIAL INSTRUMENTS

Abhijit De⋆, Avinash Gupta, and Ashish Kumar

Making right investment decisions for investor is always very crucial to get descent return (or the aggressive return depending upon the risk taking profile of investor). In this project an attempt has been made to help such investor in identifying the sectors as per his risk profile and in order to get the desired rate of return over their investment. Paper intends to develop methodology which will enable decision making for fund allocation among different financial instruments and to maximize return and to minimize risk on the investment. Two phase method is designed and formulated to solve the problem. In first phase financial instruments are categorized in different sectors. Problem is modeled using quadratic programming to identify the sectors giving descent return thus helping investor to focus on the particular sector(s). Now different weights are assigned to those sectors for allocation of funds. In second phase problem is solved for each sector to allocate assigned fund among constituent financial instruments. Developed model is applied on BSE data. It is found that FMCG, healthcare, bank and technology sectors are the most rewarding sectors where one should invest their money.

K-MEANS CLUSTERING: A NOVEL PROBABILISTIC MODELING WITH APPLICATIONS

Dipak K. Dey⋆ and Samiran Ghosh

Over the last decade a variety of clustering algorithms have evolved. However one of the simplest (and possibly overused) partition based clustering algorithm is K-means. It can

be shown that the computational complexity of K-means does not suffer from exponential growth with dimensionality rather it is linearly proportional with the number of observations and number of clusters. The crucial requirements are the knowledge of cluster number and the computation of some suitably chosen similarity measure. For this simplicity and scalability among large data sets, K-means remains an attractive alternative when compared to other competing clustering philosophies especially for high dimensional domain. However being a deterministic algorithm, traditional K-means have several drawbacks. It only offers hard decision rule, with no probabilistic interpretation. In this presentation we have developed a decision theoretic framework by which traditional K-means can be given a probabilistic footstep. This will not only enable us to do a soft clustering, rather the whole optimization problem could be recasted into Bayesian modeling framework, in which the knowledge of cluster number could be treated as an unknown parameter of interest, thus removing a severe constrain of K-means algorithm. Our basic idea is to keep the simplicity and scalability of K-means, while achieving some of the desired properties of the other model based or soft clustering approaches. The methodology will be exemplified through various examples which will include moderate to very high dimensional problems.

IMPROVED ESTIMATION UNDER STOCHASTIC LINEAR CONSTRAINTS

M. Dube and S. Chandra*

The article analyzes the performance of various improved estimators under balanced loss proposed by Zellner. Risks properties of the estimators have been derived using small disturbance asymptotic approximations and assuming error distribution to be not necessarily normal. Attempts have also been made to study the comparative performance of various estimators under balanced loss when some stochastic prior information in the form of linear restrictions binding the regression coefficients is available.

OCEANOGRAPHIC APPLICATION OF MULTIVARIATE STATISTICAL METHODS

A. A. Fernandes

BAYESIAN APPROACH TO ESTIMATION OF POSITIVE FALSE DISCOVERY RATE

Subhashis Ghosal*, Yongqiang Tang and Anindya Roy

In the recent years, multiple hypothesis testing has come to the forefront of statistical research especially due to the emergence of new fields like genomics, proteomics and fMRI. The main goal in these analyses is often identifying significant hypotheses among several thousands of hypotheses, most of which are insignificant. The classical approach to testing leads to procedures that are too conservative. More appropriate measures of errors are given by the False Discovery rate (FDR) and its variants, and the idea is to develop a multiple testing procedure to control this error. A particularly fruitful approach is based on a mixture model framework developed by Storey. We argue that in this setting, a Bayesian approach is natural and able to exploit salient features in the model to increase the efficiency of estimation. By considering a mixture of beta densities to model alternative density and putting a Dirichlet process prior on the mixing distribution, we develop a nonparametric Bayesian solution of the problem. We discuss computational techniques and convergence properties of our procedure. Through a simulation study, we demonstrate that our estimator, in general, has smaller mean squared error.

RANK CHASING: DO FRACTILES OF RETURNS AFFECT MUTUAL FUND INFLOW?

Aurobindo Ghosh

Tax liability influences investor's decision on investment choice. It is established that return chasing behavior among investors introduce a non-linearity in the performance-flow relationship. This non-linearity could be handled in a robust way if we look at the ranks rather than merely risk-adjusted returns as performance measures. We explore the distributional effects of the empirical distribution function (EDF) of risk-adjusted returns on mutual fund flow. We use smooth tests for both the unadjusted inflow distribution as well as after conditioning for risk adjusted returns to illustrate how higher order moments are changed due to investors tax exposure and regulations.

A RANDOM WALK THROUGH OLD AND NEW METHODS OF VARIABLE SELECTION

J.K. Ghosh

Variable Selection (or equivalently multiple tests about means in the orthogonal case) has emerged as one of the most important part of handling high dimensional data. There has

been an explosion of new methods, from which I choose to speak on two. I will also speak on some recent results on AIC, which is the oldest statistical method of model selection. The results on AIC are joint work with Arijit Chakrabarty (ISI) and have been published in AISM (2006).

One of the new methods is sufficient dimension reduction with no loss of information, introduced by Li in a series of papers. I will present joint work with MIchael Zhu (Purdue) and Surya Tokdar (CMU) on a new Bayesian algorithm, for which we have a consistency proof.

The second new method, based on convex optimization, appears to be a major breakthrough by Candes and Tao (Tao is a Fields Medalist of 2006). Candes and Tao call their method the Danzig selector after Danzig of linear programming fame because their method can be implemented via linear programming.

In the preparation of this talk I have benefitted from discussions with Samiran Ghosh and Surya Tokdar.

ESTIMATION, PREDICTION AND THE STEIN PHENOMENON UNDER DIVERGENCE LOSS

Malay Ghosh

We consider two problems: (1) estimate a normal mean under a general divergence loss,and (2) find a predictive density of a new observation drawn independently of the sampled observations from a normal distribution with the same mean but possibly with a different variance under the same loss. The general divergence loss includes as special cases both the Kullback-Leibler and Bhattacharyya-Hellinger losses. The sample mean, which is a Bayes estimator of the population mean under this loss and the improper uniform prior, is shown to be minimax in any arbitrary dimension. A counterpart of this result for predictive density is also proved. However, the general Baranchick class of estimators, which includes the James-Stein estimator and the Strawderman class of estimators, dominates the sample mean in three or higher dimensions for the estimation problem. An analogous class of predictive densities is defined and any member of this class is shown to dominate the predictive density corresponding to a uniform prior in three or higher dimensions.

BOUNDARY POINT DETECTOR USING KERNAL PRINCIPAL COMPONENT ANALYSIS

O.S. Deepa Gopakumar⋆ and Krishnan Namboori P.K

Principal component analysis (PCA) is a widely used statistical technique for unsuper-

vised dimension reduction. High dimensional data are often transformed into lower dimensional data via the principal component analysis (PCA) where coherent patterns can be detected more clearly. Such unsupervised dimension reduction is used in very broad areas such as meteorology, image processing, genomic analysis, and information retrieval .

Given a set of n data points in some d-dimensional space, the goal of clustering is to partition the data points into k clusters. Clustering can be considered as an instance of unsupervised techniques used in machine learning. If the data are tightly clustered or lie in nice linear subspaces then simple techniques like k-means clustering can find the clusters. However in most of the cases the data lie on highly nonlinear manifolds and a simple measure of Euclidean distance between the points in the original d-dimensional space gives very unsatisfactory results. A plethora of techniques have been proposed in recent times which show very impressive results with highly nonlinear manifolds. The important among them include Kernel Principal Component Analysis.

In this paper, we propose a simple yet novel approach BORDER (a BOundaRy points DEtectoR) by means of a kernel principal component analysis to detect the boundary points that are located at the margin of densely distributed data.

Utilizing **Reverse k-Nearest Neighbor** (RkNN) in data mining tasks will require the execution of a RkNN query for each point in the dataset (the set-oriented RkNN query). However, this is very expensive and the complexity will be $O(N^3)$ since the complexity of a single RkNN query is $O(N^2)$ time using sequential scan for non-indexed data, where N is the cardinality of the dataset. In the case where the data is indexed by some hierarchical index structure, the complexity can be reduced to $O(N^2 \log N)$. However, the performance of these index structures is often worse than sequential scan in high-dimensional space. Instead of running multiple RkNN queries, the proposed BORDER approach utilizes Gorder kNN join (or the G-ordering kNN join method) which performs KPCA technique for transformation of datasets to find the reverse k-nearest neighbors for a set of data points. Gorder is a block nested loop join method that exploits sorting, join scheduling and distance computation filtering and reduction to reduce both I/O and CPU costs. It sorts input datasets into the *G-order* and applies the *scheduled block nested loop join* on the G-ordered data. It also employs distance computation reduction to further lower the CPU costs. It doesn't require an index and handles high-dimensional data efficiently. The study shows that BORDER with Gorder kNN join using KPCA is scalable to large data size.

