Book of Abstracts*

2011 IISA Conference on Probability, Statistics, and Data Analysis

April 21 – 24, 2011
Raleigh, NC, USA

*Last revised on: April 19, 2011
Predicting response rates in display advertising [I.4.5]

Deepak Agarwal

Statistics/Machine Learning
Yahoo!

Abstract. Display advertising is an attractive alternative to traditional media like TV, Newspapers and magazines for marketers to target the audience of their choice. The intent varies – brand awareness/reach could be the primary goal of some big advertisers, immediate sales/conversions may be of paramount importance to others. An ad exchange provides an efficient and unified marketplace to connect buyers (advertisers) and sellers (publishers,) possibly through commercial intermediaries like ad-networks or publisher networks. One key statistical challenge for proper functioning of such an exchange involve estimating response rates of rare events like click on an ad, user buying a product on publisher page and so on. I will present a statistical model that estimates such response rates through a multi-level hierarchical model. Our model performs shrinkage by leveraging data aggregates along multiple hierarchies and scales to massive data sets in a map-reduce framework (billions of observations, hundreds of millions of covariates).

A class of tests for exponentiality against NBUE alternatives [C.7.7]

M Z Anis

SQC & OR
Indian Statistical Institute

Abstract. In this paper we develop a class of test statistics for testing exponentiality against NBUE alternatives. The test statistics are shown to be asymptotically normal and consistent. This class of test statistics includes the test proposed by Hollander and Proschan (1975) as a special case. Efficiency studies have also been done.

Quality improvement through third-order slope-rotatable designs over all directions [C.7.5]

G V S R Anjaneyulu

Statistics
Acharya Nagarjuna University, Guntur, AP, India

Abstract. Recent years have been an upsurge of interest in the role of statistical ideas and methods in improving the quality and productivity of industrial processes and products. Response Surface Methodology is, one such idea, useful for analyzing problems where several independent variables influence a dependent variable or response. The earlier study of response surface designs mainly emphasized the estimation of absolute response. Estimation of differences in response at different points in the factor space will often be of great importance. If differences at points closed together are involved, estimation of local slope i.e., the rates of change of the response surface is of interest. This problem, estimation of slopes, occurs frequently in practical situations, particularly in third order response surface. This enable the determination of the best operating conditions for the process i.e., the best com-
bination of the levels of the controllable factors which gives the optimum value of the third order response function. This would also enable us to determine the best way to control the process. In this paper, we made an attempt to study the role of third order slope rotatability to improve the quality of a product. Anjaneyulu et al (1993) introduced embedding in Second Order Slope Rotatable Designs and constructed the same using embedding techniques similar to those of Draper (1960), Herzberg (1967). Park (1987) studied the necessary and sufficient conditions for second order slope rotatability over all directions, Anjaneyulu et al (1997) showed that these designs have the Variance-Sum Property. Anjaneyulu et al (1995) introduced TOSRD and gave a method of construction of TOSRD using Central Composite type design points. Anjaneyulu et al (1995) introduced Third Order Slope Rotatable Design over all directions introduced by Park and Lee (1995). In this paper, an attempt is made to explain the role of embedding in TOSRDOAD for Quality Improvement.

Censored measurements and measurement errors in exposure measurements in environmental health studies [I.7.4]

Srikesh Arunajadai

Biostatistics
Columbia University

Abstract. In this talk we will explore two frequently encountered scenarios in environmental health studies. The first is the rather prominent Limit of Detection problem. In the environmental health sciences, measurements of toxic exposures or other contaminants are often constrained by a lower limit called the limit of detection (LOD). That is, one cannot measure the concentration of the toxicant with certainty below the limit of detection. In spite of reasonable warnings in the literature against their use, the non-detect exposures are often substituted with a value equal to LOD/2 or LOD/√2. This is probably because the authors who caution against the use of such substitutions often imply that they are acceptable under low proportions of censoring. In this talk we will explore the effects of such substitutions on the exposure effect in a regression model under various conditions and propose a multiple imputation based method for estimating the exposure effect. Next we explore the issue associated with urine based exposure measurements. Exposure measurements based on urine samples are affected by the dilution of urine, which varies from one patient to another. The measurement error modeling literature has never been considered to deal with urinary exposure measurements. Both toxicologists and environmental health scientists employ a concentration adjustment formula proposed nearly fifty years ago to adjust for the dilution effects. Here we discuss the effect of this adjustment on the exposure effect and propose a measurement error model based approach to estimate the exposure effect. Simulations are performed in both cases to compare the effects of the various substitutions and adjustments on the estimated exposure effect.

Testing for mean in high dimensional Gaussian time series [I.6.2]

Deepak Nag Ayyala

Department of Mathematics and Statistics
University of Maryland, Baltimore County

Abstract. In this talk, we address the problem of testing for mean vector when the data has a Multivariate Gaussian distribution. Significance of the test is emphasized by the fact that the dimension can be larger than the sample size, i.e. a 'large p and small n' problem. Several tests have been proposed when the observations are independent but they do not address dependence. The proposed test statistic is observed to perform better than the existing tests when dependence is considered and reduces to one of the existing tests in case of independence. Power study is also shown to support our results.

Burgers equation with Poissonian forcing [I.3.4]

Yuri Bakhtin

School of Mathematics
Georgia Tech

Abstract. The Burgers equation is a nonlinear hydrodynamic model describing the evolution of the velocity field of sticky dust. The particles in this kind of medium interact only when they hit and stick to each other forming clumps. Some ergodic properties of this system with white-noise forcing and mostly in compact domains are known, but there are several interesting unanswered questions, especially for unbounded domains. In this talk a new simpler model for forcing based on Poissonian point field is proposed. The advantage of this model is that although it preserves many characteristic features of the white-noise model, it is easier to work with and visualize the resulting behavior. In fact, the model can be studied by looking at optimal paths through the Poissonian environment. In the unbounded domain case, if the spatial component of the measure driving the Poisson process has finite first moment, we obtain ergodic results for this model: one force-one solution principle; existence, uniqueness and some properties of a global skew-invariant solution including its behavior at infinity and a description of its basin of forward and pullback attraction; existence and uniqueness of a stationary distribution. Even for the Burgers equation on the circle this model provides a new insight into the behavior of the global minimizer.

Bayesian hierarchical functional models for high-dimensional genomics data [I.2.2]

Veera Baladandayuthapani

Biostatistics
MD Anderson Cancer Center

Abstract. Recent advances in genomic profiling techniques such as array-based comparative genomic hybridization (array-CGH) and SNP experiments provide a high-throughput, high-resolution method to measure relative changes in DNA copy number. These experiments typically yield data consisting of profiles of fluorescence intensity ratios of test and reference DNA samples across the whole chromosomal map. One of the goals of the analysis is characterization of these profiles into calling gains (amplifications) or losses(deletions) in copy numbers. These amplifications and deletions at the DNA level are important in the pathogenesis of cancer and other diseases. We present a class of Bayesian models to model-
ing these genomic profiles using functional data analysis methods, which not only accounts for correlation of markers along the genome but also allows for subject specific deviations to answer relevant biological questions. The methods are illustrated using several real and simulated datasets.

**Nonparametric spatial models for periodontal disease data with spatially-varying non-random missingness** [I.5.5]

Dipankar Bandyopadhyay

*Division of Biostatistics and Epidemiology, Dept. of Medicine*
*Medical University of South Carolina*

Abstract. Clinical attachment level (CAL), defined as the distance down a tooth’s root that is detached from surrounding bone, measures the current state of a subject’s periodontal disease. Specifically, it is measured at 6 locations per tooth throughout the mouth (excluding the molars) during a typical periodontal exam, giving rise to a clustered set-up. It is common to summarize a subject’s disease status using a one-number summary of the CAL measurements, such as the whole-mouth average or the number of observations greater than a threshold, and then use these summaries in a least squares regression to identify important subject-level fixed effects. However, analysis of CAL data comes with several interesting challenges. CAL is non-stationary, having a symmetric distribution with small variance in the front of the mouth and a skewed distribution with large variance in the back of the mouth. Also, absent (missing) teeth are common and cannot be considered ‘missing at random’ since periodontal disease is the leading cause of adult tooth loss.

In this talk, we propose a flexible model which allows for a different, non-Gaussian CAL distribution at each location in the mouth. Rather than a simple regression on a one-number summary, we analyze the data for each subject to exploit the spatial structure of the data to improve upon the fixed-effect estimates. We also jointly model the observed CAL and location of an absent tooth in a way that allows for complex missing data mechanisms and the nature the missing data mechanism to vary in different regions of the mouth.

This is joint work with Brian Reich and Howard Bondell.

**Analysis of Sabine river flow data using semiparametric spline modeling** [I.1.4]

Soutir Bandyopadhyay

*Mathematics*
*Lehigh University*

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

**Efficient Gaussian process for large data sets** [Poster]

Anjishnu Banerjee

*Dept Of Statistical Science*

*Duke University*

Abstract. Gaussian processes (GPs) are widely used in nonparametric regression, classification and spatio-temporal process modeling, motivated in part by a rich literature on theoretical properties. However, a well known drawback of GPs that limits their use in many applications is the very expensive computation involved, requiring $O(n^3)$ computations in performing the necessary matrix inversions with $n$ denoting the number of data points. In large data sets, data storage and processing also lead to computational bottlenecks and numerical stability of the estimates and predicted values degrades with $n$. To address these problems, a rich variety of methods have been proposed, with recent options including predictive processes in spatial data analysis and subset of regressors in machine learning. The underlying idea in these approaches is to use a subset of the data, leading to questions of sensitivity to the subset and limitations in estimating fine scale structure in regions that are not well covered by the subset. Motivated by the literature on compressive sensing, we propose an alternative random projection of all the data points onto a lower-dimensional subspace. We demonstrate the superiority of this approach from a theoretical perspective and through the use of simulated and real data examples.

**Inference for change-point parameters under varying degrees of model misspecification** [1.4.4]

Moulinath Banerjee

*Statistics*

*University of Michigan*

Abstract. We introduce and study a formulation of the problem of estimating change-point models under mis-specification, where the level of mis-specification is allowed to change with the sample size at various different rates, in a standard regression model with i.i.d. data. We study what limit distributions ensue under what rates. Three non-standard limits arise: at rates slower than $n$, one obtains Chernoff’s distribution, at faster rates the minimizer of a standard two-sided compound Poisson process, and at rate $n$, a new limit distribution resembling a compound Poisson process that depends on intrinsic parameters driving the underlying sequence of models. Some connections between the three limiting regimes as well as power approximations under a fixed smooth alternative are also discussed.

**Feature extraction and model based clustering in cbir** [Poster]

*Banerjee*
Sayantan Banerjee

Statistics
North Carolina State University

Abstract. Content Based Image Retrieval (CBIR) is a system which is widely used in retrieving similar images from a large database of such on the basis of features extracted from the images themselves. CBIR systems are in much demand for accessing and organizing images in an efficient way using the information contained in image databases like medical images, satellite images to name a few. A feature extraction technique called the multi-spectral method is applied which extracts a set of color-texture based features from a database of 2000 pre-classified images. Model based clustering with Gaussian mixtures is used for our image retrieval first using the full feature set and then reducing the feature space by generating a sparse matrix. Hierarchical agglomeration techniques, EM algorithm and Bayes factors are used for proper model selection. The utility of the extracted feature set and the performance of retrieval technique proposed is compared with a previously applied technique on the same database and also the procedure is run on another database of 1000 images.

Inference for Birnbaum-Saunders distribution [C.7.7]
Prasanta Basak

Math & Statistics
Penn State Altoona

Abstract. Birnbaum-Saunders distribution has been used widely as a model in analyzing lifetime data. Some work has been done in the past on the estimation of parameters of two-parameter (BS2) and three-parameter Birnbaum-Saunders (BS3) distributions based on complete and censored samples. In this article, we develop estimation methods based on progressively Type-II censored samples from a BS3 distribution. In particular, we use EM-algorithm as well as some other numerical methods to determine maximum likelihood estimates (MLEs) of the parameters. The asymptotic variances and covariances of the MLEs from the EM-algorithm are derived by using the missing information principle. The inferential methods developed are then illustrated with some numerical examples. We also consider the interval estimation based on large-sample theory and examine the actual coverage probabilities of these confidence intervals in case of small samples by means of Monte Carlo simulations.

Space-time data fusion under error in computer model output: an application to modeling air quality [I.5.5]
Veronica J. Berrocal

Department of Biostatistics
University of Michigan

Abstract. In the last two decades a considerable amount of research effort has been devoted to modeling air quality with public health objectives. Conclusions on the relationship between exposure to air pollutants and adverse health outcomes are often contrasting, highlighting the importance of good data quality for both health outcomes and environmental exposure.
In this paper, we propose two modeling approaches to combine monitoring data with numerical model output, addressing the difference in spatial scale between the two sources of data, yielding improved prediction of exposure at point level. Extending our previous downscaler model (Berrocal et al., 2010), these new models are intended to address two potential concerns with the model output. One is potential spatial displacement in the computer model values assigned to a grid cell. Possibly, this output is appropriate for a displacement of the grid cell. The second recognizes that, with regard to improving predictive performance of the fusion at a location, there may be useful information in the outputs for grid cells that are neighbors of the one in which the location lies. The first model is a Gaussian Markov random field smoothed downscaler model that relates monitoring data and computer model output via the introduction of a latent Gaussian Markov random field linked to both sources of data. The second model is a smoothed downscaler model with spatially varying random weights defined through a latent Gaussian process that allows us to smooth the computer model output accounting for possible directionality. We applied both methods to daily ozone concentration data for the Eastern US during the summer months of June, July and August 2001, obtaining a 10%-15% predictive gain in predictive mean squared error over our earlier downscaler model (Berrocal et al., 2010).

Exact exponential bounds for generalized empirical likelihood [I.3.1]
Patrice Bertail

MODAL X et CREST
University Paris-Ouest

Abstract. A possible interpretation of empirical log-likelihood ratio is to see it as the minimization of the Kullback distance between the empirical distribution of the data \( P_n \) and a measure (or a probability measure) \( Q \) dominated by \( P_n \), under linear or non-linear constraints imposed by the model. The use of others metrics instead of the Kullback distance has been suggested by Owen (1990) and many other authors. For example, the choice of relative entropy has been investigated by DiCiccio & Romano (1990), Jing & Wood (1995) and led to "Entropy econometrics" in the econometric fields. Some results have been also obtained with Cressie-Read discrepancies (Baggerly 1998, Corcoran 1998, Newey & Smith, 2003) and led to some econometric extensions known as “generalized empirical likelihood”. Bertail, Harari & Ravaille (2005) have shown that Owen’s (1988) original method in the case of the mean can be extended to any regular convex statistical distance or phi-divergence, satisfying some natural conditions. We show that this method enjoy the same asymptotic properties as empirical likelihood for a large class of Hadamard differentiable functionals extending works of Bertail (2007). Among such divergence, a particular case has been considered in convex analysis (leadind to so called log-proximal methods). We call it quasi kullback divergence : it corresponds to a phi-divergence with phi defined as the Legendre transform of \( \epsilon x^2 + (1-\epsilon)(-x - \log(1-x)) \) for \( \epsilon \) small. We show that for this divergence it is possible to choose \( \epsilon \) so as to get Bartlett correctability but with better coverage probabilities for small \( n \) than for empirical likelihood. Moreover it is possible to get for this family of divergence exact exponential bounds, in the case of the mean, by relating the generalized empirical likelihood maximization program to the behavior of an autormalized sum. These
bounds are distribution free if the underlying distribution of the variable are symmetric and
depends on the kurtosis of the distribution in the general case. We illustrate the results with
some simulations.