MULTIVARIATE FACTORIAL DESIGNS
WHEN THE NUMBER OF FACTOR LEVELS IS LARGE

Solomon W. Harrar⋆ and Arne C. Bathke

We obtain the asymptotic distributions of different multivariate parametric and non-

parametric tests for the situation where the number of replications is limited, whereas the number of treatments goes to infinity (large k, small n case). For the parametric case, we consider the Dempster's ANOVA-type, Wilks Lambda, Lawley-Hotelling and Bartlett-Nanda-Pillai Statistics. In the nonparametric case, we propose the rank-analogs of the Dempster's ANOVA-type, Lawley-Hotelling and Bartlett-Nanda-Pillai statistics. The tests are based on separate rankings for the different variables.

We provide a finite sample approximation procedure in both the parametric and nonparametric cases. The finite performance of the tests is investigated through simulations. It turns out that the proposed nonparametric tests perform very well as compared to their parametric competitors, especially in the presence of outliers.

An example illustrates the application.

ANALYSIS OF MICROTUBULE DYNAMICS

S. Rao Jammalamadaka⋆ and M. A. Siddiqi

Microtubules are sub-cellular structures in most plant and animal cells and play a major part in cell locomotion, cell transport and cell division. Microtubule images in living cells are acquired by fluorescence microscopy. These images are in the form of a video which is essentially a stack of images taken over successive time intervals. This work analyzes microtubule activity under various treatments, using statistical techniques that have been developed for "Growth Curve" modeling. Mean growth curves are fitted and statistical tests carried out for relevant biological hypotheses for microtubule dynamicity in living cells. The analyses and conclusions have important biological significance.

LINKING A PARAMETRIC FAMILY OF DISTRIBUTIONS WITH NONPARAMETRIC QUANTILE ESTIMATION AND REGRESSION

Chris Jones

On the one hand, there is classical statistical modelling involving the fitting of parametric families of distributions; on the other, one can learn about one's data via nonparametric approaches such as kernel density estimation. In this talk, I shall explore a specific intriguing link between the two. Starting from the special case of the log F distribution, I will describe a novel way of generating four/three-parameter families of continuous univariate distributions with simple exponential tail behaviour. Skewness is to the fore. I will then consider the interaction between this family of distributions and kernel quantile estimation: the latter is maximum likelihood estimation of location in the former. And finally, I will consider the

practical ramifications of this link for kernel quantile regression. In particular, I will end up, in what is joint work with Keming Yu, with an improved version of the "double kernel local linear quantile regression" method that we had previously proposed.

BIAS-CORRECTED AIC FOR SELECTING MULTIVARIATE GMANOVA MODELS UNDER NONNORMALITY

Ken-ichi Kamo[⋆] and Hirokazu Yanagihara

We consider an AIC (Akaike's information criterion) type information criterion in multivariate GMANOVA model under nonnormality. We correct the bias of AIC by using the sum of square of predicted residuals without adding any bias correcting terms, and we obtain a new information criterion "Jackknifed AIC" (AICJ). AICJ has two well properties than AIC. First, the order of bias is improved as first order accuracy. Next, AICJ becomes the unbiased estimator of Kullback-Leibler information when the distribution of true model is normal. We will report the performance for AICJ by numerical examinations.

MULTIVARIATE PROCESS CAPABILITY USING RELATIVE IMPORTANCE OF QUALITY CHARACTERISTICS

K. G. Khadse and R. L. Shinde[⋆]

Consider a production process where quality of the product is determined by more than one-quality characteristics. We discuss the classification of quality characteristics according to their relative importance. We propose probability based multivariate process capability indices (MPCIs) which takes into account relative importance of quality characteristics. The two approaches based on (i) modified specification region and (ii) modified process region, are used to incorporate the relative importance of quality characteristics.

CERTAIN MIXED MULTIVARIATE MODELS CONNECTED WITH RANDOM SUMS AND MAXIMA, WITH APPLICATIONS TO CLIMATOLOGY AND FINANCE

Tomasz J. Kozubowski[⋆], Anna K. Panorska, and Franco Biondi

Motivated by problems arising in hydro-climatology, we consider multivariate distributions connected with X, Y, and N, where N has a discrete distribution on positive integers while X and Y are the sum and the maximum of N IID copies of a random variable W,

independent of N. Our focus are bivariate distributions of (X, N) and (Y, N) in case where N is geometric and W is exponential. We present their basic properties, including marginal and conditional distributions, joint integral transforms, infinite divisibility, stability with respect to geometric summation, and estimation, and argue why these are natural models for describing the joint behavior of magnitudes, peak values, and durations of hydro-climatic episodes. Another example with N representing the number of consecutive positive daily log-returns of currency exchange rates illustrates the modeling potential of these distributions in mathematical finance.

APPLICATION OF MULTIVARIATE TECHNIQUES
IN SELECTION AND RANKING PROBLEMS

Narinder Kumar

In this paper, a class of subset selection procedures for selecting a subset containing the best population is proposed assuming that the underlying populations differ only in their scale parameters. The population which corresponds to the smallest scale parameter is labeled as the best population. This type of formulation is usually encountered in agriculture, engineering, business etc. In agriculture when the average yield of all the variables of a crop is same, the aim is to select more consistent varieties to be recommended for small and marginal farmers. The proposed class of procedures is shown to satisfy the P*-condition and have strong monotonicity, monotonicity, and unbiasedness properties. As a result of multivariate techniques, the proposed class of procedures is compared asymptotically with the existing procedures and the results are found to be interesting. The implementation of the proposed class of procedures is illustrated with the help of existing tables and the sample size sufficient for their implementation is worked out using extensive simulation study.

ESTIMATING RELIABILITY OF A SELECTED PARETO POPULATION

Somesh Kumar

DETERMINATION OF DISCRETE SPECTRUM
IN A RANDOM FIELD

Debasis Kundu

We consider a two dimensional frequency model in a random field, which can be used to model textures and also has wide applications in Statistical Signal Processing. First we

consider the usual least squares estimators and obtain the consistency and the asymptotic distribution of the least squares estimators. Next we consider an estimator, which can be obtained by maximizing the periodogram function. It is observed that the least squares estimators and the estimators obtained by maximizing the periodogram function are asymptotically equivalent. Some numerical experiments are per- formed to see how the results work for finite samples. We apply our results on simulated textures to observe how the different estimators perform in estimating the true textures from a noisy data.

BAYESIAN ANALYSIS OF RANK DATA

Arnab Kumar Laha

In many fields like market research, opinion polls, sports etc. respondents are often asked to rank a set of items according to their preferences. We are interested in the estimation of true rank of the items when the ranks given by the respondents are subject to error. A new model for such data is proposed and based on this model we develop a Bayesian methodology for the estimation of the true rank. The posterior marginal distribution of the true rank can be conveniently obtained using (a) Sampling Importance Resampling or (b) Gibbs Sampling. An extension to the problem of estimating the true rank when the respondents provide only a partial ranking is also discussed. Several real life data sets are analysed to illustrate the methodology.

A CLASS OF GOODNESS OF FIT TESTS
FOR MARKOV RANDOM FIELDS

S.N. Lahiri⋆, and Mark Kaiser

In this talk, we formulate some goodness of fit tests for fitting Markov random field models to spatial data observed on a regular grid. The key construction is based on a new notion (called "conclique") of grouping the observations. We derive the asymptotic null distribution of the goodness of fit test statistics in two cases - (i) where the null hypothesis is simple and (ii) where it is composite, without a specified value of the model parameter. The asymptotic distribution has a non-standard form in the latter case.