**An adaptive particle allocation scheme for off-line sequential importance sampling algorithms** [I.5.2]

Anindya Bhadra

*Statistics*

*Texas A&M University*

Abstract. In presence of extreme observations, evaluation of conditional likelihood by a
Sequential Monte Carlo filter (also known as a particle filter) suffers from high variance at
the corresponding time points, resulting in poor Monte Carlo estimate of the overall like-
lihood. This paper develops a novel strategy for particle allocation for off-line Sequential
Monte Carlo (SMC) based filters, in order to reduce the overall variance of the likelihood
estimate. The complications arising from the intractability of the actual SMC variance is
handled indirectly by modeling the conditional log likelihood of the observations as an au-
torgressive process. We demonstrate numerical results on both simulated and real data sets
where adaptive particle allocation results in close to 50 overall variance over the naive equal
allocation of particles at all time points. The technique presented in this article is quite
general and applicable to many off-line SMC based optimization techniques.

**Two philosophies for random graphs and networks: local weak convergence and scaling limits** [I.1.2]

Shankar Bhamidi

*Statistics and Operations Research*

*UNC Chapel Hill*

Abstract. The last few years have witnessed an explosion in the number of mathematical
models for random graphs and networks, as well as models for dynamics on these network
models. In this context I would like to exhibit the power of two well known philosophies in
attacking problems in random graphs and networks:

(i) Local weak convergence: The idea of local neighborhoods of probabilistic discrete
structures (such as random graphs) converging to the local neighborhood of limiting infinite
objects has been known for a long time in the probability community and has proved to be remarkably effective in proving convergence results in many different situations. Here
we shall give a wide range of examples of the above methodology. In particular, we shall
show how the above methodology can be used to tackle problems of flows through random
networks, where we have a random network with nodes communicating via least cost paths to
other nodes. We shall show in many popular models this methodology allows us to prove the
convergence of the empirical distribution of edge flows, exhibiting how macroscopic order
emerges from microscopic rules. Also, we shall show how for a wide variety of random
trees (uniform random trees, preferential attachment trees arising from a wide variety of
attachment schemes, models of trees from Statistical Physics etc) the above methodology

*Bhadra* 8  
*Bhamidi*
shows the convergence of the spectral distribution of the adjacency matrix of these trees to a limiting non random distribution function.  

(ii) Scaling limits: For the analysis of critical random graphs, one often finds that properly associated walks corresponding to the exploration of the graph encode a wide array of information (including the size of the maximal components). In this context we shall extend work of Aldous on Erdos-Renyi critical random graphs to the context of inhomogeneous random graph models. If time permits we shall describe the connection between these models and the multiplicative coalescent, arising from models of coagulation in the physical sciences.

**Simplex factor models for multivariate unordered categorical data**

[I.5.7]

Anirban Bhattacharya

*Statistical Science*

*Duke University*

Abstract. Gaussian latent factor models are routinely used for modeling of dependence in continuous, binary and ordered categorical data. For unordered categorical variables, Gaussian latent factor models lead to challenging computation and overly complex modeling structures. As an alternative that is more natural for unordered categorical data, we propose a novel class of simplex factor models. In the single factor case, the model treats the different categorical outcomes as independent with unknown marginals. The model can characterize highly flexible dependence structures parsimoniously with few factors, and as factors are added, any multivariate categorical data distribution can be accurately approximated. Using a Bayesian approach for computation and inferences, a highly efficient Gibbs sampler is proposed that scales well with increasing dimension, with an adaptive Gibbs step enabling selection of the number of factors. Theoretical properties are described and we evaluate the approach through simulation examples. Applications are described for modeling dependence in nucleotide sequences and prediction from high-dimensional categorical features.

**The order-restricted homogeneous RC model for ordered contingency tables: estimation and testing for fit**

[I.1.3]

Bhaskar Bhattacharya

*Mathematics*

*Southern Illinois University Carbondale*

Abstract. The RC model is well-known for modeling association in an ordered contingency table. We consider a square contingency table when there is a one-to-one correspondence between the categories of the row and column variables and the parametric scores reflect the ordinality of the categories. The maximum likelihood estimates of these scores are found. Testing for the order-restricted homogeneous RC model for fit is discussed. Two well-known data sets are used to illustrate the proposed procedure. Simulation is used to compare the performance of the proposed procedure with a benchmark.

**Nonparametric statistics on manifolds-by examples and applications**

Bhattacharya

9

Bhattacharya
Rabi Bhattacharya

Abstract. The general theory of nonparametric statistics on manifolds M presented here is of recent origin. It builds much of its framework on the notion of the Frechét mean of a probability measure $Q$, namely, the point on the manifold which minimizes the expected squared distance from a random variable with distribution $Q$. The nonparametric methods are intrinsic or extrinsic, depending on the distance used on $M$. The extrinsic distance, which is emphasized here for simplicity, is the distance induced from a good embedding of $M$ in an Euclidean space. In examples, it is often the case that the nonparametric methods yield sharper inference than their parametric counterparts provide. We begin with two of the simplest non-flat manifolds, namely the circle $S^1$ and the sphere $S^2$. An application to paleomagnetism is considered for the analysis of the movements of the earth’s magnetic poles over geological times, and the related area of continental drifts. The majority of our applications here are on landmark based shape spaces. Here a set of $k$ measurements at $k$ appropriate locations of a planar or 3-D object are centered, scaled, and rotated, to define the shape of the object, which is invariant under Euclidean rigid motions as well as scaling. One application is to discriminate between the male and female of a species of mammals by means of shapes of their skulls, when projected on a plane. Because of mathematical and statistical problems associated with the notion of a 3-D shape defined above, a new 3-D shape invariant under all orthogonal transformations is more appropriate, and extrinsic analysis is extended to it. As an application we consider a matched pair problem for the diagnosis of glaucoma based on this notion of shape of the inner eye. Among other applications are digital recognition based on affine shapes, and machine vision, or robotics, using the notion of projective shapes. This talk is based on joint work with Vic Patrangenaru and Abhishek Bhattacharya, and is supported in part by the NSF grant DMS 0806011.

Bayes and the contrapositive [I.7.2]
Sudip Bose

Abstract. Induction problems have been analyzed using probabilistic methods at least since the times of Laplace and Bayes. The work of Broad and Jeffreys included interesting results for finite populations. More recently, Berger, Bernardo and Sun (2009) have considered this problem. We present some recent research related to this area and also discuss this in a two-by-two setting.

Elliott-kalton stochastic differential games associated with the infinity Laplacian [I.3.4]
Amarjit Budhiraja

Abstract. Induction problems have been analyzed using probabilistic methods at least since the times of Laplace and Bayes. The work of Broad and Jeffreys included interesting results for finite populations. More recently, Berger, Bernardo and Sun (2009) have considered this problem. We present some recent research related to this area and also discuss this in a two-by-two setting.
University of North Carolina, Chapel Hill

Abstract. A two-player zero-sum stochastic differential game, defined in terms of an m-dimensional state process driven by a one-dimensional Brownian motion, played until the state exits the domain, is studied. The players' controls enter in the diffusion coefficient and also as an unbounded drift coefficient in the state process dynamics. We show that the game has value, and characterize the value function as the unique viscosity solution of an inhomogeneous infinity Laplace equation. Joint work with R. Atar

Bayesian proportional hazards model for current status data with monotone splines [C.1.1]
Bo Cai

Epidemiology and Biostatistics
University of South Carolina

Abstract. The proportional hazards model is widely used to deal with time to event data in many fields. However, its popularity is limited to right-censored data, for which the partial likelihood is available and the partial likelihood method allows one to estimate the regression coefficients directly without estimating the baseline hazard function. In this paper, we focus on current status data and propose an easy-to-implement Bayesian approach under the proportional hazards model. Specifically, we model the baseline cumulative hazard function with monotone splines leading to only a finite number of parameters to estimate while maintaining great modeling flexibility. An efficient Gibbs sampler is proposed for posterior computation relying on a data augmentation through Poisson latent variables. The proposed method is evaluated by simulation studies and compared to a constrained maximum likelihood method. Uterine fibroid data from an epidemiological study are analyzed as an illustration.

Kernel machine approach to risk prediction with genomewide association studies [I.1.1]
Tianxi Cai

Biostatistics
Harvard School of Public Health

Abstract. As genetic studies of human diseases progress, it is becoming increasingly evident that genetics often play a major and complex role in many types of diseases. Therefore, the complexity of the genetic architecture of human health and disease makes it difficult to identify genomic markers associated with disease risk or to construct accurate genetic risk prediction models. Accurate risk assessment is further complicated by the availability of a large number of markers that may be predominately unrelated to the outcome or may explain a relatively small amount of genetic variation. Often, standard prediction models merely rely on additive or marginal relationships between the markers and the phenotype of interest. Marginal association based analysis has limited power in identifying markers truly associated with disease, resulting in a large number of false positives and false negatives. Simple additive modeling does not perform well when the underlying structure of associa-
tion involves interactions and other non-linear effects. Additionally, these methods do not make use of information that may be available regarding genetic pathways or gene structure. We propose a multi-stage method relating possibly predictive markers to the risk of disease by first forming multiple gene-sets based on certain biological criteria. By imposing a naive Bayes kernel machine model, we estimate gene-set specific risk models that relate information from each gene-set to the outcome. In the second stage, we aggregate information across all gene-sets by adaptively estimating the weights for each gene-set via a regularization procedure. The KM framework efficiently models the potentially non-linear effects of predictors without specifying a particular functional form. Estimation and predictive accuracy is further improved with kernel PCA approximation to reduce the degrees of freedom in the first stage and with adaptive regularization in the second stage to remove non-informative regions from the final prediction model. Prediction accuracy is assessed with bias-corrected ROC curves and AUC statistics. Numerical studies suggest that the model performs well in the presence of non-informative regions and both linear and non-linear effects.

Bayesian statistics in a regulatory environment: innovative statistics at the food and drug administration [Special II]
Gregory Campbell

Center for Devices and Radiological Health
U.S. Food and Drug Administration

Abstract. Bayesian designs and analyses are part of an increasing number of premarket submissions to FDA’s Center for Devices and Radiological Health. This innovative initiative, which began in the late 1990s, takes advantage of good prior information on safety and effectiveness that is often available for studies of the same or similar recent generation medical devices. The adaptive approach can be facilitated using a Bayesian design, even for non-informative priors, and can provide insight into stopping not only for success or futility but also using predictive probabilities to curtail patient enrollment. Decision rules are based on the posterior probability calculations and are often confirmed through simulation of the operating characteristics of the designs. Accomplishments of this initiative are reviewed. A number of important lessons in the design, conduct and analysis of medical devices studies that have been learned in the course of the Bayesian initiative at CDRH are explored. Insights into success are addressed as well as challenges for the future.

Two stage adaptive lasso for [I.7.3]
Mehmet Caner

Economics
North Carolina State University

Abstract. The paper proposes adaptive lasso estimators of Zou (2006) to select instruments and then do estimation in simultaneous equations framework. This will extend the theories in econometrics in several ways. First, this will be selecting only strong instruments and eliminating the weak ones. Then we can select the model through adaptive lasso also in the second stage. Compared to statistics literature, there are extensions. First we extend the
univariate adaptive lasso to multivariate lasso. Then we also find near minimax optimal risk bounds in non Gaussian case. This is for iid data. We also establish bounds in the non iid case as well. Adaptive lasso achieves the optimal risk bound in iid case but for independent heterogeneous data (unconditional heteroskedasticity) this does not hold. The proofs for these are different from Zou (2006).

An estimating equation approach for analyzing dependent over-dispersed count data [I.6.5]

N. Rao Chaganty

Department of Mathematics and Statistics
Old Dominion University

Abstract. For dependent and over-dispersed count data, parameter estimation and inference using the method of maximum likelihood can be achieved with copula-based statistical models with negative binomial marginal distributions. For situations where the regression parameters or other univariate parameters are of main interest and not the dependence, we propose an estimating equation approach known as "weighted scores." This method is based on weighting univariate marginal score functions, where the weight matrices are estimated by initially fitting a Gaussian copula model. Asymptotic and small sample efficiency calculations show that the weighted scores method is robust and nearly as efficient as maximum likelihood for other fully specified copula models. An example is given to illustrate the use of our approach to analyze data on health care utilization based on family characteristics. This is a joint work with Dr. Aristidis K. Nikoloulopoulos and Dr. Harry Joe.

Estimating optimal dynamic treatment regimes with shared decision rules across stages: an extension of q-learning [I.7.4]

Bibhas Chakraborty

Biostatistics
Columbia University

Abstract. Dynamic treatment regimes (DTR) are individually tailored treatments. They offer a framework to operationalize the adaptive multistage decision making in clinical practice, thus providing an opportunity to improve such clinical decision making. Formally, a DTR is a set of treatment decision rules, one per stage; each decision rule takes a patient’s treatment and covariate history as input, and outputs a recommended treatment. In some studies, these decision rules are shared across stages of treatment, i.e. the decision rule at each stage is the same function of the history available at that stage. Because of this sharing phenomenon, recursive methods like Q-learning (and recursive g-estimation in Structural Nested Mean Models) that move backward through stages are not suitable; simultaneous estimation techniques are necessary. In this paper, we propose a novel simultaneous estimation procedure for the optimal DTR with shared decision rules, adjusting for non-regularity via thresholding. We compare the performance of the proposed method to other available approaches through simulation.
BARN: Bayesian additive regression network [I.4.5]
Sounak Chakraborty

Statistics
University of Missouri-Columbia

Abstract. Bayesian ensemble learning otherwise known as the Bayesian additive regression tree (Chipman, et al. 2005) is a pioneering work in terms of adopting the philosophy of the "slow learners" under a Bayesian setup. In this paper we describe Bayesian ensemble methods for classification and regression. Our model is based on neural network architecture. Using an ensemble of small neural networks we develop our Bayesian ensemble networks for regression and classification problems. The number of required networks are selected adaptively using a prior. A Bayesian back-fitting algorithm is proposed to fit the model. Success of our methods are demonstrated based on simulated and real data sets.

Large time behaviors of quantum Markov semigroups [C.7.6]
Mou-Hsiung (Harry) Chang

Mathematical Sciences
U.S. Army Research Office

Abstract. This paper investigates large time behavior of a quantum Markov semigroup of bounded linear operators acting on a von Neumann algebra defined on a complex Hilbert space. Large deviation principle as well as necessary and sufficient conditions for existence of an invariant state are established using integral representation of normal states.

The extended Fay-Herriot model: a nonparametric small area framework [I.2.4]
Ansu Chatterjee

School of Statistics
University of Minnesota

Abstract. Building on the Fay-Herriot model and Henderson’s method of obtaining best linear unbiased predictor (BLUP), we propose an extended small area model that does not require an explicit two-level parametric modeling. We discuss how to measure prediction accuracy under such a non-parametric framing of the Fay-Herriot model. We show some robustness properties of this extended Fay-Herriot model, and show how dependent small area data may be handled. Examples based on climate sciences are discussed.

Higher order properties of the bootstrap for adaptive lasso estimators [I.5.3]
Arindam Chatterjee

Stat-Math Unit
Indian Statistical Institute, New Delhi

Abstract. In this paper, we consider higher order properties of the bootstrap method for the
The search for missing heritability and some statistical problems
[Special III]
Nilanjan Chatterjee

Biostatistics Branch
Division of Cancer Epidemiology and Genetics

Abstract. Recent genome-wide association studies have led to the identification of many susceptibility loci for a variety of complex traits of quantitative and qualitative nature. Although a large fraction of heritability of many of the individual traits still remain unexplained, there are now enough discoveries to begin assessing the ‘genetic architecture’ of polygenic traits in an empirical fashion. In the first part of this talk, I will describe a set of statistical tools that can use data from existing studies to estimate the number of underlying susceptibility loci for a trait and the distributions of various related parameters, such as effect sizes and allele frequencies. These methods are applied to analyze data for a variety of traits to estimate the number of undiscovered susceptibility loci, sample size required for future studies for further discoveries, potential utility of additional discoveries for risk-prediction and to examine the relationship between effect-size and allele frequencies. In the second part of the talk, I will describe a novel approach to meta-analysis that can combine association signals from a set of heterogeneous, but possibly related, phenotypes to explore common genetic links between them. An application of the method is illustrated using data from GWAS of 30,000 cases and 30,000 controls across six different cancer sites.

An empirical likelihood based approach to incorporate sampling weights and population level information [I.3.1]
Sanjay Chaudhuri
Statistics and Applied Probability
National University of Singapore

Abstract. We consider methods to include sampling weights in an empirical likelihood based estimation procedure to augment population level information in sample-based statistical modelling. Our estimator uses conditional weights and is able to incorporate covariate information both through the weights and the usual estimating equations. We show that the estimates are strongly consistent, asymptotically unbiased and normally distributed. Moreover, they are more efficient than other methods. Our framework provides additional justification for inverse probability weighted score estimators in terms of conditional empirical likelihood. We give two applications to demographic hazard modelling by combining

Chatterjee
Chaudhuri
birth registration data with complex survey data to estimate annual birth probabilities.

A unified approach for nonparametric evaluation of agreement in method comparison studies [I.6.4]

Pankaj Choudhary

Mathematical Sciences
University of Texas at Dallas

Abstract. We present a nonparametric methodology for evaluation of agreement between multiple methods of measurement of a continuous variable. Our approach is unified in that it can deal with any scalar measure of agreement currently available in the literature, and can incorporate repeated and unreplicated measurements, and balanced as well as unbalanced designs. Our key idea is to treat an agreement measure as a functional of the joint cumulative distribution function of the measurements from multiple methods. This measure is estimated nonparametrically by plugging-in a weighted empirical counterpart of the joint distribution function. The resulting estimator is shown to be asymptotically normal under some specified mild assumptions. A closed-form expression is provided for the asymptotic standard error of the estimator. This asymptotic normality is used to derive a large-sample distribution-free methodology for simultaneously comparing the multiple measurement methods. The small-sample performance of this methodology is investigated via simulation. The asymptotic efficiency of the proposed nonparametric estimator relative to the normality-based maximum likelihood estimator is also examined. The methodology is illustrated by applying it to a blood pressure data set involving repeated measurements from three measurement methods.

Generalized fiducial inference for normal mixed linear models [Poster]

Jessi Cisewski

Statistics and Operations Research
UNC - Chapel Hill

Abstract. While linear mixed modeling methods are foundational concepts introduced in any statistical education, adequate general methods for estimation of confidence intervals involving models with more than a few variance components is lacking, especially in the unbalanced setting. Generalized fiducial inference provides a framework that accommodates this absence of methodology. Under the fabric of generalized fiducial inference along with sequential Monte Carlo methods, we present an approach for quantile estimation for both balanced and unbalanced Gaussian mixed linear models. This approach applies to situations where the data are measurements with a degree of uncertainty or where the data are intervals. We compare the proposed method to classical and Bayesian results in the literature in a simulation study of two-fold nested models and two-factor crossed designs with an interaction term. The proposed method is found to be competitive or better when evaluated on empirical coverage and average length of confidence intervals.

Medain cross validation for complex models and heavy tailed errors
Bertrand Clarke  
*Medicine*  
*U. of Miami*

Abstract. We present an alternative to the usual cross-validation (CV) by taking the median of the squared residuals rather than the average. We present extensive computation support to argue that median CV outperforms regular CV in the linear models context when the error terms have heavy tails or when the models are complex in the sense of having many terms and especially many small terms rather than a few leading terms.

**Soil mapping for very large datasets** [Special III]

Noel Cressie  
*Department of Statistics*  
*The Ohio State University*

Abstract. The ability to take many observations at precisely known spatial locations has transformed agriculture into a spatial science. An important component of precision agriculture is its intersection with pedometrics. Maps of soil properties are in great demand, but there is a point where too much data can “break” the algorithms associated with statistically optimal (kriging) maps. In this research, a geostatistical methodology is developed that relies on highly flexible non-stationary spatial covariances. After estimating the covariance function, exact kriging can be carried out for very large datasets (on the order of tens of thousands to hundreds of thousands of data). The methodology is applied to total counts obtained from gamma radiometer readings in several fields of the Nowley Farm, New South Wales. This is joint research with Emily Kang.

**Statistical inference for visual methods for exploratory data analysis** [I.3.2]

Dianne Cook  
*Statistics*  
*Iowa State University*

Abstract. Plots of data often provoke the response “is what we see really there”. Statistical significance of “discoveries” can be measured by having an impartial human viewer compare the plot of the real dataset with collections of plots of null datasets: plots take on the role of test statistics, and human cognition the role of statistical tests. This is a simple but rigorous protocol that provides inferential validity. It is modeled after the “lineup”, popular from criminal legal procedures. Another protocol modeled after the “Rorschach” inkblot test well-known from (pop-)psychology will help analysts acclimatize to random variability before being exposed to the plot of the real data.

A new R package, nullabor, allows these protocols to be administered by ourselves, the data analysts. In this talk we will discuss our experience with using these protocols as part of a data analysis, and the results of human subject experiments to quantify the power of
the lineup protocol.

Joint with Heike Hofmann, Hadley Wickham, and Andreas Buja.

**Inference on treatment effects from a randomized clinical trial in the presence of premature treatment discontinuation: the synergy trial** [Special I]

Marie Davidian

Statistics
North Carolina State University

Abstract. The Superior Yield of the New Strategy of Enoxaparin, Revascularization, and GlYcoprotein IIb/IIIa inhibitors (SYNERGY) trial was a randomized, open-label, multi-center clinical trial comparing two anticoagulant drugs (enoxaparin and unfractionated heparin, UFH) on the basis of various time-to-event endpoints. In contrast to those of other studies of these agents, the primary, intent-to-treat analysis did not find sufficient evidence of a difference, leading to speculation that premature discontinuation of the study agents by some subjects might have attenuated the treatment effect. As is the case in such trials, some subjects discontinued (stopped or switched) their assigned treatment prematurely, either because occurrence of an adverse event or other condition under which discontinuation was mandated by the protocol or due to other reasons, e.g., switching to the other treatment at his/her provider’s discretion (with more subjects switching from enoxaparin to UFH than vice versa). In this situation, interest often focuses on the difference in survival distributions had no subject discontinued his/her assigned treatment, inference on which is often attempted via standard analyses where event/censoring times for subjects discontinuing assigned treatment are artificially censored at the time of discontinuation. However, this and other common ad hoc approaches may not yield reliable information because they are not based on a formal definition of the treatment effect of interest. We use SYNERGY as a context in which to describe how such an effect may be conceptualized properly and to present a statistical framework in which it may be identified, which leads naturally to the use of inverse probability weighted methods.

This is joint work with Min Zhang (University of Michigan), Butch Tsiatis (NCSU), and Karen Pieper and Ken Mahaffey (Duke Clinical Research Institute)

**Pitman closest estimators based on convex linear combinations of two order statistics** [C.1.1]

Katherine Davies

Department of Statistics
University of Manitoba

Abstract. Comparisons of best linear unbiased estimators with some other prominent estimators have been carried out over the last fifty years since the ground breaking work of Lloyd (1952). Recently, Pitman closeness comparison of estimators based on order statistics have been carried out by Balakrishnan et al. (2009, 2011). In this talk, we discuss the Pitman closest estimation based on convex linear combination of two order statistics, which does shed
some light with regard to the estimation of population median in the case of even sample size.

Model selection by testing for the presence of small-area effects, and application to area-level data [I.2.4]

Gauri S Datta

Statistics
University of Georgia

Abstract. The models used in small-area inference often involve unobservable random effects. While this can significantly improve the adaptivity and flexibility of a model, it also increases the variability of both point and interval estimators. If we could test for the existence of the random effects, and if the test were to show that they were unlikely to be present, then we would arguably not need to incorporate them into the model, and thus could significantly improve the precision of the methodology. In this work we suggest an approach of this type. We develop simple bootstrap methods for testing for the presence of random effects, applicable well beyond the conventional context of the natural exponential family. If the null hypothesis that the effects are not present is not rejected then our general methodology immediately gives us access to estimators of unknown model parameters and estimators of small-area means. Such estimators can be substantially more effective, for example because they enjoy much faster convergence rates than their counterparts when the model includes random effects. If the null hypothesis is rejected then the next step is either to make the model more elaborate (our methodology is available quite generally) or to turn to existing random effects models.

Predicting patient survival from proteomic profile using MALDI-TOF mass spectrometry data [I.7.1]

Susmita Datta

Department of Bioinformatics and Biostatistics
University of Louisville

Abstract. Predicting survival times of patients with the proteomic profile of bodily fluids for example blood, plasma, serum etc., affected cells and tissues has been of interest in biomedical study. Here, we consider predicting survival time from the proteomic profiles of patient sera using Matrix-assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) data of non-small cell lung cancer (NSCLC) patients. Due to much larger dimension of proteomic features in a mass spectrum compared to the study sample size traditional linear regression modeling on the log-transformed failure times is not feasible. Hence, we consider latent factor and regularized/penalized methods such as least absolute shrinkage and selection operator(LASSO), partial least squares (PLS), elastic net regularization and sparse partial least square(SPLS) for fitting such model in order to predict patient survival. However, it is expected that there are differences between the performances of these models. In this work, we study the performances of all these models in terms of fit to the patient survival time and prediction of the future survival times. Right censoring in the data is handled through a residual based multiple imputation. We conclude that more complex methods
like SPLS and Elastic Net performs well provided the tuning parameters are chosen carefully.

**Estimation of small area poverty indicators using skew-normal linear mixed models** [I.2.4]

Mamadou S. Diallo

*Mathematics and Statistics*

*Carleton University*

Abstract. Eradication of extreme poverty and hunger is the first of the Millennium Development Goals established by the United Nations. Availability of reliable statistics on peoples living conditions is a basic requirement for the achievement of this goal. However, information collected from national surveys is often limited and allows reliable estimation only for large regions or large population subgroups. For areas with small (or even zero) sample sizes it is necessary to employ indirect estimation methods that can lead to reliable estimates by borrowing information across related areas through linking models. In this talk, our focus is on small area estimation of poverty measures that are complex non-linear functions of the values of a welfare variable, in particular on a class of measures used by the World Bank that covers poverty incidence, poverty gap and poverty severity as special cases. Molina and Rao (2010) use the empirical Bayes (EB) method to obtain efficient EB estimators of small area poverty measures, based on a nested error linear regression linking model relating a transformed welfare variable to auxiliary variables available for all population units. Molina and Rao studied the case of normal errors and the inferences depend on the normality assumption. In this talk we relax normality by assuming skew normal errors (either random effects or unit errors or both) which cover a variety of distributions including normal. We obtain EB inferences for poverty indicators under skew normal and present simulation results. The proposed methodology is applicable to general non-linear parameters.

Joint work with Kalyan Das, J. N. K. Rao

**Bayesian ensemble methods for survival prediction in gene expression data** [I.2.2]

Kim-Anh Do

*Department of Biostatistics*

*MD Anderson Cancer Center*

Abstract. We propose a Bayesian ensemble method for survival prediction in high-dimensional gene expression data. We specify a fully Bayesian hierarchical approach based on an ensemble sum-of-trees model and illustrate our method using three popular survival models. Our non-parametric method incorporates both additive and interaction effects between genes, which results in high predictive accuracy compared to other methods. In addition, our method provides model-free variable selection of important prognostic markers based on controlling the false discovery rates; thus providing a unified procedure to select relevant genes and predict survivor functions. We assess the performance of our method on several simulated and real microarray data sets. We show that our method selects genes potentially related to the development of the disease as well as yields predictive performance that is very competitive.

*Diallo* 20  

*Do*
to many other existing methods.

**Conditions for equivalence between Mallows distance and convergence to stable laws** [Poster]

Chang C.Y. Dorea

*Mathematics*

*Universidade de Brasilia*

Abstract. Convergence in Mallows distance is of particular interest when heavy-tailed distributions are considered. For $1 \leq \alpha < 2$, it constitutes an alternative technique to derive Central Limit type theorems for non-Gaussian $\alpha$-stable laws. In this note, we further explore the connection between Mallows distance and convergence in distribution. Conditions for their equivalence are presented.

**Identifiability of linear structural equation models** [I.6.1]

Mathias Drton

*Department of Statistics*

*University of Chicago*

Abstract. Structural equation models are multivariate statistical models that are defined by specifying noisy functional relationships among random variables. This talk treats the classical case of linear relationships and additive Gaussian noise terms. Each linear structural equation model is associated with a graph and corresponds to a polynomially parametrized set of positive definite covariance matrices. A basic problem is to determine which models are identifiable. In other words, the problem is to determine the graphs for which the polynomial parametrization map is injective, possibly only generically so. I will discuss recent results on this problem based on joint work with Jan Draisma, Rina Foygel and Seth Sullivant.