A GCV APPROACH TO BANDWIDTH
SELECTION IN POSITRON IMAGE
TOMOGRAPHY IMAGE RECONSTRUCTION

Ranjan Maitra

A generalized cross-validation (GCV) approach to bandwidth estimation for Positron

Emission Tomography (PET) image reconstruction is developed in the context of generalized deconvolution. Results on eigendecomposition of symmetric one- and two-dimensional circulant matrices are derived and their computational implications in making reconstruction bandwidth estimation a simple extension of standard filtered backprojection studied. The approach extends readily to radially asymmetric smoothing kernels, widening the class of filters available for practical PET reconstruction. Performance evaluations on a class of idealized one-dimensional generalized deconvolution problems and PET phantom are excellent and in real-time. Finally, a practical application of the methodology to a PET study is presented.

BAYESIAN HIERARCHICAL SPATIALLY CORRELATED FUNCTIONAL DATA ANALYSIS WITH APPLICATION TO COLON CARCINOGENESIS

Bani K. Mallick

In this talk we present new methods to analyze data from an experiment using rodent models to investigate the role of p27, an important cell cycle mediator, in early colon carcinogenesis. The responses modeled here are essentially functions nested within a two-stage hierarchy. Standard functional data analysis literature focuses on a single stage of hierarchy and conditionally independent functions with near white noise. However, in our experiment, there is substantial biological motivation for the existence of spatial correlation *among* the functions, which arise from the locations of biological structures called colonic crypts: this possible functional correlation is a phenomenon we term *crypt signalling*. Thus, as a point of general methodology, we require an analysis that allows for functions to be correlated at the deepest level of the hierarchy. Our approach is fully Bayesian and uses Markov Chain Monte Carlo methods for inference and estimation. Analysis of this data set gives new insights into the structure of p27 expression in early colon carcinogenesis and suggests the existence of significant crypt signalling. Our methodology uses regression splines, and because of the hierarchical nature of the data, dimension reduction of the covariance matrix of the spline coefficients is important: we suggest simple methods for overcoming this problem.

PRICE PUZZLE UNDER VAR

Kumarjit Mandal

Generally it is found in the VAR models relating to monetary variables that when the nominal interest rate increases due to a shock in the system the inflation rate increases at the

first instance. This presents a puzzle in monetary economics. In this study an attempt has been made to mitigate the price puzzle with inflation data based on consumer price index and estimating the VAR with local projection.

AQL BASED MULTIATTRIBUTE SAMPLING SCHEME
Anup Majumdar

Irrespective of the type of product, evaluation of conformity to specified requirements of its quality characteristics is an integral part of quality assurance. Although they form a set of necessary verification activities almost at all stages of production, these activities, known as inspection do not add value to the product on their own and are to be kept at their minimum. The sampling inspection where a portion of a collection of product units is inspected on a set of characteristics with a view to making decision about acceptance or otherwise becomes relevant in this context. The number of elements of the set of characteristics to be verified for this purpose is unlikely to be just one in most of the practical situations. We may call this as multiattribute inspection as employed for verification of materials procured from outside and further at all stages of production, through semi finished and finished or assembly stages to final despatch to the customers. At all such stages consecutive collections of products called lots, are submitted for acceptance or alternative disposition.

In this context one observes that the defectives with respect to different characteristics may in some situations be considered as jointly independent, whereas in some situations, occurrences of one type of defect may preclude occurrence of any other type. In some other situations, it is more natural to count number of defects for each unit of product so that the quality of a lot is expressed as an ordered set of average number of defects per unit for each type of defect instead of proportion defectives in an aggregate. We have attempted to develop the procedure of sampling inspection by attribute with this scope and purpose in mind. One may call them as multiattribute single sampling plans(MASSP).

For the present discussions we restrict ourselves to establish a sampling scheme in line with available international standards tabulated on the basis of Acceptable Quality Level(AQL) in such mul-tiattribute situations. These Standards in general prescribe that separate plans are to be chosen for the different classes of attributes.

We examine the consequences of constructing a sampling plan by this prescribed method in a multiattribute situation. We first consider the effective producer's risks. Secondly, it has been thought as reasonable to expect that the Operating Characteristic (OC) function should be more sensitive to the changes in the defect level of more important attributes, particularly, in a situation, where unsatisfactory defect level occurs due to more serious type of defects. A measure of sensitivity has been defined, for this purpose. Plans constructed with all possible practically useful combinations of AQL for three attributes have been examined

for both the as dicussed. It has been found that the features observed depict a picture far from the ideal on both the counts for most of the plans.

Further, it has been discovered that these two desirable properties remain generally dissatisfied for the class of plans (we call these as MASSP of C kind), where we accept a lot if the number of defectives observed in the sample is less than the respective acceptance numbers stipulated separately for each attribute depending upon the AQL. We have, therefore, introduced a sampling scheme consisting of plans with same sample size as used by the International standards for a given lot size, but with a different acceptance criterion. (we call this as MASSP of A kind). For these plans the OC function is more sensitive to the changes in the defect level of more important attributes characterized by low AQL values. Further, they can be deigned to ensure a reasonable producer's risk. These plans are tabulated for ready use by the industries, encountering the need for multiattribute inspection schemes.

ON S.N. ROY'S LEGACY TO MULTIVARIATE ANALYSIS

Kanti V. Mardia

I describe here my connection with two of the major contributions of S.N. Roy: namely the Jacobians of complicated transformations for various exact distributions, rectangular coordinates and the Bartlett decomposition. Their applications have appeared in directional statistics, shape analysis and now in statistical bioinformatics.

PROGNOSTIC MODELS WITH CLINICAL AND BIOMARKER DATA: A MULTIVARIATE GRWOTH CURVE PERSPECTIVE IN CLINICAL TRIALS

Sati Mazumdar*, Stewart J. Anderson, Feng-shou Ko and Qianyu Dang

Among the many contributions to multivariate statistical methods made by Dr. S.N. Roy was the extension of growth curve models to the case where two or more measurements are considered at each time point (Potthoff and Roy, *Biometrika*, 313-316, 1964). In the years since that contribution, many variations of that extension have been proposed both for normally and nonnormally distributed data. This paper begins with two types of prognostic models using clinical and biomarker measures obtained in clinical treatment trials. The first type uses trajectories of clinical and biomarkers obtained during the acute treatment phase of an illness and investigate whether these trajectories have predictive ability for recurrences in the maintenance phase of the treatment trial. The statistical problems lie in

the estimation of these trajectories from multivariate (continuous and discrete), irregularly spaced, often missing, growth type measures and accounting for estimation errors in using them as predictors in prognostic models. The methods are illustrated and evaluated with data from a maintenance treatment trial for major depression in an elderly population and simulation studies. The second type involves a method to identify longitudinal biomarkers using frailty models in survival analysis. The method allows random effects to be present in both the longitudinal biomarker and underlying survival function. We use simulations to explore how the number of individuals, number of time points and the functional form of the random effects influence power to detect the association of a longitudinal biomarker and the survival time. An application using prothrombin index as biomarker for survival of liver cirrhosis patients is provided. Possible extensions with multivariate longitudinal measures, both clinical and biomarkers are discussed.

SYMMETRY, IG-SYMMETRY AND R-SYMMETRY

Govind S. Mudholkar

The symmetry, and more generally invariance, have central roles in the theory and applications of statistics. Mudholkar and Natarajan defined and studied the IG-symmetry and demonstrated that it plays an intriguing role in the inverse Gaussian theory. This concept, although mathematically well defined, lacks conceptual clarity. The notion of R-symmetry, which is physically transparent, is proposed as a remedy. It is studied, related to both the basic symmetry and IG-symmetry and an analogue of Wintner's classical result is established by showing that the product-convolution of R-symmetric unimodal distributions is R-symmetric unimodal. Furthermore its application to the power functions of the best unbiased tests for the significance of IG means is discussed.

ASYMPTOTIC DISTRIBUTIONS OF M-ESTIMATORS IN A SPATIAL REGRESSION MODEL UNDER SOME FIXED SPATIAL SAMPLING DESIGNS

Kanchan Mukherjee

In this talk, we consider M-estimation of the regression parameter in spatial multiple linear regression model. We discussed the consistency and the asymptotic normality of the M-estimators when the data-sites are generated by a class of deterministic spatial sampling schemes. It is shown that scaling constants of different orders are needed for asymptotic normality under different spatial sampling schemes considered here. Results are established

for M-estimators corresponding to certain non-smooth score functions including Huber's function and the sign functions.