**Variable selection in semiparametric regression model for right censored data** [I.5.6]

Pang Du

*Statistics*

*Virginia Tech*

Abstract. Recent biomedical studies often measure two distinct sets of risk factors: the first is the low dimensional clinical and environmental measurements, and the second is the high dimensional gene expression measurements. For prognosis studies with right censored response variables, we propose a semiparametric regression model whose covariate effects have two parts. The first part takes a nonparametric form for the low dimensional covariates and the second part takes a parametric form for the high dimensional covariates. An effective penalized variable selection approach is developed. The selection of parametric covariate effects is through an iterated Lasso approach, for which we prove the selection consistency property. The nonparametric component is estimated using a sieve approach. An empirical model selection tool for the nonparametric component is derived based on the
Kullback-Leibler geometry. Numerical studies show that the proposed approach has satisfactory finite-sample performance. An application to a lymphoma study illustrates the method.

Factor partition models for high-dimensional mixed domain data [I.1.1]
David Dunson

Statistical Science
Duke

Abstract. There is increasing interest in many application areas in developing flexible joint models for mixed domain data. For example, data may consist of not only vectors of categorical, count and continuous variables but also document text, images, and curves of various types. Traditional methods of joint modeling based on Gaussian latent factors or random effects fall short in such settings. Alternatives have been proposed using nonparametric Bayes methods that place Dirichlet process priors on a joint random effects distribution. Such models induce dependence through defining a discrete mixture model for each type of data, with the mixture component allocation (cluster) for a subject being identical across all domains. Unfortunately, global clustering is restrictive and can lead to introducing an over-abundance of clusters. We propose a new class of factor partition models that instead allow separate but dependent clustering in each domain through generalizing a class of simplex factor models that we recently developed for efficient nonparametric modeling of high-dimensional unordered categorical data. Properties are described and we develop a highly efficient MCMC algorithm for posterior computation that scales well with increasing dimension. The methods are illustrated through applications to a variety of settings including prediction from high-dimensional objects (curves, text, images, etc) allowing higher order interactions.

A few seedlings of research [I.2.1]
Rick Durrett

Math
Duke U

Abstract. We will describe recent research done during the ongoing SAMSI program concerning the processes taking place on random networks.

On the determination of optimal manpower reserves at three nodes in series [C.1.1]
R. Elangovan

Statistics
Annamalai University

Abstract. The welfare and prosperity of any country depends up on the natural resources available in that country. But the proper utilization of the natural resources depends up on the availability of manpower of the appropriate type. In modern days due to the advance-
ment of civilization, progress of science and technology the need for the manpower of skilled personnel, has become imperative. But the training of the individuals according to the needs of skilled manpower of various natures has ended in the complex policies of manpower management. There are many industries and organizations where the skilled personnel are to be recruited and they must be given prior training before employment. In the area of Software and Information Technology the training and induction of the right type of personnel is a pressing problem. In this paper, we consider the optimal solution for the manpower to be kept as reserve inventory at two different nodes in series. This is based on the inventory model for finding the optimal size of reserve inventory at two successive stations in series. The models developed in this paper also provides a tool for assessing the manpower profile and predicting future manpower development on an industry wide basis.

A ROAD to classification in high dimensional space [I.5.6]
Yang Feng

Statistics
Columbia University

Abstract. For high-dimensional classification, it is well known that naively performing the Fisher discriminant rule leads to poor results due to diverging spectra and noise accumulation. Therefore, researchers proposed independence rules to circumvent the diverse spectra, and sparse independence rules to mitigate the issue of noise accumulation. However, in biological applications, there are often a group of correlated genes responsible for clinical outcomes, and the use of the covariance information can significantly reduce misclassification rates. The extent of such error rate reductions is unveiled by comparing the misclassification rates of the Fisher discriminant rule and the independence rule. To materialize the gain based on finite samples, a Regularized Optimal Affine Discriminant (ROAD) is proposed based on a covariance penalty. ROAD selects an increasing number of features as the penalization relaxes. Further benefits can be achieved when a screening method is employed to narrow the feature pool before hitting the ROAD. An efficient Constrained Coordinate Descent algorithm (CCD) is also developed to solve the associated optimization problems. Sampling properties of oracle type are established. Simulation studies and real data analysis support our theoretical results and demonstrate the advantages of the new classification procedure under a variety of correlation structures. A delicate result on continuous piecewise linear solution path for the ROAD optimization problem at the population level justifies the linear interpolation of the CCD algorithm.

Some recent results on high-dimensional approximations with computable error bounds [I.5.4]
Yasunori Fujikoshi

Mathematics, Faculty of Science
Hiroshima University

Abstract. In the recent study of multivariate statistical analysis there are much interests in approximations under high-dimensional framework when dimension p and sample size n go
to infinity together with the ratio p/n converging a limit. These approximations are revising to make the classical multivariate methods useful in wide range of relations between p and n. Furthermore, such approximations are useful in extending multivariate statistical theory in high-dimensional situations when p is comparable with n or even exceeds it.

In this paper we first review some results on high-dimensional approximations of statistics including MANOVA tests, MANOVA roots, canonical correlations, tests for covariance matrices, statistics with a class of moments, and linear discriminant functions with ridge-type estimator for the covariance matrix. Next we discuss with their errors. Most results supply the so-called order estimates only. However, such estimates do not give information on actual errors for given values n, p and other parameters. Ideally, we wish to have computable error bounds, in addition to order estimates. We show that such computable error bounds are given for some high-dimensional approximations of tests with a class of moments, which include LR tests for mean vectors and covariance matrices, and linear discriminant functions.

**Statistical physics, interpolation method and scaling limits in sparse random graphs** [I.1.2]
David Gamarnik

*Sloan School of Management and Operations Research Center
MIT*

Abstract. Recently a powerful approach was developed in the physics of spin glasses based on the interpolation idea. Among other things, the interpolation method was used to prove the existence of the so-called free energy thermodynamic limits for several spin glass models, including Sherrington-Kirkpatrick, Viana-Bray and random K-SAT models.

We propose a simple combinatorial approach to the interpolation method and demonstrate it’s applicability in context of combinatorial optimization problems on sparse Erdos-Renyi and random regular graphs. Specifically we establish the existence of scaling limits for maximum independent sets, MAX-CUT, coloring and K-SAT problems. For these models, we show that the optimal values appropriately normalized, converge to a limit with high probability (w.h.p.), as the size of the underlying graph diverges to infinity. In the context of independent set model this resolves an open problem posed by Aldous in 2000. Our simpler combinatorial approach allows us to work with the zero temperature case (optimization) directly, which simplifies the analysis substantially. Additionally, using our approach, we establish the large deviations principle for the satisfiability property for constraint satisfaction problems such as coloring, K-SAT. The talk will be completely self-contained. No background on random graph theory/statistical physics is necessary.

Joint work with Mohsen Bayati (Stanford) and Prasad Tetali (Georgia Tech).

**Optimal buffer size and dynamic rate control for a queueing network with reneging in heavy traffic** [I.4.1]
Arka Ghosh

*statistics
Iowa State University*
Abstract. We address a control problem associated with a single server Markovian queue with customer abandonment with an infinite horizon discounted cost under heavy traffic. The controls are: buffer size for the queue and a state-dependent service rate. By solving the limiting diffusion control problem explicitly, we propose a suitable control for queueing network and prove its asymptotic optimality. This optimal solution also identifies the parameter regimes where the infinite buffer size is optimal.

**Genomic applications of multivariate multiple testing procedures**
[I.7.1]
Debashis Ghosh

*Statistics*
*Penn State*

Abstract. There has been a lot of recent statistical and bioinformatic literature dealing with the situation where one is faced testing multiple hypotheses from high-dimensional “-omics” data. In this talk, we describe genomic problems in which for each feature, there is a vector of test statistics or equivalently p-values. The goal then centers on identification of “interesting” features that investigators wish to followup on further. Such problems arise naturally when considering genomic meta-analysis problems, outlier profile analysis using diverse data types and determining periodic genes. Consideration of such settings requires extension of standard multiple testing procedures, such as the Benjamini-Hochberg procedure, to deal with “multivariate p-values.” We describe generalizations of the B-H procedure based on geometric ideas as well as concepts from spacings. We describe some new methods for this problem and assess their finite-sample properties using simulated and real data.

**Bayesian and non-bayesian high dimensional model selection: a quick survey of basic issues, some theory and computation**
[I.2.5]
Jayanta K Ghosh

*Department of Statistics*
*Purdue University*

Abstract. To make the talk accessible, I will begin with AIC and BIC, then go on to survey some recent problems in high dimensional model selection, illustrating with a toy example on SNP’s and Lasso, an old example due to M. Stone with some new results, stability of computation of Bayes Factors, etc. The talk is partly a survey and partly based on new work in three recent papers, two with Ritabrata Dutta, one with him and Gosia Bogdan.

**Capturing species at a certain abundance level in a multi-species assemblage with unknown heterogeneity**
[I.5.1]
Kaushik Ghosh

*Mathematical Sciences*
*University of Nevada, Las Vegas*

Abstract. In a sample of mRNA species counts, sequences without duplicates or with small
numbers of copies are likely to carry information related to mutations or diseases and can be of great interest. However, in some situations, sequence abundance is unknown and sequencing the whole sample to find the rare sequences is not practically possible. In addition, the existence of rare sequences in the whole sample rules out the implementation of multinomial distribution, a computationally more friendly distribution function, to model the mRNA species counts. To collect mRNA sequences of interest, or more generally, species of interest, we propose a two-phase Bayesian sampling method with these concerns addressed. The first phase of the design is used to infer sequence (species) abundance through a cluster analysis applied to a pilot data set. The clustering method is built upon a multivariate hypergeometric model with a Dirichlet process prior for species relative frequencies. The second phase, through Monte Carlo simulations, infers the sample size to collect a certain number of species of particular interest. Due to the dependence between variables and the limitation of Dirichlet process, efficient posterior computing methods are proposed. The developed approach is demonstrated and evaluated via simulations. An mRNA segment data set is used to illustrate and motivate the proposed sampling method.

**Benchmarked small area estimators** [Special I]

Malay Ghosh

*Statistics  
University of Florida*

Abstract. The paper considers benchmarking issues in the context of small area estimation. We find optimal estimators within the class of benchmarked linear estimators under either external or internal benchmark constraints. This extends existing results for both external and internal benchmarking and also provides some links between the two. In addition, necessary and sufficient conditions for self-benchmarking are found for an augmented model. Most results of this paper are found using ideas of orthogonal projection.

**A strong law for the rate of growth of long latency periods in cloud computing service** [I.1.4]

Souvik Ghosh

*Department of Statistics  
Columbia University*

Abstract. Cloud-computing shares a common pool of resources across customers at a scale that is orders of magnitude larger than traditional multi-user systems. Constituent physical compute servers are allocated multiple ”virtual machines” (VM) to serve simultaneously. Each VM user should ideally be unaffected by others’ demand. Naturally, this environment produces new challenges for the service providers in meeting customer expectations while extracting an efficient utilization from server resources. We study a new cloud service metric that measures prolonged latency or delay suffered by customers. We model the workload process of a cloud server and analyze the process as the customer population grows. The capacity required to ensure that average workload does not exceed a threshold over long segments is characterized. This can be used by cloud operators to provide service guaran-
tees on avoiding long durations of latency. As part of the analysis, we provide a uniform large-deviation principle for collections of random variables that is of independent interest.

**A Bayesian semiparametric model for the analysis of sequence counts data.** [I.4.2]

Michele Guindani

*Biostatistics*

*UT MD Anderson Cancer Center*

Abstract. Many data resulting by the use of modern sequencing technologies can be summarized by the frequencies of observed features. As examples, we can cite digital gene expression data in next generation sequencing studies, or T cell receptors sequence data in Immunological studies. Those sequences are often characterized by overdispersion, since many low frequency measurements are observed together with highly abundant items. Hence, estimating the distribution of the frequency counts and their relative abundance is of particular importance to understand the underlying biological phenomena.

We propose a Bayes nonparametric model to address the problem of estimating the true abundances in a recorded library of SAGE (Serial Analysis of Gene expression) data and in a T-cell repertoire. Then, we extend the model to take into account the more general comparison of recorded frequencies across several libraries and conditions and exemplify our methodology on SAGE data from a known set of libraries sampled from gastrointestinal healthy and tumor tissues.

**Adaptive FWER and FDR control under block dependence** [I.4.3]

Wenge Guo

*Mathematical Sciences*

*New Jersey Institute of Technology*

Abstract. In the context of multiple hypotheses testing, the proportion of true null hypotheses among all nulls often plays an important role, although it is generally unknown a priori. In adaptive procedures this proportion is estimated and then incorporated into conventional procedures to derive more powerful methods to control the familywise error rate (FWER) or the false discovery rate (FDR). Recently, a number of adaptive Bonferroni methods controlling the FWER and adaptive BH methods controlling the FDR have been introduced in the literature. However, until now, in a non-asymptotic setting (with the number of hypotheses not being infinitely large), the FWER or FDR control of all practical adaptive procedures is proved only under independence, although extensive simulation studies have suggested that such adaptive procedures perform well under certain type of dependence structure. In this talk, several variants of the conventional adaptive Bonferroni and BH methods will be presented, along with proofs that these newly suggested procedures provide ultimate control of the FWER or FDR under block dependence among the underlying p-values. Results of simulation studies comparing the performances of these adaptive procedures with the conventional FWER and FDR controlling procedures under dependence will also be presented.
Exploring new models for population prediction in detecting demographic phase change for sparse census data [C.7.5]

Arindam Gupta

Statistics
Burdwan University, India

Abstract. Logistic model has some limitations when applied for developing countries. In such situation the relative growth rates (RGR) exhibit some unusual trends (increasing, primary increasing and then decreasing). To tackle those situations we extend the logistic law by incorporating nonlinear positive and negative feed back terms. Here we have assumed that RGR is a function of size and time separately. The time covariate model has some key advantages than the size covariate model. It can detect the demographic phase change point at which a developing country switches over towards a developed one.

A study of the asymptotic properties of the lasso estimator for correlated data [I.2.5]

Shuva Gupta

statistics division, mathematics dept
Northern Illinois University

Abstract. In this talk we investigate asymptotic properties of penalized weighted least squares estimators in regression models with a large number of variables and correlated errors with weak dependency structure (Doukhan 1996). We show that the estimate can be used for consistent selection of the true regression covariates, even if the number of variables exceeds the sample size. We establish the asymptotic distribution of the estimators, and show that under appropriate choices of the tuning parameters the limiting distribution is multivariate normal. This generalizes the result of Knight and Fu (2000), obtained for regression models with independent errors, to the case of weakly correlated errors.

Optimal progressive type-II censoring schemes for nonparametric confidence intervals of quantiles [I.3.3]

David Han

Department of Management Science & Statistics
University of Texas at San Antonio

Abstract. In this talk, optimal progressive censoring schemes are examined for the nonparametric confidence intervals of population quantiles. The results obtained can be universally applied to any continuous probability distribution. By using the interval mass as an optimality criterion, the optimization process is free of the actual observed values from the sample and needs only the initial sample size n and the number of complete failures m. Using several sample sizes combined with various degrees of censoring, the results of the optimization are presented for the population median at selected levels of confidence (99, 95, and 90%). With the optimality criterion under consideration, the efficiencies of the worst progressive Type-II censoring scheme and ordinary Type-II censoring scheme are also examined in comparison.
to the best censoring scheme obtained for fixed $n$ and $m$.

**Kernel averaged predictors for spatio-temporal processes** [I.1.4]

Matthew Heaton

*Statistical Science*
*Duke University*

Abstract. For spatio-temporal processes, predictors from multiple locations affect the response at a separate location. For example, predictors such as precipitation, temperature, pollution emissions, etc. are often used to explain ground-level ozone production. Due to weather and other factors, however, the relationship between these predictors and ozone is not confined to a single spatial location or time period as is often assumed. Again, the effect of pollution on mortality is spatially and temporally lagged because mortality does not, typically, occur at time of exposure. Here, kernels are proposed as a tool to properly weight predictor surfaces in spatio-temporal regression models. The kernels are assumed to be parametric with parameters that are estimable from the data. Distributional results are provided for the case of a univariate predictor and response in the Gaussian process setting. Additionally, relations to previously proposed models and computational details are discussed. The methodology is demonstrated on simulated data as well as in an application to ozone levels as explained by temperature.

**The lomax distribution as a failure model under various loss** [C.7.6]

Amal Helu

*Mathematics*
*University of Jordan*

Abstract. Based on progressively Type-II censored samples, the maximum likelihood, the uniformly minimum variance unbiased (UMVU), and the Bayes estimators for the unknown parameter and the failure function of the Lomax model are derived. The Bayes estimators are obtained based on both the symmetric (Squared error and Absolute difference) loss functions, and asymmetric (LINEX, General entropy, and Logarithmic) loss functions, this was done with respect to the conjugate prior. A comparison between these estimators is provided. A real data example consists of data from Iowa 65+ Rual Health Study (RHS) are used for illustrating the proposed methods.

**Interactive statistical graphics for data exploration** [I.3.2]

Heike Hofmann

*Statistics*
*Iowa State University*

Abstract. Interactive graphics for looking at data has been promoted by J.W. Tukey since the 1970s. The difference between then and now is that we now have the computational power and the tools to make interactive graphics part of mainstream statistical practice. Based on R, the statistical lingua franca, we present packages that allow the regular user on a standard
machine to make use of interactivity in their every day data analysis. Instantaneous graphical feedback allows selection and linked brushing between plots, interactive querying and other plot-specific interactions. The framework for this interactive graphics system is open and allows users to adjust and extend existing tools.

We will introduce the cranvas package and show case studies.

**A Bayesian approach to seasonal adjustment of long memory time series** [I.6.2]

Scott Holan

*Statistics*

*University of Missouri*

Abstract. Research into long memory processes has recently spread to the modeling of seasonality through the use of generalized exponential (GEXP) time series models. This talk describes the GEXP model and introduces the new Seasonal Fractional Exponential (SFEXP) model. We explore the fit of these models to economic time series data and present an application of seasonal long memory modeling to the problem of seasonal adjustment. In particular, we discuss a structural approach to obtaining component models for seasonal and trend in the context of long memory, and use these models to obtain minimum mean square error signal extraction estimates. The approach we propose is fully Bayesian and thus naturally quantifies the uncertainty in the signal extraction estimates. Finally, this technique is illustrated on several economic time series.

This is joint work with Tucker S. McElroy (United States Census Bureau)

**An algorithm for multivariate isotonic regression** [I.1.3]

Xiaomi Hu

*Mathematics and Statistics*

*Wichita State University*

Abstract. The mean vectors of q multivariate normal populations form a p by q parameter matrix. Often the prior knowledge imposes separate restrictions on the elements in each rows. The MLE of the parameter matrix under such restrictions is a matrix projection with respect to a norm induced from an inner product that is the sum of the inner products of columns. This projection is called a multivariate isotonic regression when the restrictions are order restrictions.

A row-modification algorithm for the computation of the projection is proposed. This algorithm converts the problem of a matrix projection to that of vector projections with simplified norm systems. The procedure produces a sequence of matrices. It is shown that the sequence is convergent, and it converges to the desired projection. With a computer program that implements the algorithm several numeric examples will be illustrated to show some of the futures of the algorithm.

**Statistical issues in defense and national security** [I.1.5]

Aparna V. Huzurbazar
Statistical Sciences Group
Los Alamos National Laboratory

Abstract. The Statistical Sciences Group at Los Alamos National Laboratory is involved in a broad range of methodological and applied work in a wide range of areas. These include complex computer model evaluation, visualization and analysis of massive streaming data, reliability for engineering systems, uncertainty quantification, classification and anomaly detection, and machine learning, to name a few. I will give an overview of the group’s work and some of the specific issues that arise in the national security setting.

Multivariate association and dimension reduction: with an environmental application [I.2.5]
Ross Iaci
Mathematics
College of William and Mary

Abstract. In this talk, a new nonparametric method to recover relationships between two sets of multivariate random vectors by finding the vector projections that minimize an L2 distance between each projected vector and an unknown function of the other is discussed. The unknown functions are estimated using the Nadaraya-Watson smoother; thereby enabling the procedure to detect both linear and nonlinear relationships. The method is equivalent to Canonical Correlation Analysis (CCA) when only linear relationships exist and hence, term the method Generalized Canonical Analysis (GCA). Extensions to multiple sets and groups of multiple sets is also discussed and a bootstrap procedure, which is not specific to GCA, is developed to detect the number of significant relationships. For an environmental dataset from Los Angeles County, we apply our GCA multiple set methodology to study the nature of the relationships between mortality, weather and pollutants. In the analysis we detect the existence of both linear and nonlinear relationships between the vectors, and use these recovered associations to build a dimension reduced time series regression model for mortality. This illustrates the usefulness of our method, not only in the analysis of this particular dataset, but possibly in many other applications.

Analysis of mortality rates of Virginia: a Bayesian approach [Poster]
Rasika Jayatillake
Mathematics and Statistics
Old Dominion University

Abstract. Over the decades mortality rate has proved to be an important statistic that is used by professionals in many areas such as policy studies, actuaries, and epidemiology. As a result of many studies it is a known that the death rates are affected by the demographic of the subjects. Age and gender are two such well known factors that need to be considered when interpreting or calculating mortality. In this study death rates are modeled considering age, gender, and region as covariates. The region/spatial effect can be considered as a factor that represents the general socio-economic and environment conditions. The usual parametric modeling of mortality can prove to be complicated especially when one needs to
consider correlation between spatial regions. The advantage of Bayesian methods is that this model can be viewed as a hierarchical model and with appropriate prior models and algorithms such as Markov chain Monte Carlo (MCMC) methods these correlations can be easily incorporated. This study investigates the application of hierarchical Bayesian modeling on mortality data of Virginia obtained from CDC WONDER online database for the period of 2004-2006. The death rates are considered as independent variables with binomial distribution and their logits are modeled with regional and age-sex effects as covariates. Counties and independent cities of Virginia are considered as spatial regions. The correlations that are expected to exist between regions and consecutive age groups are modeled using spatial Gaussian Markov random field (GMRF) priors and intrinsic Gaussian priors respectively. The posterior distribution samples are generated using MCMC algorithms which use a mixture of Metropolis Hasting and Gibbs sampling techniques. The regional effects are further investigated using choropleth maps. Results of the model indicate that there exist certain clusters of regions in terms of estimated mortality rates.

**Bayesian hierarchical models for RNA-Seq data** [I.2.2]

Yuan Ji

*Department of Biostatistics*

*UT MD Anderson Cancer Center*

Abstract. I will present two projects related to the analysis of RNA-Seq data. In the first project, we tackle the problem of read mapping of multireads that can be potentially mapped to more than one genomic locations. In the second project, we discuss the classical problem of finding genes with differential expression. Both projects are based on RNA-Seq technology, a recently developed and fast growing high-throughput sequencing platform. Bayesian hierarchical models are developed and posterior inferences are performed to map the multireads for project one and to call differential expression for project two.

**Statistical inference of protein structure and function** [Plenary II]

Michael Jordan

*Statistics*

*University of California*

Abstract. The study of the structure and function of proteins raises many problems that offer challenges and opportunities for statistical research. I will overview my experiences in several such problem domains, ranging from domains where off-the-shelf ideas can be fruitfully applied to domains that require new thinking. These are: (1) the identification of active sites in enzymes; (2) the modeling of protein backbone configurations; (3) the prediction of molecular function based on phylogeny; (4) joint inference of alignment and phylogeny.

**Estimation of square of location parameter of an exponential distribution with known coefficient of variation** [C.7.5]

Dr. Harshada Joshi

*Dept of Mangement (JSPM)*
Pune University, India

Abstract. The crux of this paper is to discuss the problem of estimating the square of location parameter \( \Theta \) (i.e. \( \Theta^2 \)) of an exponential distribution when the coefficient of variation is known without error. Several estimators have been proposed with their properties. Numerical illustrations are given in the support of the present study.

**A control chart for multivariate autoregressive processes** [C.1.1]
A. A. Kalgonda

*Department of Statistics*
*The New College, Kolhapur (Affiliated to Shivaji University, Kolhapur)*

Abstract. In recent years, due to automation of measurement and data collection system, a process can be sampled at higher rates, which ultimately leads to autocorrelation. Consequently, when autocorrelation is present in the data it can have serious impact on the performance of classical control charts. Control procedures using Shewhart, MCUSUM and MEWMA charts for first order autoregressive (VAR(1)) processes are available in the literature. But time series modeling of the process data is not always straightforward and timely revision of order of process is required. Naturally, if the order of the process becomes inadequate to describe the process, the above mentioned control chart procedures for VAR(1) process may fail in its performance. Therefore in this article I propose a generalized approach of multivariate control chart for k-th order vector autoregressive process. The proposed chart not only detects the status of the process but it also helps in identifying the variable(s) responsible for the out of control state. An example illustrating these facts is given. Key words and Phrases: Statistical process control, Multivariate process monitoring, Autoregressive process.

**Non-explosiveness and immortal particle for a catalytic branching process** [I.2.1]
Min Kang

*Mathematics*
*NC State University*

Abstract. We study the existence and asymptotic properties of a conservative branching particle system driven by a diffusion with smooth coefficients for which birth and death are triggered by contact with a set. Sufficient conditions for the process to be non-explosive are given in relation to the eikonal equation. In the Brownian motions case the domain of evolution can be non-smooth, including Lipschitz, with integrable Martin kernel. The results are valid for an arbitrary number of particles and non-uniform redistribution after jump. Additionally, with probability one, it is shown that only one ancestry line survives. In special cases, the evolution of the surviving particle is studied and we will discuss explicit transition function for two particle system on a half line if time permits.

**Exact test of Hardy-Weinberg equilibrium in case-parent triad designs a la fisher via algebraic statistics** [I.7.2]

Kalonda 33

Kasala
Subramanyam Kasala
Mathematics and Statistics
University of North Carolina, Wilmington

Abstract. Case-parent triad designs are useful in genetic studies, especially when one is studying childhood diseases with complex phenotypes. The parents of the child act as controls for an assessment of risk. The design involves a specific bi-allelic marker with alleles A and a, and an affected child is genotyped at this marker. The genotypes of the child's parents are also determined at this marker. Thus a triad consists of a child and child's parents and their genotypes. Models have been developed to assess the risk of the disease when a child carries two copies of A and also when child carries one copy of A. One of the critical questions in triad design is to test the hypothesis of Hardy-Weinberg Equilibrium at the marker of interest. In the literature, such tests are carried out obliquely. In this presentation, we will develop an exact test of Hardy-Weinberg Equilibrium using algebraic statistics methodology. A key idea is to look at the fiber of all possible triad data sets with the same allele frequencies as those of the given triad data set. We will work out a Markov base for the fiber and then develop an MCMC algorithm to estimate the p value under the hypothesis of Hardy-Weinberg Equilibrium.

Submodel selection in linear regression [C.1.1]
D.N Kashid
Statistics
Shivaji University, Kolhapur, India

Abstract. Submodel selection problem is one of the important problems in regression analysis. Various submodel selection methods are available in the literature like Mallows Cp-statistic, Stepwise selection method etc. All these methods assume that the error distribution is normal and regression parameters are estimated by least squares method. In case, if error term follows non-normal distribution, outlier observation or multicollinearity is present in the data then the performance of the least squares estimator is poor. Consequently, submodel selection method based on least squares estimator selects wrong subset and this submodel gives the misleading inference. In the last decade, few methods are available in the literature for dealing the above situation. Ronchetti and Staudte (1994) have proposed robust version of Mallow’s Cp called RCp, Kashid and Kulkarni (2002) suggested more general criterion called Sp-statistic for outlier data. In this article, different estimation methods are used for regression parameter estimation. Based on these estimators, we propose the submodel selection method. The performance of the proposed method is evaluated through the simulation study. Also real data set is used for illustrating the procedure.

Analytical methods for non-proportional hazards data [I.2.3]
Amarjot Kaur
Clinical Biostatistics
Merck Research Labs

Abstract. Cox proportional hazards (PH) and Log-rank test models are the commonly used
methods in the analysis of time-to-event data. When the PH assumption does not hold, the score test based on Cox PH model is no longer optimal. The loss of efficiency can be marginal to substantial depending upon the magnitude of the violation of PH assumption. The purpose of this investigation is to examine the performance of the commonly available methods to analyze data with non-proportional hazards, such as weighted Kaplan-Meier test, piecewise exponential model, gamma frailty model, and weighted log-rank test. Performances of these methods would be compared with Cox PH model and log-rank test in terms of asymptotic relative efficiency along with empirical power using a simulation study for different scenarios of non-proportionalities.

Likelihood ratio tests for continuous monotone hazards with unknown change point [Poster]
Dong-Yun Kim

Statistics
Virginia Tech

Abstract. We consider the likelihood ratio tests (LRT) for two continuous monotone hazards with unknown change point. We establish the convergence in distribution and weak convergence of LRT. Simulation study shows that the proposed tests compare favorably to other existing tests.