MULTI-RESPONSE EXPERIMENTS AND STOCHASTIC OPTIMISATION

S. P. Mukherjee

Traditional Response Surface Analysis aims at locating optimum levels of continuously varying factors through the use of contour plots of the fitted response and optimum-seeking techniques. In some experiments, a need exists to optimize several responses-possibly conflicting and most often inter-related-simultaneously . Usually a compromised solution is worked out. Solution methods for Dual Response Systems (DRS) -an important class of multi-response experiments-adopt various non-linear programming algorithms which optimize one response, may be the most important, subject to constraints on the other. Such solutions are subject to two limitations. Firstly, these ignore sampling variations in the fitted responses or-in other words-treat the fitted responses as deterministic functions of levels of the two factors. Secondly, these are not applicable where both the responses are equally important as objective functions in an optimization exercise . One way to take care of the second problem is to set some goals for the values of the two responses and to formulate the problem as one in goal programming. In many experiments it is possible to set such goals. Regarding sampling variations in fitted responses, we need to look upon the responses as random objective functions . Even if we follow the first approach of constrained optimization and we take the expected value of the more important response as the deterministic objective function , we have a problem in chance constrained programming (CCP). To solve a problem formulated this way we need to convert the chance constraint into its deterministic equivalent and this calls for an easily tractable distribution for the corresponding response function. Stochastic Goal programming remains a complex problem. Starting with an assumed multivariate normal distribution of the estimated regression coefficients, one can work out the distributions of fitted responses, conveniently in the case of linear or quadratic responses. These will be pretty involved and to derive quantiles of such a distribution-as will be needed in a CCP-and to carry out a simulation exercise for this purpose will have to begin with some assumed valued of the parameters in the response function distributions. This paper simply comes up with different possible formulations of the DRS problem, hints at possible solution techniques and points out direct general ical level-to the multi-response situations.

ON OPTIMAL ESTIMATING FUNCTIONS
IN THE PRESENCE OF NUISANCE PARAMETERS

Parimal Mukhopadhyay

When there is only one interesting parameter μ_1 and one nuisance parameter μ_2 Godambe and Thompson (1974) showed that the optimal estimating function for μ_1 essentially is a linear function of the μ_1-score, the square of the μ_2-score and the derivative of μ_2-score with respect to μ_2. Mukhopadhyay (2000b) generalized this result to m nuisance parameters. Mukhopadhyay (2000, 2002 a, b) obtained lower bounds to the variance of regular estimating functions in the presence of nuisance parameters. Taking cue from these results we propose a method of finding optimal estimating function for μ_1 by taking the multiple regression equation on μ_1 score and Bhattacharyya's (1946) scores with respect to μ_2. The result is extended to the case of m nuisance parameters.

GENETIC ALGORITHMS

C.A. Murthy

This lecture deals with the basic principles and the formulation of Genetic Algorithms. The convergence theorem for elitist model of genetic algorithms will be stated. Some applications of Genetic algorithms will be discussed.

CHARACTERIZING MULTIVARIATE LIFE
DISTRIBUTIONS BY RELATIONSHIP BETWEEN
MEAN RESIDUAL LIFE AND FAILURE RATES

N Unnikrishnan Nair and Sudheesh Kumar Kattumannil*

Characterization of univariate life distributions by relationship between failure rate and mean residual life has been attempted by many authors, including those for specific distributions and general class of probability laws in the discrete and continuous cases. These help in modeling and analysis of lifetime data. However there is no general analogous result in the multivariate case, though some results are available for specific distributions in the bivariate case . In the present paper , we establish some characterizations of a general class of multivariate distributions, based on identities connecting multivariate failure rates and mean residual life.

BAYESIAN INFERENCE FOR A STRATIFIED CATEGORICAL VARIABLE ALLOWING ALL POSSIBLE CATEGORY CHOICES

Balgobin Nandram

In many sample surveys, there are items that require respondents to make at least one choice. For example, in the Kansas Farm Survey (conducted by the Department of Animal Sciences at Kansas State University), livestock farmers in Kansas were asked "What are your primary sources of veterinary information?" By level of education, the sources are professional consultant, veterinarian, state or local extension service, magazines, and feed companies and representatives, and the farmers were allowed to pick as many sources that apply. The analyses of such survey data are complex because an individual is allowed to make at least one choice, the number of individuals with none of these choices is unknown or not reported, and the categorical table with mutually exclusive categories are typically sparse. We use a Bayesian product multinomial-Dirichlet model to fit the count data both within and across education levels. We estimate the proportions of individuals with each choice, show how to select the best choice, and show using the Bayes factor how to test that these proportions are the same over different levels of farmers' education. Our Bayesian procedure uses a sampling based method with independent samples.

CANONICAL CORRELATION ANALYSIS IN FUNCTIONAL MRI

Rajesh R. Nandy

Detection of activation using functional MRI is often complicated by the low contrast-to-noise ratio in the data. The primary source of the difficulty is the fact that for neuronal activities that are subtle, the signal can be hidden inside the inherent noise in the data. Classical univariate methods based on t-test or F-test are susceptible to noise, as they fail to harness systematic correlations in evoked responses within neighboring voxels. A more sensitive post-processing method for fMRI data analysis based on canonical correlation analysis will be presented which may obtain more accurate statistical maps of subtle activations. Results comparing canonical correlation analysis and univariate methods will be presented using simulations and real data.

NONPARAMETRIC TEST FOR HOMOGENEITY OF OVERALL VARIABILITY

Hon Keung Tony Ng* and Ashis SenGupta

In this paper, we propose a nonparametric test for homogeneity of overall variabilities for two multidimensional populations. Monte Carlo simulation study is used to evaluate the

performance of the proposed nonparametric procedure. Comparison between the proposed procedure and the asymptotic parametric procedure based on generalized variances is made when the underlying populations are multivariate normal. We observed that even though the nonparametric procedure is not as powerful as the parametric procedure under normality, however, it is a reliable and powerful test for comparing overall variabilities under other multivariate distributions such as the multivariate Cauchy and the multivariate exponential distributions, even with small sample sizes. An example from a educational study is used to illustrate the proposed nonparametric test.

A CONSERVATIVE TEST FOR MULTIPLE COMPARISON BASED ON HIGHLY CORRELATED TEST STATISTICS

Yoshiyuki Ninomiya⋆ and Hironori Fujisawa

In genetics and epidemiology, we often encounter a large number of highly correlated test statistics. The most famous conservative bound for multiple comparison is the Bonferroni's bound, which is suitable when the test statistics are independent but not when the test statistics are highly correlated. This paper proposes a new conservative bound, which is easily calculated without multiple integration and is a good approximation when the test statistics are highly correlated. The performance of the proposed method is evaluated by simulation study and real data analysis.

SCATTER MATRICES, KURTOSIS AND INDEPENDENT COMPONENTS

H. Oja⋆, S. Sirkiä and J. Eriksson

In the independent component analysis (ICA) it is assumed that the components of the multivariate independent and identically distributed observations are linear transformations of latent independent components. The problem then is to find the (linear) transformation which transforms the observations back to independent components. In the talk the ICA is discussed and it is shown that, under some mild assumptions, two scatter matrices may be used together to find the independent components. The scatter matrices must have the so called independence property. See Oja et al. (*Austrian Journal of Statistics*, **35**, 175-189, 2006). The independent components are given in a kurtosis order. Different possible choices (robust and non-robust) of the two scatter matrices are discussed. The theory is illustrated with several examples.

CLASSIFICATION FOR TWO CORRELATED NORMAL POPULATIONS: DISTRIBUTION OF QUADRATIC DISCRIMINANT STATISTIC AND SOME MONTE CARLO STUDIES

S.H. Ong⋆, B.W. Yap and C. Low

Classification for two jointly distributed normal populations has been examined theoretically by a number of authors who have derived various asymptotic results. In this presentation the distribution of the quadratic discriminant function when two populations are dependent with proportional covariance matrices is derived. However, due to the complicated expression of the density function of the classification statistic, the performance of the classification rule for two jointly distributed multivariate normal populations with equal covariance matrix for (a) covariate classification, in the presence of (b) censored observations and (c) outliers will be considered through Monte Carlo simulation studies. The study is motivated by a real life data set on trace elemental concentration of breast cancer tumours. A necessary and sufficient condition for the positive definiteness of the partitioned covariance matrix has been derived in order to facilitate the simulation study.

PEAK MAXIMA DISTRIBUTION AND ITS DERIVED PEAK FACTOR FOR WIND LOAD COMBINATION

S. Nadaraja Pillai⋆, Y. Tamura

In this paper, a method for calculating the peak factor to estimate the maximum column normal stress due to the wind load combination is derived. The weighted distribution which shows almost same density of maxima distribution as the weighted narrow band given by S.O. Rice. The density of maxima even considering the bandwidth parameter ϵ in the Longuet-Higgins, also cannot predict the peak factor for the combination of loads . Considering this distribution with weighted reduced variate of the normal stess, the peak factor is derived. The distribution of maxima calculated from $\nu_Y(x) = \int_0^\infty \ddot{y} P_{Y\dot{Y}\ddot{Y}}(x, \dot{y}, \ddot{y}) d\ddot{y}$ where is the joint probability density of three variables . The derived function is weighted for the density of maxima

$$\mu_Y(x) = \frac{1}{2\pi}\sqrt{\frac{m_4}{m_0 m_2}}\left\{\frac{\epsilon}{2\pi}\exp\left(-\frac{x^2}{2\epsilon^2 m_0}\right) + \sqrt{1-\epsilon^2}\frac{x}{\sqrt{m_0}}\exp\left(-\frac{x^2}{2m_0}\right)\phi\left(\frac{x}{\sqrt{m_0}}\frac{\sqrt{1-\epsilon^2}}{\epsilon}\right)\right\}.$$

The above equation can be weighted with the x^n to form the weighted density distribution is then considered to derive the Peak factor. This peak factor can be used to estimate the maximum value for any random processes having the non narrow banded process. This peak factor and the experimental results are compared and discussed.

CLASSIFICATION AND MODEL AVERAGING: FIRST STAGES IN DEVELOPING PREDICTIVE MODELS FOR DISEASE ONSET

Robert H. Podolsky* and Jin-Xiong She

High-throughput approaches such as microarrays are being used in attempts to discover disease biomarkers, including markers of disease onset. While much work has focused on multivariate classification methods for $p >> n$, two issues have received much less attention. First, any model identified may represent a false discovery for several reasons, and not necessarily due to overfitting. Second, the development of prognostic markers requires prospectively measured subjects, a study design that is cost prohibitive in high-throughput studies. Here, we focus on one solution to developing multivariate models as markers for disease onset. This solution involves an aspect of experimental design and incorporates ensemble classification methods in which the predictions of multiple "weak" multivariate classification models are averaged. The predicted error rates of the averaged models compares well with prediction analysis for microarrays (PAM), a method that has been shown to perform well with microarray data. One advantage to this method is that the identified models represent candidate models that can be further evaluated, and removing models does not necessarily reduce classification accuracy. This approach is applied to a cross-sectional study with a unique design modification aimed at identifying biomarkers for the onset of type 1 diabetes, with promising results.

MULTI-MODEL ENVIRONMENT AS RATIONAL APPROACH FOR DRUG DESIGN – AN EXPERIENCE WITH CP-MLR

Yenamandra S. Prabhakar

In isolation, a data point is only a qualified number. A collection of such qualified numbers makes a variable or descriptor. In models, each and every variable communicate with all other variables. A meaningful inter- and intra-variables communications result in the evolution of models with predictive and diagnostic value. In this connection, application of quantum chemical and graph theory to chemical structures has result in several hundreds of descriptors to characterize the molecule from different perspectives. Many of these structure indices carry different information. They have a significant role in providing direction to the design of chemotherapeutic agents. Also, when dealing with large number of descriptors in the modeling studies, for the optimum utilization of contents of the generated datasets, it is necessary to identify different models as well as information rich descriptors corresponding to the phenomenon under investigation. More over, in QSAR studies, each model may address

different sub-structural regions and attributes in the predictive and diagnostic aspects of the chosen phenomenon. A study of population of such models provides scope to understand the diagnostic aspects of different sub-structural regions and in averaging and extrapolating the predictive aspect beyond the individual models. Combinatorial Protocol in Multiple Linear Regression (CP-MLR) (Y.S. Prabhakar, *SAR Comb. Sci.* **22**, 583-595, 2003) is a variable selection approach and generates multiple QSAR models to address the structure-activity relations in terms of different sub-structural regions and attributes in predicting the activity. Experiences of some recent QSAR models generated using the CP-MLR approach will be discussed.

SMALL AREA ESTIMATION WITH AUXILIARY SURVEY DATA

N.G.N. Prasad

Statistical agencies like Statistics Canada and U.S. Bureau of the Census conduct large scale surveys to provide accurate statistics at the national level. However, data from such surveys are being used to provide estimates for smaller domains such as provinces, states, or different racial and ethnic subgroups. In view of small sample sizes in such small domains, standard procedures adopted to give estimates at the national level may produce unacceptable standard errors for the estimates for the domains. In such situations to improve estimates for small domains auxiliary variables from administrative records are often used as covariates. This article considers multivariate methods to borrow information from several sample surveys based on stratified random sampling design to improve estimates for smaller domains. Results from a simulation study will be presented.

RESPONDENT-GENERATED INTERVALS IN SAMPLE SURVEYS

S. James Press and Judith M. Tanur⋆

This article brings together research on the Respondent-Generated Intervals (RGI) approach to recall in factual sample surveys. Additionally presented is new research on the use of RGI in opinion surveys. The research combines Bayesian hierarchical modeling with various cognitive aspects of sample surveys.

ESTIMATION OF THE MULTIVARIATE BOX-COX TRANSFORMATION PARAMETERS

Mezbahur Rahman⋆ and Larry M. Pearson

The Box-Cox transformation is a well known family of power transformations that brings

a set of data into agreement with the normality assumption of the residuals and hence the response variables of a postulated model in regression analysis. This paper implements the multivariate Newton-Raphson method in estimating the transformation parameters and gives a new method of estimation by maximizing the multivariate Shapiro-Wilk statistic. Simulation is performed to compare the two methods in case of the bivariate transformations.

LINEAR REGRESSION FOR RANDOM MEASURES

M. M. Rao

If X, Y are random variables, a classical problem of Ragnar Frisch is to characterize X, Y such that the regression function $E(Y|X) = g(X)$ is linear, $g(X) = aX + b$. This problem has been worked by many people, and solutions were given under various assumptions. A further generalization is considered here. Let $Z(\cdot)$ be a random measure on the class of bounded Borel sets B_0 of the real line R where $Z(A)$ is integrable, and Z is σ-additive with independent values on disjoint sets. Conditions for $E(Z(A)|Z(B)) = aZ(B)$ for all $A, B \in B_0$ are given if the values of $Z(\cdot)$ are α symmetric stable random variables for $1 \leq \alpha \leq 2$ where a depends only on A and B. Moreover extensions to integrals Y^f for continuous f supported by bounded sets of reals, defined by the random measure Z are also obtained.

STATISTICAL EIGEN-INFERENCE FROM LARGE WISHART RANDOM MATRICES

Raj N Rao⋆, Alan Edelman, James Mingo and Roland Speicher

The asymptotic behavior of the eigenvalues of a sample covariance matrix is described when the observations are from a zero mean multivariate (real or complex) normal distribution whose covariance matrix has population eigenvalues of arbitrary multiplicity. In particular, the asymptotic normality of the fluctuation in the trace of powers of the sample covariance matrix from the limiting quantities is shown. Concrete algorithms for computing the limiting quantities and the covariance of the fluctuations are presented using the framework of free probability and second order freeness, respectively.

Tests of hypotheses for the population eigenvalues are developed and a parametric technique for inferring the population eigenvalues is proposed that exploits this asymptotic normality of the trace of powers of the sample covariance matrix. Monte-Carlo simulations are used to demonstrate the superiority of the proposed methodologies over classical techniques and the robustness of the proposed techniques in high-dimensional, (relatively) small sample size settings. Non-parametric extensions are briefly discussed.

The results are obtained by a new construction called second order freeness. Second order freeness extends to fluctuations of random matrices what Voiculescu's first order freeness did for the problem of calculating the eigenvalue distributions of sums and products of random matrices. We will sketch the main ideas of second order freeness and present our second order analogue of the R-transform, which allows us to effectively calculate the fluctuations of A+B from the fluctuations of A and B.

EXACT DISTRIBUTIONS OF TEST STATISTICS AND THE R-FUNCTION

P.N. Rathie

The existing test statistics for the real and complex multivariate normal cases are mentioned along with the work done on the exact and approximate distributions so far. A few new results are also exposed in this paper along with the authors recent results in terms of the R-function, beta and logarithmic series expansions. The test statistics $\Sigma = \Sigma_0, \mu = \mu_0$ and $\Sigma = \Sigma_0$ for the real and complex multinormal situations are especially discussed. A few open problems on complex integration with branch points are pointed out.