Adaptive fractional polynomial modeling [C.1.1]
George J. Knafl

School of Nursing
University of North Carolina at Chapel Hill

Abstract. Fractional polynomials have been used in regression models to account for non-linear dependence of outcome variables on primary predictors. Usually, the set of power transforms is limited to a small set, but the powers can be determined adaptively from among an unlimited set of powers. An approach is formulated for adaptively estimating regression models based on fractional polynomials in available predictors and their interactions. Models are evaluated using k-fold likelihood cross-validation (LCV) scores and compared using LCV ratio tests analogous to likelihood ratio tests. Fractional polynomial models are selected using heuristic search techniques guided by tolerance parameters controlling how much of a penalty (or reduction) in LCV scores can be tolerated at individual stages of the search process. Adaptive models can be generated for both means and variances/dispersions. LCV scores can be based on true likelihoods or extended likelihoods such as extended quasi-likelihoods for modeling dispersions, pseudolikelihoods for modeling dependence in conditional models, and their combination for modeling dependence and dispersion in conditional models. Univariate outcomes can be analyzed using adaptive linear, logistic, and Poisson regression. Multivariate continuous outcomes can be modeled using adaptive marginal models assuming multivariate normality under alternate correlation structures like exchangeable and autoregressive. These marginal models induce associated conditional models which can be compared to other conditional models for the data. Multivariate continuous
and categorical outcomes can be modeled using adaptive conditional models. Adaptively selected predictors for conditional models include averages of subsets of other outcome values for subjects. When these subsets include only previous outcomes, the models are adaptive transition models. Adaptively selected predictors for conditional models also include averages of other time-varying predictor values for subjects. These adaptive regression methods have been implemented in a SAS macro. Analyses of selected data sets are also presented demonstrating these methods.

A general class of nonparametric subset selection procedures [C.7.6]
Narinder Kumar

Abstract. In this paper, a general 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 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. Similarly in animal husbandry more consistent breeding methods are preferred among the breeding methods with the same average. A business concern using different advertising methods to enhance the sales prefers the advertising methods with more consistent sales among the methods having the same average sales. The proposed class of procedures is shown to satisfy the P*-condition and have strong monotonicity, monotonicity, and unbiasedness properties. This class also includes some of the existing procedures as its members which do exist in literature under similar settings.

Using multivariate techniques, the proposed class of procedures is compared asymptotically with existing procedures and the results are found to be interesting in terms of asymptotic relative efficiency. Implementation of the proposed class of procedures is illustrated with the help of existing tables and the sample size sufficient for their implementation is determined using extensive simulation study.

A nonparametric method to estimate the number of component processes [I.3.5]
Subrata Kundu

Abstract. In recent past, estimating the number of component processes in a system of superimposed IID renewal processes has gained some interest with its wide ranging applications in software reliability and other areas. There are an unknown number ν of IID renewal processes each having a renewal distribution given by, say F. We observe only the superimposition of these ν component processes up to a time τ. If an event from a particular process occurs before time τ, then this process is detected and further events from this process by time
\( \tau \) can be identified with the process. In software reliability application, this identification is done by a mechanism known as recapture debugging (Nayak, 1988). The objective is to estimate \( \nu \) based on the above observation. In this work, we suggest a simple nonparametric method for the estimation of \( \nu \), which seems to perform well in terms of bias and asymptotic normality. We also investigate, via simulation, the effect on the estimate of \( \nu \) due to departure of the renewal distribution from an assumed model. Our simulation results demonstrate that violation from the exponential distribution can result in serious bias in the estimate of \( \nu \).

**Single factor transformation priors for density regression** [C.7.5]

Suprateek Kundu

*Biostatistics*  
*UNC Chapel Hill*

Abstract. Although discrete mixture modeling has formed the backbone of the literature on Bayesian density estimation incorporating covariates, the use of discrete mixtures leads to some well known disadvantages. Avoiding discrete mixtures, we propose a flexible class of priors based on random nonlinear functions of a uniform latent variable with an additive residual. These priors are related to Gaussian process latent variable models proposed in the machine learning literature. For density regression, we model the response and predictor means as distinct nonlinear functions of the same latent variable, thus inducing dependence through a single factor. The induced prior is shown to have desirable properties including large support and posterior consistency. We demonstrate advantages over Dirichlet process mixture models in a variety of simulations, and apply the approach to an epidemiology application.

**Bounds on correlations for multinomial random vectors with specified marginals** [Poster]

Raghavendra Rao Kurada

*Mathematics and Statistics*  
*Old Dominion University*

Abstract. Abstract: It is well known that for non-Gaussian vectors, Pearson correlations have bounds that depend on the univariate marginal distributions. The upper and lower bounds, known as Frchet bounds, correspond to the attainable extremal positive and negative dependence, and are usually more restrictive than the positive definite range. In this talk we will systematically study the ranges of the correlations between a pair of multinomial random vectors both for structured and unstructured correlation matrices. Some extensions to the trivariate case and construction of joint distributions for dependent multinomial random vectors with specified means and correlations will also be discussed. Our results are useful in simulations and for estimating dependence in categorical data.

**Data management, design and analysis challenges in the prevention of perinatal sepsis (PoPs) randomized control trial, Soweto, South Africa** [C.7.6]

*Kundu*  37  *Kuwanda*
Locadiah Kuwanda

Department of Science Technology/National Research Foundation: Vaccine Preventable Diseases
University of Witswatersrand

Abstract. The World Health Organization highlights neonatal sepsis as the fourth leading cause of death among children <5 years, and the second leading cause of neonatal death. An inexpensive and feasible intervention, to cleanse the birth canal during labor and to wash the newborn at birth with the wide-spectrum microbicide chlorhexidine, holds great promise as a neonatal sepsis prevention tool. Randomized trials done in Malawi and Egypt had design flaws that led to limited uptake, and the efficacy of the intervention remains under debate. Our study (PoPS) was a single center, prospective, randomized controlled trial evaluating the efficacy of 0.5% chlorhexidine wipes of the birth canal during labour and neonatal skin after delivery at preventing serious neonatal peripartum infections. The purpose of this paper is to describe the design characteristics and challenges of the trial and of our method of dealing with the lack of double-blind outcome assessment. To deal with the complexity of having 13 data collection forms and about 30 data tables we developed a relational database and used a double data entry system for cleaning and editing the database. Because the trial was not double-blinded, we elected to monitor some baseline characteristics by treatment group to detect any potential problems with implementation of the randomization procedures. Propensity scores analysis was planned to address imbalances. Serious Adverse Events (SAEs) were monitored and analyzed separately for mothers and infants. The primary analysis included all births and twin births were assumed to be independent events. An interim analysis was undertaken to evaluate the balance between the treatment arms and to decide on whether to stop the trial for compelling efficacy. At the conclusion of the trial, balance was found acceptable. The rate of neonatal sepsis within three days of birth did not differ between the chlorhexidine (3.5%; 141 of 4,058) and placebo arms (3.7%; 148 of 4,036; p=0.65).

Understanding clinical trial data through graphical exploration [I.3.2]
Mani Lakshminarayanan

Biostatistics and Research Decision Sciences
Merck Research Laboratories

Abstract. Research and development in pursuit of safe and efficacious pharmaceutical products involves generation of huge amount of animal and human data. In particular, the amount of data generated from clinical trials involving volunteers and patients provide wealth of information on clinical endpoints defined to investigate the efficacy and safety of investigational drugs and their comparators. Collected data from these trials that include efficacy endpoints, dosing, adverse events, lab data, medical history and others are then presented as summary tables and figures in clinical study reports and integrated summaries. Though most of the data are reported as standardized outputs for ease of reproducible presentations and to meet regulatory guidance, graphics can play a crucial role in exploratory analysis in identifying emerging issues prior to creating new hypothesis. Graphical presentations are also effective in elucidating known issues which might enable rapid and accurate interpretation of the
complex clinical data. In this talk, we will explore the use of graphics, both simple and complex multi-panel displays, in reporting clinical trial data. We will also make reference to standard graphical output that are created in support of clinical study reports.

Title... [Poster]
Seonjoo Lee

Abstract. ...

On hyperplane alignment for linear and nonlinear sufficient dimension reduction [I.2.5]
Bing Li

Statistics
Pennsylvania State University

Abstract. We introduce a Hyperplane Alignment (HA) approach that can be used for both linear and nonlinear sufficient dimension reduction. The basic idea is to divide the response variables into slices and use a modified form of support vector machine to find the optimal hyperplanes that separate them. These optimal hyperplanes are then aligned by the principal components of their normal vectors. It is proved that the aligned normal vectors provide an unbiased, square-root n consistent, and asymptotically normal estimator of the sufficient dimension reduction space. The method is then generalized to nonlinear sufficient dimension reduction using the reproducing kernel Hilbert space. In that context, the aligned normal vectors become functions and it is proved that they are unbiased in the sense that they are functions of the true nonlinear sufficient predictors. We compare HA with other sufficient dimension reduction methods by simulation and in real data analysis, and through both comparisons firmly establish its practical advantages.

Bayesian regularization via the graph Laplacian prior [I.7.3]
Fan Li

Department of Statistical Science
Duke University

Abstract. Regularization is an important approach to prevent over-fitting in regression. Under the Bayesian paradigm, many regularization techniques correspond to imposing certain shrinkage prior distributions on the regression coefficients. Existing Bayesian methods usually assume independence between explanatory variables a priori. In this article, we propose a novel Bayesian approach, which explicitly models the dependence structure between variables through a graph Laplacian matrix. We generalize the graph Laplacian to allow both positive and negative correlations. A prior distribution for the graph Laplacian is then proposed, which allows conjugacy and thereby greatly simplifies the computation. We show that the proposed Bayesian model leads to proper posterior distribution. Connection is made be-
between the proposed method and some existing regularization approaches, such as the Lasso, the Elastic Net, the OSCAR, and ridge regression. An efficient MCMC method based on parameter augmentation is developed for posterior computation. Finally, we demonstrate the method through simulation studies and a real data analysis.

**Variational Bayesian canonical correlation analysis for the integrated analysis of genetical systems biology data** [Poster]

Yan Ling

*Virginia Bioinformatics Institute*

*Virginia Tech*

Abstract. Genetical Systems Biology (GSB) combines the study of genetic variation (genome-wide association or linkage study) with Systems Biology by profiling the same study subjects not only for DNA variants (Single Nucleotide Polymorphisms or SNPs, Copy Number Polymorphisms) but also for genome-wide gene expression and other “omics” variables (DNA methylation, regulatory RNA expression, proteomics). GSB studies enable us to elucidate the genetic control of gene expression, regulatory RNA expression and DNA or methylation, to annotate DNA polymorphisms implicated in previous Genome-Wide Association Studies (GWAS) for particular diseases, and to infer key control genes and pathways causally underlying a disease or biomedical trait of interest. It is well-known that the expression profiles of groups of genes share common regulators (represented for example by SNPs), and these common regulators are more easily identified when associated with a group of expression profiles rather than with the expression of individual genes. Several approaches to associating DNA variants with groups of expression profiles have recently been proposed, including Geronemo/Lirnet (Su-In Lee et al. 2006/2009) which iterate between gene clustering and inferring each group’s regulators, and a Bayesian partition method (Wei Zhang et al. 2010). Canonical Correlation Analysis (CCA) is a classical multivariate approach to discover linear relationships between two sets of variables from datasets with sample sizes strongly exceeding the number of variables ($N >> p$). Recently, a probabilistic framework for CCA (Francis R. Bach and Michael I. Jordan 2005) has been developed, which enables a Bayesian approach to CCA of two or more groups of variables. Also recently, sparse Bayesian and non-Bayesian regression methods have been developed, allowing application to $N << p$ datasets. We have implemented Bayesian CCA for data integration in GSB using deterministic algorithms (e.g., a variational algorithm) and an extension to discrete variables, and we are comparing our method to non-Bayesian sparse CCA and sparse Partial Least Squares. Here we present initial results from this comparison.

**Bayesian adaptive design for Phase I/II trials** [1.2.3]

Yu Lou

*Discovery Biometrics*

*GSK*

Abstract. The cost of discovering and developing new medicines has increased dramatically in recent years. Bayesian approaches have the potential to reduce development costs...
and increase the trial success rate. These approaches are well suited to flexible development strategies tailored to meet an asset’s unique challenges (i.e., differential development). Specific questions can be addressed using much more focused studies prior to large investments into full development programs. Bayesian approaches provide a process for balancing risk mitigation and cost vs. speed of development. Specifically, Bayesian-based adaptive designs can be used in the conduct of Phase I and II clinical trials to (i) effectively incorporate into the design current knowledge on the efficacy and safety of candidate medicines; (ii) predict possible outcomes from candidate clinical trial designs and select the design yielding the highest probability of reaching a decision; and (iii) stop the trial early when accumulating data are sufficient to reach a decision. The basis for decision making must be established as part of the clinical trials design discussions. Examples based on our research and development of antiviral agents will be presented.

Multiscale geometric analysis of noisy low-dimensional point clouds in high dimensions [I.5.7]
Mauro Maggioni
Mathematics and Computer Science
Duke University

Abstract. We discuss techniques for the geometric multiscale analysis of intrinsically low-dimensional point clouds. We first show how such techniques may be used to estimate the intrinsic dimension of data sets, then discuss a novel geometric multiscale transform, based on what we call geometric wavelets, that leads to novel approximation schemes for point clouds, and dictionary learning methods for data sets. Finally, we apply similar techniques to model estimation when points are sampled from a measure supported on a union of an unknown number of unknown planes of unknown dimension.

Threshold estimation based on a p-value framework in dose response and regression settings [I.2.6]
Atul Mallik
Statistics
University of Michigan

Abstract. ...

Statistical methods for differential equation models with applications to toxicology [I.2.6]
Siddhartha Mandal
Biostatistics
UNC Chapel Hill

Abstract. Stochastic and deterministic differential equations are used to describe a wide variety of biological and physiological phenomena. For example, in Physiologically based pharmacokinetic(PBPK) models, differential equations explain the absorption, distribution,
metabolism and excretion (ADME) of a compound in the human or animal body. Usual approaches for parameter estimation in such situations include non-linear least squares and Bayesian hierarchical modeling. However, a common challenge with these problems is the lack of explicit equations/models that relate response variable to the explanatory variables. Recent functional data analysis methods indicate the use of basis functions to bypass this problem. This talk focuses on estimation and inference of the model parameters, taking into account the variability within and between multiple subjects, while exploiting the structure implied by the system of differential equations. Large sample behavior of the parameter estimates are also explored. Application of the methods are shown using simulated and real life data on compartmental and state space models.

Object oriented data analysis [I.4.6]
J. S. Marron
Statistics & O. R.
U. N. C.