In 1989, the author proposed the R-function to obtain computable exact distributions of various test statistics not covered by the generalized hypergeometric functions, such as G and H functions.

In this paper we also indicate a new generalization of the R-function, a detailed study of which will result in the unification of various problems in Statistics and Physics.

GENERATION OF MULTIVARIATE DENSITIES

R N Rattihalli

Consider a p-variate C contoured unimodal density function f with model value 0. By using contour transformation we obtain a new family of C_δ contoured density functions $\{g(x, \delta) : \delta \in \Delta\}$, where Δ is a suitable set of parameters. The density f is a member of this family. Some properties of $g(x, \delta)$ are studied. Further location and scale parameters can be introduced.. Such a model can be used for the analysis of data.

HIERARCHICAL MODAL CLUSTERING
BASED ON THE TOPOGRAPHY
OF HIGH-DIMENSIONAL MIXTURES

Surajit Ray

With the advent of new high throughput technologies in scientific areas such as medical imaging and genomics, the need for efficient data analyses is ever increasing. This necessitates the development of statistical methodology in a non-standard setup, often referred to as the high dimensional low sample size (HDLSS) setup. In this talk I will propose new methodologies for addressing two important aspects of analyzing high-dimensional data: (i) Finding contextually relevant partitions and dimension reduction, (ii) Feature extraction and asymptotics for HDLSS data.

Multivariate mixtures provide flexible methods for both fitting and partitioning high-dimensional data. But in reality the true clusters may not arise from standard statistical distributions. We propose a new concept of modal clusters, where we start from a conventional mixture analysis or a kernel based density estimator, and cluster together those components whose contributions are actually unimodal. Ray and Lindsay (2005) show that the topography of multivariate mixtures, can be analyzed rigorously in lower dimensions by use of a ridgeline manifold that contains all critical points as well as the ridges of the density. Based on this fundamental result, we have developed a comprehensive modal clustering technique, which uses a MM algorithm (generalized version of EM) to find the modes within the ridge line manifold. Additionally, as different levels of smoothing provide different aggregations of data points, modal clustering also lays the foundation for model based hierarchical clustering.

In the last part of my talk I will demonstrate a new technique for feature extraction in HDLSS setup. This is based on the geometry and asymptotics of large dimension and small sample size. Further, based on the special geometric structure imposed by these datasets, we propose a novel technique for model-based bi-clustering. Performance of these two newly proposed methods will be demonstrated through applications on medical image segmentation, clustering of gene expression data and other simulated data sets.

MULTIVARIATE QUALITY MANAGEMENT
INTERMS OF A DESIABILITY MEASURE

Dilip Roy

For a multi-processing system, processing effects may get captured through a vector measure of the quality characteristics of the product. However, when the end user of the product

receives the same use-quality of the product is mostly perceived in terms of a combined level of satisfaction arising out of the individual satisfactions and their interrelationships. In this sense, a user examines the overall utility or the desirability of the product.

The proposed work examines different aspects of the desirability measure and obtains the distribution of a combined measure of desirability, both under null and non-null set-up. The concept of zonal polynomials has been used in this process to present the derived probability functions. Further, suitability of the combined measure has been examined to install a control mechanism for ensuring quality in the production system. Use of the desirability value has also been explored to describe the process of buying from a product category through a binary modelling approach.

THE APPLICATION OF PARTICLE FILTERS FOR TRACKING AND RECOGNITION OF DYNAMICAL EVENTS IN VIDEO

Amit K. Roy-Chowdhury⋆, A. SenGupta, and B. Song

The focus of this paper is on application of particle filters for tracking and recognition of objects in video. Objects can be represented as multi-dimensional vectors comprised of different features, like landmark points, color or contour. Tracking these objects in video requires robustness to occlusions, clutter and sensor noise. Particle filters are particularly suited for this purpose as they can handle non-linear dynamical models describing the motion of the objects and multi-modal noise distributions that arise due to occlusion and clutter.

In this paper, we propose an integrated approach to tracking and recognition of human activities. The tracks, obtained using particle filters, are used to recognize the activity, and the recognized activity drives the tracking in the next frame of the video by specifying the appropriate dynamical model. The recognition is achieved by comparing the tracked feature vectors against learned models for different activities in a Bayesian hypothesis testing scenario. We show some preliminary theoretical results that justify this integrated tracking and recognition approach and demonstrate its promise. Specifically, we show that if the proper activity model is chosen in the recognition phase, the tracking error is minimized using this model. We also present some results of our approach on real-life video sequences.

SPATIO-TEMPORAL MODELING OF OCEAN TEMPERATURE AND SALINITY

Sujit K. Sahu

The world's climate is to a large extent driven by the transport of heat and fresh water

in the oceans. Regular monitoring, studying, understanding and forecasting of temperature and salinity at different depths of the oceans are a great scientific challenge. Temperature at the ocean surface can be measured from space. However salinity cannot yet be measured by satellites, although systems are being developed, and space-based measurements can only ever give us surface values. Until recently temperature and salinity measurements within the ocean have had to come from expensive research ships. The Argo float program, described below, has been funded by various nations around the world to collect actual measurements and rectify this problem.

The primary objective of this paper is to model data obtained from Argo floats in the North Atlantic Ocean during the year 2003. The purpose is to build a model and develop methodology for constructing annual prediction maps at three different depths. In so doing we tackle various modeling and computational issues. The statio-temporal data sets are completely misaligned: no two measurements are recorded at the same location because of the moving floats. Moreover, there is no temporal regularity in the data. As a result many current strategies for modeling time series data from fixed monitoring sites are not appropriate here. In addition it can be expected that the underlying processes are non-stationary and anisotropic in space and time due to variations arising from different sources e.g. latitude and time of the year. Lastly, typical data sets are quite large adding to the computational burden.

In this paper we consider a Bayesian hierarchical model describing the spatio-temporal behavior of the joint distribution of temperature and salinity levels. The space-time model is obtained as a kernel-convolution effect of a single latent spatio-temporal process. Additional terms in the mean describe non-stationarity arising in time and space. We use predictive Bayesian model selection criteria to choose and validate the models. We obtain annual prediction surfaces and the associated uncertainty maps. We develop different models for three different depths of the north Atlantic ocean. The Markov chain Monte Carlo methods are used throughout in our implementation. More work is needed to unify the three models at different depths into a single hierarchical model. (This is joint work with Peter Challoner.)

ASYMPTOTIC THEORY OF SOME STATISTICS IN CANONICAL CORRELATION ANALYSIS WHEN THE DIMENSION AND THE SAMPLE SIZE ARE LARGE

Tetsuro Sakurai

This paper is concerned with asymptotic theory of some basic statistics including (i) canonical correlations, test statistics for (ii) dimensionality and (iii) additional information between $\boldsymbol{x} : p_1 \times 1$ and $\boldsymbol{y} : p_2 \times 1$, based on a sample of size $N = n + 1$. The asymptotic distributions of these statistics have been extensively studied when the dimensions p_1 and

p_2 are fixed and the sample size N tends to infinity. However, these approximations become worse as p_1 or p_2 is large in comparison to N. We derive asymptotic distributions of these statistics when both the dimension and the sample size is large. Let $\boldsymbol{x} = (x_1, ..., x_{p_1})^{\mathrm{T}}$ and $\boldsymbol{y} = (y_1, ..., y_{p_2})^{\mathrm{T}}$ be two random vectors having a joint $(p_1 + p_2)$-variate normal distribution with a mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_x, \boldsymbol{\mu}_y)^{\mathrm{T}}$ and a covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix}, \quad \boldsymbol{\Sigma}_{xx} : p_1 \times p_1, \ \boldsymbol{\Sigma}_{yy} : p_2 \times p_2, \ p_1 < p_2.$$

Let S be the sample covariance matrix formed form a sample of size $N = n + 1$ of $\boldsymbol{x}, \boldsymbol{y}$. Corresponding to a partition of $\boldsymbol{x}, \boldsymbol{y}$ we partition S. Let $\rho_1 \geq \cdots \geq \rho_{p_1}$ and $r_1 \geq \cdots \geq r_{p_1}$ be the population and sample canonical correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$.

The main study for (i) is to derive asymptotic distribution of $f(r_i^2)$ under a high-dimensional framework such that p_1 is fixed, $p_2 \to \infty$, $n \to \infty$, $m = n - p_2 \to \infty$ and $c = p_2/n \to c_0 \in (0, 1)$. Then the result is given by

$$\sqrt{n}\sqrt{1-c}(f(r_i^2 - f(\widetilde{\rho}_i^2)) \to N(0, \sigma_{n,p}(\widetilde{\rho}_i^2)),$$

where " $\to$ " denotes convergence in distribution, $\widetilde{\rho_i^2} = \rho_i^2 + c(1 - \rho_i^2)$. It is assumed that $f(r_i^2)$ be a function r_i^2 such that the third derivarive is continuous at a neighborhood of $r_i^2 = \widetilde{\rho}_i^2$. The effectiveness of our high dimensional results can be demonstrated through simulation experiments. The results for (i) and (ii) are based on a joint work with Professor Y. Fujikoshi.

NONPARAMETRIC INEQUALITY MEASURE BASED ON RANKS

Abhishek Sarkar*, Debdeep Pati, and Anirban Bhattacharya

The aim of the project is to study the presence of inequality among the states of India with respect to several attributes. Given the ranks of 19 states of India on 8 different attributes during a period of time summarized in a 'rank matrix' $\boldsymbol{A}_{19 \times 8}$ whose each column is a permutation of 1 to 19, our problem involves proposing a statistic that would be an indicator of inequality in the country for that period of time. Next, on the basis of data on several time instants, we would like to conclude if there is any trend in inequality amongst the states over the past decade. First of all we tried out several common measures in this regard e.g. the Friedman's statistic, a statistic involving the Cayley's distance, etc., but they reflected the inequality only in a crude way in the sense that they are likely to inflate or deflate mainly due to the inflation or deflation of the inter-column distances of $\boldsymbol{A}$. So we have modified the statistic in such a way that it must represent how the matrix as a whole deviates from the perfect inequality situation i.e., matrices having identical columns. To this end, we

proposed a statistic which represents the minimum of the minimum number of transpositions required to transform our given matrix into a perfect inequality matrix. Clearly, we could interpret smaller values of the statistic as more inequality in the sense that it is closer to perfect inequality situation. Computing the aforesaid statistic for large dimensional matrices through complete enumeration is beyond the scope of a modern computer. Also no closed form expression of our proposed statistic for an arbitrary rank matrix is available. So we took resort to several combinatorial optimization techniques though for two fairly restrictive situations we were able to give an exact solution to the proposed statistic. Again the objective function (i.e. the minimum number of transpositions required to transform the given rank matrix into a perfect inequality matrix) that we are going to minimise is likely to have several local minima and hence standard optimisation techniques like gradient based random search or the Metropolis Hastings algorithm being stuck at a local minima is bound to give inaccurate results. So we used the Simulated Annealing algorithm which is similar to hill-climbing or gradient search with a few modifications. We have particularly chosen the algorithm as it works well with functions having lots of local minima. We got pretty accurate results using simulated annealing. We also obtained the bootstrap distribution of our statistic and as apprehended, we were able to detect high inequality amongst the states of India in all the four rounds of survey conducted by the NSSO.

CONSTRAINED STATISTICAL INFERENCE FOR HIGH-DIMENSION LOW SAMPLE SIZE MODELS AND ROY'S UNION-INTERSECTION PRINCIPLE

Pranab Kumar Sen

In high dimension (K) low sample size (n) environments $(K >> n)$, often, nonlinear, inequality, order, or general shape constraints crop up in rather complex ways. As a result, optimal statistical procedures based on the likelihood principle (LP) may not exist in closed or manageable forms or may not even exist. While some of these complex statistical inference problems can be treated in suitable asymptotic setups, the curse of dimensionality (i.e., the $K >> n$ environment), with often n small, calls for a different asymptotics route (wherein K is indefinitely large but n fixed), having anitipated different statistical perspectives. S.N. Roy (1953) introduced the union-intersection principle (UIP) which has its genesis in the LP and is more flexible and amenable in the $K >> n$ environment. This scenario is appraised with a few important applications in neuronal spike-train models and genomics. Nonstationarity as well as likely lack of spatial independence along with the dimensional constraint $K >> n$ create impasses for adoption of standard robust statistical inference, and hence, some alternative robust inference procedures based on the UIP are considered.

ESTIMATION OF INTEGRATED COVOLATILITY FOR ASYNCHRONOUS ASSETS IN THE PRESENCE OF MICROSTRUCTURE NOISE

Rituparna Sen

Measurement of integrated covolatility is very important in finance for various reasons like portfolio management, calculating value at risk and hedge ratios, value derivative products that depend on two or more asets, etc. When data on 2 securities in finance are observed non-synchronously (eg trades take place at random times which do not coincide and we are using 5 min data that report the last observed trade), then the realized covariation (RC) measure becomes smaller as we increase the sampling frequency. This was observed empirically by Epps (1976). Hayashi and Yoshida (2004) propose an alternative estimator that takes care of this problem using tick-by-tick data. However in the presence of microstructure noise, this estimator can have very high variance. Mykland and Zhang (in preparation) advocate the use of RC as the estimator in the presence of microstructure noise but under very restrictive assumptions to avoid the problem of nonsynchronicity. In statistical terms, for high frequncy asynchronous data with microstructure noise, the Hayashi-Yoshida estimate (HY) has no bias but very high variance while RC has high bias. We propose an estimator which equals RC for low frequency and asymptotically equals HY for high frequency. Then we can choose an optimal frequency by some bias-variance trade-off criterion. Since this estimator is not based on tick-by-tick data (unlike HY), it renders itself to variance reduction techniques, like subsampling. We shall present the problem, properties of our estimator and some simulation and real data examples.

A NEW REGRESSION MODEL FOR CIRCULAR VARIABLES WITH APPLICATION TO MALAYSIAN WIND DATA

Ashis SenGupta and Abdul Ghapor Hussin*

By a circular random variable, we mean a variable whose values can be mapped to the circumference of a circle. This random variable must be analyzed by techniques differing from those appropriate for the usual linear (defined on the real line) variables because the circumference is a bounded closed space, for which the concept of origin is arbitrary or un-defined. Recently Mardia & Downs (2002) proposed a regression curve to model relationship between two circular variables jointly distributed on the torus. We discuss two regression models for circular variables. The first directly uses regression with circular variables. The second model uses an approach based on some link function. We also consider the relevance of circular normal (or, equivalently, von Mises) conditionals distribution for this framework.

Parameter estimates are obtained by minimizing circular errors, instead of the least-squared method usually employed for linear data. A measure of fit is provided based on the mean circular error. Our approach is illustrated with an application to the modeling and analysis of the Malaysian wind direction data recorded at two different locations, the south and the north Malaysia, during the southwest monsoon season.

THE MULTIVARIATE TUKEY-KRAMER TYPE MULTIPLE COMPARISON PROCEDURES AMONG MEAN VECTORS

Takashi Seo

In this paper, conservative simultaneous confidence intervals for multiple comparisons among mean vectors in multivariate normal populations are considered. As the conservative procedure, the multivariate Tukey-Kramer type multiple comparison procedures for pairwise comparisons and for comparisons with a control are presented. In particular, the affirmative proof of the multivariate generalized Tukey conjecture in the case of four mean vectors is presented. Further, the upper bounds of coverage probabilities for the conservativeness of the multivariate Tukey-Kramer type procedures are also given. Finally, numerical results by Monte Carlo simulations are given for some selected values of parameters.

ON A CLASS OF PBIB DESIGNS USEFUL IN SAMPLING

John Stufken

The class of block designs now known as polygonal designs has received attention in the literature due to its connection to Balanced Sampling Plans for Excluding Contiguous Units, which were introduced in Hedayat, Rao, and Stufken (JSPI, 1988). A polygonal design for ν varieties, k blocks ($k < \nu$) and minimum distance $\alpha + 1$ is a binary block design in which varieties i and j appear together in precisely λ blocks, say, if $\alpha < |i-j| < \nu - \alpha$, while they do not appear together in any block otherwise. For $\alpha = 0$, this class of designs consists precisely of the class of balanced incomplete block designs, while the designs are partially balanced incomplete block designs for $\alpha > 0$. The presentation will provide a brief overview of our current knowledge of these designs. The focus will be on the existence and construction of polygonal designs for $\alpha > 0$, on extensions to higher dimensions, and on open problems.