Abstract. Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. Recent developments in medical image analysis motivate the statistical analysis of populations of more complex data objects which are elements of mildly non-Euclidean spaces, such as Lie Groups and Symmetric Spaces, or of strongly non-Euclidean spaces, such as spaces of tree-structured data objects. These new contexts for Object Oriented Data Analysis create several potentially large new interfaces between mathematics and statistics. Even in situations where Euclidean analysis makes sense, there are statistical challenges because of the High Dimension Low Sample Size problem, which motivates a new type of asymptotics leading to non-standard mathematical statistics.

A generalized bivariate scale-invariant test for location based on slopes [I.6.6]
Sunil Mathur
Mathematics
University of Mississippi

Abstract. A generalized test is proposed for bivariate one-sample location problem. The proposed test is based on the slopes obtained by using two variates. The asymptotic distribution of the proposed test statistic is found. The proposed test statistic does not depend on the covariance structure of the population, and is scale-invariant. The proposed test statistic is compared with Hotelling’s $T^2$ test statistic, Mardia’s (1967) test statistic, Wilcoxon’s (1945) rank sum test statistic, and Peters-Randles (1991) test statistic using Monte Carlo technique. For non-normal distributions, the proposed test statistic performs better than most of the test statistics under consideration, some of which are based on two-samples, for almost all the shifts in the location. Application of the test is also illustrated using a bivariate data set.
**Use of baseline responses in 2x2 crossover trials** [I.2.3]
Devan V. Mehrotra

*Clinical Biostatistics*  
*Merck Research Laboratories*

Abstract. The 2x2 crossover design is commonly used for both early and late stage clinical trials across multiple therapeutic areas. It is preferred over the corresponding parallel groups design on the presumption that the smaller sample size requirements for the crossover design will result in cost and/or time efficiencies. We caution that this may not always be true. Moreover, in many cases, baseline values for the response of interest are assessed prior to either the first period only or both periods of the crossover design, and the treatments are subsequently compared on the basis of within-subject changes from baseline. We take a critical look at when, why, and how baseline measures should be used, and we contrast several so-called ”baseline-adjusted” analysis options. Theory and simulations reveal that the use of baseline responses for the 2x2 crossover design is often counterproductive from a cost/precision perspective, notably so for some popular analysis options. Data from two crossover trials will be used to reinforce the key points.

**Revisiting constrained parametric regression** [I.1.3]
Mary Meyer

*Statistics*  
*Colorado State University*

Abstract. For the classical least-squares regression model it is often the case that a set of linear inequality constraints may be imposed on the parameters. Examples include shape restrictions such as when a quadratic regression function must be monotone over a range of the predictor variable, or a regression surface that must be monotone in both predictors, as well as order restrictions on a set of population means. We present a general proof that when the constraints are met, tests using constrained alternatives have higher power than the corresponding tests with unconstrained alternatives. In addition, we propose a test for the constrained versus unconstrained model, which is shown to have good large-sample properties, and simulations demonstrate that small-sample behavior compares favorably to existing tests.

**Likelihood inference for lognormal data with left truncation and right censoring with an illustration** [Poster]
Debanjan Mitra

*Mathematics and Statistics*  
*McMaster University*

Abstract. The lognormal distribution is quite commonly used as a lifetime distribution. Data arising from life-testing and reliability studies are often left truncated and right censored. Here, the EM algorithm is used to estimate the parameters of the lognormal model based on left truncated and right censored data. The maximization step of the algorithm
is carried out by two alternative methods, with one involving approximation using Taylor
series expansion (leading to approximate maximum likelihood estimate) and the other based
on the EM gradient algorithm (Lange, 1995). These two methods are compared based on
Monte Carlo simulations. The Fisher scoring method for obtaining the maximum likelihood
estimates shows a problem of convergence under this setup, except when the truncation
percentage is small. The asymptotic variance-covariance matrix of the MLEs is derived by
using the missing information principle (Louis, 1982), and then the asymptotic confidence
intervals for scale and shape parameters are obtained and compared with corresponding
bootstrap confidence intervals. Finally, some numerical examples are given to illustrate all
the methods of inference developed here.

A non-parametric approach for comparing survival distributions
with propensity score adjustment [I.7.4]
Nandita Mitra

Biostatistics
University of Pennsylvania

Abstract. The Cox proportional hazards model is most commonly used in the analysis
of censored survival data. However, in many situations, the proportional hazards model
assumptions are not satisfied. We propose a nonparametric test for the equality of the
conditional hazard function between comparison groups that allows for the adjustment of
covariates using an estimated propensity score. This conditional k-sample logrank test does
not require the assumption of proportional hazards, a parametric specification of the relative
risk function, or randomization of group assignment. We conduct a series of Monte Carlo
simulation studies to compare the operating characteristics of this statistic against the un-
conditional log-rank statistic, the conditional log-rank statistic proposed by Venkatraman
and Heller (2004), the score test from the Cox model and the robust score tests proposed
by Lin and Wei (1989) and Kong and Slud (1997). We demonstrate our approach with an
observational study of intrahepatic cholangiocarcinoma patients obtained from the Surveil-
 lance, Epidemiology, and End Results (SEER) database that compares the survival benefit
of men who received radiation therapy versus those who did not. Numerous covariates such
as race, age, and disease stage are adjusted for via the propensity score.

On threshold estimation problem in visual field analysis [I.4.4]
Debashis Mondal

Statistics
University of Chicago

Abstract. The introduction of Humphrey’s field analyzer marks an important beginning in
automated perimetry. Visual field analysis currently constitutes a significant clinical ap-
proach to detecting eye defects, including screening, diagnosing, and monitoring of retinal
and other neuro-ophthalmologic disorders. The Humphrey’s field analyzer produces a map
of threshold light intensities, which are interpreted to have a point-wise 50% chance of being
detected by the patient over the visual field. In this talk, I will discuss some details about
how measurements are taken in Humphrey’s field analyzer and provide a critical statistical analysis of the estimated light thresholds. Even though Humphrey’s field analyzer does not fully disclose how the measurements are taken and how the statistics it yields are computed, I will discuss why estimates of threshold light intensities can be a point of contention, and how estimation of threshold light intensities can benefit from existing statistical literature on threshold estimation. Next, I will present a model-based approach to estimate threshold maps of light intensities using Markov random field priors. If time permits, I will discuss further statistical challenges. This is a joint work with Guillaume Pouliot.

**Comparative effectiveness research from a statistician’s perspective**  
[Plenary I]  
Sally C. Morton  

*Biostatistics*  
*University of Pittsburgh*

Abstract. The American Recovery and Reinvestment Act (ARRA) of 2009 provided $1.1 billion for “comparative effectiveness research” or CER. Which healthcare intervention “works best, for whom, and under what circumstances?” (Clancy and Slutsky, 2009). This is the central question of CER, which encompasses all patients, including those traditionally underrepresented in clinical research such as children or individuals with comorbid conditions, and ideally helps address health disparities. CER studies may include both the generation of new evidence and the synthesis of existing evidence. As a result, CER methods comprise a variety of statistical design and analysis methodologies such as randomized controlled trials, observational studies, patient registries, adaptive designs, propensity scores, meta-analysis and modeling. In this talk, I will discuss comparative effectiveness research from a statistician’s perspective, and elucidate both its promise and challenge in helping patients, clinicians, and policy-makers reach more informed healthcare decisions.

**Bayesian semiparametric analysis of two-phase studies of gene-environment interaction**  
[I.4.2]  
Bhramar Mukherjee  

*Biostatistics*  
*University of Michigan*

Abstract. Two phase stratified sampling (Neyman, 1938) is an efficient alternative to the traditional cohort and case-control designs. Under two-phase sampling, the disease outcome (D) and some relatively inexpensive covariate (Z) information is collected in phase I. At phase II, more expensive covariate (X) information is collected on a subset selected from phase I samples. Often stratified sampling with strata defined by levels of D and Z is implemented at the second stage. Two-phase design has many potential applications in case-control studies that explore interplay of genetic and environmental risk factors (G x E studies). We present a semiparametric Bayesian approach that can handle multiple genetic factors, environmental exposures, their interaction in a retrospective likelihood-based framework under two-phase sampling. The method has an automated variable selection feature and can relax
gene-environment independence in a data-adaptive way. The methods are applied to explore
the interaction between statins and genes on the cholesterol synthesis and lipid metabolism
pathway in an ongoing study of colorectal cancer. Joint work with Jaeil Ahn, Stephen B.
Gruber and Malay Ghosh.

**Efficient bias correction for MRI image denoising** [I.4.6]
Partha Sarathi Mukherjee

*Statistics*
*University of Minnesota, Twin Cities*

Abstract. Magnetic resonance imaging (MRI) is a popular radiology technique that is used
for visualizing detailed internal structure of the body. Observed MRI images are generated
by the inverse Fourier transformation from received frequency signals of a MR scanner sys-
tem. Previous research has demonstrated that random noise involved in the observed MRI
images can be adequately described by the so-called rician noise model. Under that model,
the observed image intensity at a given pixel is a non-linear function of the true image in-
tensity and two independent zero-mean random variables with a same normal distribution.
Because of such a complicated noise structure in the observed MRI images, denoised im-
ages by conventional denoising methods are usually biased, and the bias would reduce image
contrast and affect negatively on subsequent image analysis. Therefore, it is important to
address the bias issue properly. To this end, several bias correction procedures have been
proposed in the literature. Most of them are proposed based on intuition. In this paper,
we study the rician noise and the corresponding bias-correction problem systematically, and
propose a new and more accurate bias-correction formula. Numerical studies show that it
works well in various applications.

**A tale of two manifolds** [I.5.7]
Sayan Mukherjee

*Statistical Science*
*Duke University*

Abstract. The focus is on the problem of supervised dimension reduction (SDR). We first
formulate the problem with respect to the inference of a geometric property of the data, the
gradient of the regression function with respect to the manifold that supports the marginal
distribution. We provide an estimation algorithm, prove consistency, and explain why the
gradient is salient for dimension reduction. We then reformulate SDR in a probabilistic
framework and propose a Bayesian model, a mixture of inverse regressions. In this modeling
framework the Grassman manifold plays a prominent role.

**Application of wavelets in quantification of protein co-localization
in intensity images of variable dimension** [I.7.2]
Nitai Mukhopadhyay

*Department of Biostatistics*
*Virginia Commonwealth University*
Abstract. Protein co-localization is an important tool in optical microscopy for disease diagnostics. Historically identification and quantification of protein co-localization has been done rather subjectively by experienced researchers. Manders (1993) and later Costes et al (2004) provided a statistical approach to estimate and test for significant co-localization. However, the method suggested does not apply immediately to images of unequal dimension. We illustrate the limitation of straightforward application of correlation statistic in this case, and suggest a wavelet based approach that would address the problem with wider scope. Simulated images with various degree of co-localization would be used in demonstrating the performance of the method. Also comparative analysis of the proposed method with direct correlation approach will be shown on cancer tissue image.

High dimensional classification & variable selection: comparison density approach

Subhadeep Mukhopadhyay

Statistics
Texas A&M University

Abstract. In this paper, we develop a new statistical framework for high dimensional classification using quantile machinery. The main building blocks of our approach are based on (i) establishing a connection between optimal Bayesian classifier and comparison density, (ii) estimating the comparison density using robust orthonormal score functions based on ranks through the mid-distribution transform that can handle continuous, discrete as well as categorical predictor variables, (iii) ranking and variable selection using Wilcoxon type generalized score coefficients and Kullback-Leibler information number, and (iv) using log comparison density estimates of influential variables along with highly scalable efficient multi-class classification algorithm to produce estimates of posterior probabilities. We present multiple illustrative examples to explain how this method gives various insights for pattern mining. We also compare our method with various state-of-art classification methods to demonstrate the advantages and usefulness of the proposed methodology.

Bayesian modelling and computation for one-day cricket

Saman Muthukumarana

Department of Statistics
University of Manitoba

Abstract. In one-day international cricket matches, there are an endless number of questions that are not amenable to experimentation or direct analysis but could be easily addressed via simulation. For example, on average, would England benefit from increasing the number of runs scored by changing the batting order of their third and sixth batsmen? As another example, what percentage of time would India be expected to score more than 350 runs versus Australia in the first innings? A good simulator for ODI cricket is required to provide reliable answers to such questions. In this talk, I will present a Bayesian latent variable approach to model the outcomes can occur on each ball that is balled.

Given that only a finite number of outcomes can occur on each ball that is bowled, a
A discrete generator on a finite set is developed via a Bayesian latent variable model where the outcome probabilities are estimated from historical data involving one-day international cricket matches. The conditional probabilities depend on many factors including the batsman, the bowler, the number of wickets lost, the number of balls bowled, current score of the match, the opposing team and the innings. The model is more realistic in that any batsman/bowler combinations are considered. In addition, we also provide a method of generating runs in the second innings by considering dynamic change aggressiveness of batsmen due to current situation of the match. The methodology was tested using cross-validation procedures. The proposed simulator appears to do a reasonable job at producing realistic results. The simulator allows a team to easily investigate the results of making changes to the batting and bowling orders.

**Fisher information in censored samples from the Downton bivariate exponential distribution** [I.3.5]

H. N. Nagaraja

*Division of Biostatistics, College of Public Health*

*The Ohio State University*

Abstract. We develop a simple approach to finding the Fisher information matrix (FIM) for a single pair of order statistic and its concomitant, and Type II right, left, and doubly censored samples from an arbitrary bivariate distribution. We use it to determine explicit expressions for the FIM for the three parameters of the Downton bivariate exponential distributions for Type II censored samples. We evaluate the FIM for various parameter values and sample sizes, and determine its limiting form as the sample size increases. We discuss implications of our findings to inference based on censored samples from the Downton bivariate exponential distribution. (Joint work with Dr. Qinying He, Southwestern University of Finance and Economics, Chengdu, Sichuan, China)

**Estimation of parameters in the context of repeated canonical correlation analysis** [I.5.4]

Dayanand Naik

*Mathematics & Statistics*

*Old Dominion University*

Abstract. We address the problem of canonical correlation analysis of longitudinal data when the data have a Kronecker product covariance structure. Using structured correlation matrices we model the dependency of repeatedly observed data. Recent work of Srivastava, Nahtman, and von Rosen (2008, Mathematical Methods of Statistics, 17(4), 357-370) developed an iterative algorithm to determine the maximum likelihood estimate of the Kronecker product covariance structure for one set of variables. We implement and generalize their method to estimate the covariance parameters in the context of canonical correlation analysis. Testing of hypothesis problems using the likelihood ratio test statistics is also explored. This is a joint work with Dr. Ray McCollum.
Bootstrap based preliminary test estimation for the parameter of power series distribution [C.7.7]
K. M. Nalawade

*Dhananjayrao Gadgil College of Commerce*
*Affiliated to Shivaji University, Kolhapur*

Abstract. A problem of estimation of the parameter of Power Series Distribution, in the presence of a prior point information about the parameter is considered. Asymptotic Preliminary Test Estimator (APTE) for the parameter is proposed. APTE is a conditional estimator and depends on the weighted estimator of Maximum Likelihood Estimator (MLE) and prior point information. The weight is usually considered to be a constant free from the observed data. Here I discuss a bootstrap based estimator of the weight and study performance of such APTE. Simulation study for Binomial and Poisson distributions shows that Bootstrap-based APTE performs better than the MLE for a wide range of parameter space. Some examples are provided to illustrate the proposed method.