CLASSIFICATION METHOD FOR DIRECTIONAL DATA WITH APPLICATION TO NIGERIAN SKULL

Fidelis I. Ugwuowo

There are many well-known methods applied in classification problems for linear data with both known and unknown distributions. Here we deal with classifications involving data on torus or cylinder. A new method involving a generalized likelihood ratio test is developed for classification in two populations using such directional data. The approach assumes that one of the probabilities of misclassification is known. The procedure is constructed by applying Gibbs sampler on the conditionally specified distributions. A parametric bootstrap approach is also presented. An application to data involving linear and circular measurements on human skulls from two tribes in Nigeria is given.

STOCHASTIC MODEL OF BIRTH WEIGHT IN CHILD HEALTH

K. Senthamaraikannan and D. Nagarajan*

The birth weight is one of the important factors of child health, because growth is a good indication of a newborn's general health. A child who is growing well is generally healthy, while poor growth may be a sign of a problem. The attentions of researchers in recent years are focused about the birth weight of child health. This paper analyses the stationary distribution of birthweight using the spectral expansion of the transition probability matrix.

ML ESTIMATION IN A VISUAL ACUITY MODEL WITH MISSING OBSERVATIONS

Kunio Shimizu*, Masanori Okamoto, and Mihoko Minami

Olkin and Viana (1995) studied ML and large-sample estimates for a model in which (X, Y_1, Y_2) are jointly normally distributed with (Y_1, Y_2) exchangeable and (X, Y_i) having a common correlation. The model was applied to the analysis of visual acuity data. Here each Y indicates a single measurement of visual acuity in each eye and X age as an additional measurement. We study ML and REML estimates for the model with missing observations. Missing patterns considered are models of trivariate X-missing, bivariate Y-missing and trivariate Y-missing. Asymptotic variance stabilizing transformations are given and simulation study is performed to evaluate large-sample variances.

ON SOME ASPECTS OF DIALLEL CROSS DESIGNS WITH CORRELATED STRUCTURE

Karabi Sinha

The purpose of this work is to present some theoretical results for analysis of diallel cross designs in blocked situations with a possible correlation structure within each block.

TRIMMED ANOVA AND ITS MULTIVARIATE EXTENSION

Deo Kumar Srivastava⋆ and Govind S. Mudholkar

It is well known that in case of non-normal multivariate populations Hotelling's T^2 and the normal theory multivariate analysis of variance (MANOVA) procedures are unsatisfactory with respect to both the type I error control and power properties; e.g. see Seber (1984), Mudholkar and Srivastava (2000, 2001). In this talk we first outline development of trimmed-ANOVA, which is a robust modification of the analysis of variance for comparing k means based on the trimmed means in place of the means. This robust ANOVA procedure is then integrated in the J Roy's stepwise approach as modified by Mudholkar and Subbaiah (1980) to develop robust alternatives to the normal theory MANOVA procedure. The operating characteristics of the proposed robust procedures are evaluated using extensive Monte Carlo simulations. The robust procedures demonstrate reasonable type I error control and enhanced power at nonnormal alternatives with minimal loss of power in the presence of normality. The robust MANOVA procedure also readily yields multiple comparisons and simultaneous confidence intervals. The new MANOVA procedure utilizes relatively familiar univariate tests and avoids additional distributional problems.

ANALYZING MULTIVARIATE DATA WITH FEWER OBSERVATIONS THAN THE DIMENSION

Muni S. Srivastava

In DNA microarray data, gene expressions are available on thousands of genes of an individual but there are only few individuals in the dataset. Although, these genes are correlated, most of the statistical analyses carried out in the literature ignore this correlation without any justification. For example if in one-sample problem it is found that the data supports the sphericity hypothesis about the covariance matrix Σ, that is $\Sigma = \sigma^2 I_p$, for some unknown σ^2 and $p \times p$ identity matrix I_p, then any inference on the mean vector may ignore the correlations between the genes, and use the univariate methods. However, it has

not been done in the literature. Similarly, if the covariance matrix is a diagonal matrix, the univariate methods with unequal variances for the componets may be used for any inference. For example, Dudoit et al. (2002) assumed that the covariance matrix is a diagonal matrix in their classification procedure without verifying that the data support this assumption. In fact, it is shown in Srivastava (2006) that the data do not support this assumption.

Throughout this article, we shall assume that the sample size N is less than the dimension p. This in turn implies that no tests invariant under the nonsingular linear transformations exist for testing the hypothesis on the mean vector in one-sample, and mean vectors in two-sample and many samples, the so called MANOVA problem (Lehmann, 1959, p. 318). Similarly, likelihood ratio tests for the hypothesis on the covariance matrix or the covariances matrices in two or more than two populations do not exist. Thus various test criteria have been recently proposed to verify the assumptions made on the covariance matrix or matrices in two or more than two populations. Similarly, several test criteria have been proposed for the inference on the mean vectors. In this article, we review these procedures.

APPLICATION OF MULTIVARIATE METHODS IN MARKETING RESEARCH

T. Srivenkataramana

Marketing Research (M. R.) is an objective and formal process of systematically obtaining, analyzing and interpreting data for actionable decision making in marketing. Many M.R. investigations wish to determine (a) Association among several variables (b) Predictability of dependent variables (c) existence of cluster patterns etc. The present paper mentions a few novel problems in the M. R. Scenario and discusses applicability of multivariate methods for the same.

DISTRIBUTION OF EIGENVALUES AND EIGENVECTORS OF WISHART MATRIX WHEN THE POPULATION EIGENVALUES ARE INFINITELY DISPERSED

Akimichi Takemura[*] and Yo Sheena

We consider the asymptotic joint distribution of the eigenvalues and eigenvectors of Wishart matrix when the population eigenvalues become infinitely dispersed. We show that the normalized sample eigenvalues and the relevant elements of the sample eigenvectors are asymptotically all mutually independently distributed. The limiting distributions

of the normalized sample eigenvalues are chi-squared distributions with varying degrees of freedom and the distribution of the relevant elements of the eigenvectors is the standard normal distribution.

We also discuss approximations to the distribution of the eigenvalues of Wishart distribution based on tube formula approach.

EVALUATION AND COMPARISON OF GENE CLUSTERING METHODS IN MICROARRAY ANALYSIS

Anbupalam Thalamuthu, Indranil Mukhopadhyay*,
Xiaojing Zheng and George C. Tseng

Microarray technology has been widely applied in bio-logical and clinical studies for simultaneous monitoring of gene expression in thousands of genes. Gene clustering analysis is found useful for discovering groups of correlated genes potentially co-regulated or associated to the disease or conditions under investigation. Many clustering methods including hierarchical clustering, K-means, PAM, SOM, mixture model-based clustering, and tight clustering have been widely used in the literature. Yet no comprehensive comparative study has been performed to evaluate the effectiveness of these methods.

In this paper, six gene clustering methods are evaluated by simulated data from a hierarchical log-normal model with various degrees of perturbation as well as four real data sets. A weighted Rand index is proposed for measuring similarity of two clustering results with possible scattered genes (i.e. a set of noise genes not being clustered). Performance of the methods in the real data is assessed by a predictive accuracy analysis through verified gene annotations. Our results show that tight clustering and model-based clustering consistently outperform other clustering methods both in simulated and real data while hierarchical clustering and SOM perform among the worst. Our analysis provides deep insight to the complicated gene clustering problem of expression profile and serves as a practical guideline for routine microarray cluster analysis.

MULTIVARIATE GEOMETRIC DISTRIBUTION–PROPERTIES AND RELATED INFERENCE PROBLEMS

R. Vasudeva

In this paper we introduce Multivariate Geometric Distribution and establish some properties. We also discuss the problems of estimation, testing, classification etc.

ITERATIVE BIAS CORRECTION ON CROSS-VALIDATION

Hirokazu Yanagihara⋆ and Hironori Fujisawa

It is known that the cross-validation (CV) criterion is a second-order unbiased estimator of the risk between the candidate model and the true model. In this talk, we show that a 2k-th higher-order unbiased estimator can be proposed by using a liner combination of the leave-one-out,...,leave-k-out CV criteria. The distinguishing point is that the proposed idea can give a smaller bias than a jackknife method without any analytic calculation. We verify by numerical experiments that the proposed estimator has a smaller bias than other criteria including the generalised information criterion (GIC), the extended information criterion (EIC), the ordinary CV criterion, and so on.