Parametric inference for system lifetime data with signatures available [I.3.3]
Tony Ng

*Department of Statistical Science*
*Southern Methodist University*

Abstract. In this talk, the statistical inference of the lifetime distribution of component based on observing the system lifetimes with signature available is discussed. A general proportional hazard rate model for the lifetime of the components is considered, which includes some commonly used lifetime distributions. Different estimation methods for the proportional parameter are discussed. Monte Carlo simulation study is used to compare the performance of these estimation methods and recommendations are made based on these results.

Dynamic software solutions for clinical data analysis, review and reporting [I.3.2]
Michael O’Connell

*Analytics*
*TIBCO*

Abstract. Dynamic software tools for visualization and predictive analytics enable proactive decision making, rather than retrospective tracking of results. This shortens the time between critical development gates and enables key milestones to be met sooner, while simultaneously managing safety risk.

This presentation covers the use of dynamic software tools and predictive analytics for review/analysis of both clinical/safety and trial management/operations data. This includes (a) in-stream and unblinded review of adverse events, labs, vitals etc. with in-built statistical analysis for assessing AE treatment emergence; and (b) prediction of trial enrolment and re-
source management along with site monitoring and management for effective trial operations.

**Bayesian nonparametric regression models for right-censored data**  
[Poster]  
Muhtarjan Osman  
*Statistics*  
*North Carolina State University*  

Abstract. The celebrated Cox's proportional hazard model is one of the most frequently used regression models in survival analysis. In some data situations, however, assuming proportional hazard is not realistic. For example, the case with crossing survival curves certainly violates the proportional hazard assumption. In this paper, we proposed a Bayesian nonparametric model for the conditional hazard function based on Bernstein polynomials. The most important feature of the proposed Bayesian method is its capability of modeling the uncertainty about the order of Bernstein polynomials using reversible jump Markov Chain Monte Carlo (RJ-MCMC). This has been a great challenge for the proposed method from a frequentist perspective because of the lack of cross-validation procedures for censored data in regression settings. The method is further illustrated using both simulated data and the data from two cancer trials.

**The Aldous diffusion on continuum trees** [I.2.1]  
Soumik Pal  
*Mathematics*  
*University of Washington*  

Abstract. Consider a tree whose every non-leaf vertex has degree three (binary branching). A Markov chain runs on the space of such trees by picking a leaf at random, removing its edge and the branch point, and reinserting them on a randomly chosen edge among those remaining. David Aldous in his 1999 ICM address asks what would be a "diffusion" limit of such a chain on the space of continuum trees. We will show facets of this limiting diffusion and its various properties.

**Portfolio optimization with CVaR: a copula based method** [I.6.7]  
Tao Pang  
*Mathematics*  
*North Carolina State University*  

Abstract. In this paper, we consider the portfolio optimization problem with Conditional Value-at-Risk (CVaR) as the risk measurement. Our objective is to minimize the CVaR of the portfolio under certain constrains. Copula based risks have been taken into consideration. Further, we consider a time series heteroscedastic model under the GARCH framework for this optimization problem. We will find the efficient frontier for the portfolio when the marginal distribution of each asset is generalized skewed t distribution.
Bayesian closed surface fitting through tensor products [I.5.7]
Debdeep Pati
Statistical Science
Duke University

Abstract. Closed surfaces provide a useful model for 3-d shapes, with the data typically consisting of a cloud of points in $\mathbb{R}^3$. The existing literature on closed surface modeling focuses on frequentist point estimation methods that join surface patches along the edges, with surface patches created via Bézier surfaces or tensor products of B-splines. However, the resulting surfaces are not smooth along the edges and the geometric constraints required to join the surface patches lead to computational drawbacks. In this article, we develop a Bayesian generative probability model for closed surfaces, which accounts for uncertainty in choosing coordinates and generating the surface given these coordinates. The model is based on tensor products of a cyclic basis with the control points modeled through a tensor decomposition resulting in infinitely smooth surface realizations. We impose sparsity on the tensor fibers providing a sparse representation of the control points. We avoid joining of the surface patches by the use of cyclic bases developed by Róth et al. 2009. Theoretical properties of the support of our proposed prior are studied and it is shown that the posterior achieves the optimal rate of convergence under reasonable assumptions on the prior. Several simulation examples are considered and the methods are applied to real data.


Bivariate zero inflated power series distribution [C.1.1]
M. K. Patil
P.V.P.Mahavidyalaya, India

Abstract. Many researchers have discussed zero-inflated univariate distributions. These univariate models are not suitable, for modeling events that involve different types of counts or defects. To model several types of defects, multivariate Poisson model is one of the appropriate models to use. This can further be modified to incorporate inflation at zero and we can have multivariate zero-inflated Poisson (MZIP) distribution. There are several ways to construct MZIP distributions. Ching-Shang et al. (1999) have discussed various types MZIP models and investigated their distributional properties. It appears that majority of the study in the literature is restricted to Poisson distribution and its extension to multivariate set up. Relatively less has been reported for the family of distributions containing other distributions. In the present article, I introduce a new Bivariate Zero Inflated Power Series Distribution and discuss inference related to the parameters involved in the model. I also discuss the inference related to Bivariate Zero Inflated Poisson Distribution (BZIPD). The model has been applied to a real life data. Extension to k variate zero inflated power series distribution is also discussed.

Nonparametric inference using additional information [C.7.7]
Pitambar Y. Patil

Department of Statistics, Devchand College, Arjunnagar, Kohlapur, India
Shivaji University, Kohlapur, India

Abstract. Vardi (1982) considered the problem of estimation of cumulative distribution function (c. d. f.) $F$ using the additional information through length biased variables. In this paper we consider the same problem, when additional information is through a c. d. f. $G(\cdot)$, that is a functional of $F$. For example particular form of $G$ may be a power of $F$ or a truncated / censored version of $F$. One can also address the related two sample problems. The performance of the procedures will be carried out may be based on extensive simulations.

Measuring asymmetry [I.6.4]
Prakash N Patil

School of Mathematics
The University of Birmingham

Abstract. It is general practice to make assertions about the symmetry or asymmetry of a probability density function based on the coefficients of skewness. Since most of the coefficients of skewness are designed to be zero for a symmetric density, they do, overall, provide an indication of symmetry. However, skewness is primarily influenced by the tail behavior of a density function, and the skewness coefficients are designed to capture this behavior. Therefore, they do not calibrate asymmetry in the density curves. To measure asymmetry in the density curves, we first provide two new symmetry conditions for a continuous probability density function; of the two new conditions, one is only a necessary condition, while the other is necessary and sufficient. These conditions are then used to produce weak and strong asymmetry measures of a continuous probability density function on the scale of -1 to 1. We show through examples that both measures do an admirable job of capturing the visual impression of asymmetry of a continuous density function, and briefly discuss the estimation of these coefficients from a sample.

Regularized Hotelling’s $t^2$ for high-dimensional hypothesis tests [I.7.3]
Debashis Paul

Statistics
University of California, Davis

Abstract. We consider the problem of testing for the difference between means for two populations when the observations are samples from multivariate normal distributions. A classical test procedure for this hypothesis is based on the Hotelling’s $T^2$ statistic which involves the inverse of the sample covariance matrix. Under settings where the sample size is smaller than the dimensionality of the observations, this statistic is not defined. A regularized version of the test is considered which is applicable regardless of the dimensionality. We investigate asymptotic properties of the test statistic and compare the test performance with a number of existing methods through simulation examples. We also formulate a nonparametric version of the test, to deal with the problem of missing data as well as the issue of robustness. The tests are shown to effectively control the Type I error rate and maintain a good power.
property in the presence of complex correlation structures.

**Isotropic order among a core set of cell-cycle transcripts is conserved between budding and fission yeast** [I.7.1]

Shyamal D. Peddada

*Biostatistics Branch*

*NIEHS*

Abstract. A cell division cycle in eukaryotes is a well coordinated process where cells go through four distinct phases before dividing. Genes with periodic expression are known as cell-cycle genes. To discover evolutionarily conserved cell-cycle genes, biologists are often interested in identifying genes that are periodic in multiple organisms. There has been considerable discussion in the literature on genes that are periodic in fission yeast Schizosaccharomyces pombe (S. Pombe) as well as in budding yeast Saccharomyces cerevisiae (S.Cerevisiae). Recent microarray studies suggest that very few genes are periodic in both yeasts. A conservative estimate of the number of genes in this "core set" is as small as 35. Phase angle of peak expression of a cell-cycle gene is of great interest since it may be associated with the gene’s function during the cell division cycle. Since the cell division cycle is a carefully orchestrated process, one may hypothesize that the relative order of time to peak expression (known as the isotropic order) among the core set of genes is conserved in the two species of yeast. Using a novel statistical methodology, we demonstrate that the isotropic order is conserved for at least 28 out of 35 (80%) genes in the core set. On the other hand, very few fission yeast genes (about 33%) appear to satisfy the isotropic order of the corresponding human orthologs. We discover seven cell-cycle genes whose relative order is potentially conserved between humans, fission and budding yeasts. Methodology described in this talk could help biologists formulate and test other similar hypotheses.

**Estimating and improving power in genome-wide association studies** [I.7.1]

Vasyl Pihur

*Department of Biostatistics*

*Johns Hopkins University*

Abstract. Genome-wide association studies have been successful in identifying thousands of common variants associated with complex traits. Exploring human genome in an unbiased way for possible associations, however, comes at a cost of performing millions of tests. While controlling the error rates is of primary importance, maximizing power should not be overlooked. In the first part of the talk, I will present a novel iterative Benjamini-Hochberg method for improving power in multiple testing via estimating the number of null hypotheses among all hypotheses being tested. The second part of the talk will address the issues of estimating power for next-generation genome-wide association studies by proposing a simple model for the effect sizes of less common variants. Having a model for the effect size allows to assess how much power a study with a given sample size has, as well as answer the question of how much more trait variability these less frequent variants could potentially
**Diffusion limits and optimal scaling of Markov Chain Monte Carlo algorithms** [I.4.1]

Natesh Pillai

*Department of Statistics*

*Harvard University*

Abstract. MCMC (Markov Chain Monte Carlo) algorithms are an extremely powerful set of tools for sampling from complex probability distributions. Understanding and quantifying their behavior in high dimensions thus constitute an essential part of modern statistical inference. In this regard, most of the research efforts so far were focused on obtaining estimates for the mixing times of the corresponding Markov chain. In this talk we offer a new perspective for studying the efficiency of commonly used algorithms. We will discuss optimal scaling of MCMC algorithms in high dimensions where the key idea is to study the properties of the proposal distribution as a function of the dimension. This point of view gives us new insights on the behavior of the algorithm, such as precise estimates of the number of steps required to explore the target distribution, in stationarity as a function of the dimension of the state space. In the first part of the talk, we will describe the main ideas and discuss recent results on high dimensional target measures arising in the context of statistical inference for mathematical models representing physical phenomena. In the second part of the talk, we will discuss the Hybrid Monte Carlo Algorithm (HMC) and answer a few open questions about its efficiency in high dimensions. We will also briefly discuss applications to parallel tempering, Gibbs samplers and conclude with concrete problems for future work.

**Testing for modality, residual empirical process and weighted sums for time varying processes** [I.5.3]

Wolfgang Polonik

*Statistics*

*University of California at Davis*

Abstract. In the context of a time-varying AR-model, we construct a test for modality of the variance function that under appropriate assumptions is asymptotically distribution-free, even though non-parametric estimation is involved. Simulation studies and applications to real data sets illustrate the behavior of the test. The large sample analysis of the test statistic hinges on properties of both residual empirical processes and weighted sum processes based on non-stationary (time varying) time series and indexed by function classes. The theoretical properties of these processes are of independent interest. For instance, it will be shown that under appropriate assumptions non-parametric estimation does not influence the asymptotic distribution of the residual process. An exponential inequality for weighted sums of time-varying processes is also derived.

This is joint work with G. Chandler, Pomona College.

**Weakly dependent time series and application to extremes** [I.6.2]
Abstract. Politis and Romano (1994) established the subsampling estimator for converging statistics when the underlying sequence is strongly mixing. Bertail et al. (2004) applied this work to subsampling estimators for distributions of diverging statistics. In particular, they constructed an approximation of the distribution of the sample maximum without any information on the tail of the stationary distribution. However, the assumption on the strong mixing properties of the time series is sometimes too strong as for the class of first-order autoregressive sequences with uniform marginal distribution introduced and studied by Chernick (1981): let \((X_t)_{t \in \mathbb{Z}}\) be uniform AR(1) process defined recursively as

\[
X_t^{(r)} = \frac{1}{r}X_{t-1}^{(r)} + \varepsilon_t,
\]

where \(r \geq 2\) is an integer, \((\varepsilon_t)_{t \in \mathbb{Z}}\) are iid and uniformly distributed on the set \(\{0, 1, \ldots, r-1\}\) and \(X_0^{(r)}\) is uniformly distributed on \([0, 1]\). Andrews (1984) and Ango-Nze and Doukhan (2004) give arguments to derive that such models are not mixing. The results of Bertail et al. (2004) can not be used for this class of processes although the normalized sample maximum has a non-degenerate limiting distribution.

Let \(M_t = \max(X_1, \ldots, X_t)\) be the sample maximum defined for the process \((X_t)_{t \in \mathbb{Z}}\) given in (1). Chernick (1981) shows that different limit than the one of the three extreme value type distributions can be expected for the appropriately normalized \(M_t\):

\[
\lim_{t \to \infty} \Pr(t(1 - M_t) \leq x) = 1 - \exp\left(-\frac{1}{r}(r-1)x\right), \quad \forall x \geq 0,
\]

(see Theorem 4.1 in Chernick (1981)).

This paper is aimed at weakening the strong mixing condition assumed in Bertail et al. (2004). We work under the weak dependence condition for time series in terms of definition given in Doukhan and Louhichi (1999).

Doukhan and Louhichi (1999) introduced a weak dependence framework that turns out in particular to apply to the previous processes and that widely improves the amount of potentially usable models. This dependence structure is addressed in the paper. We introduce smooth and rough subsampling estimators for the distribution of converging statistics and studied their asymptotic properties. We consider two subsampling schemes based on overlapping and non-overlapping samples. Further, we consider subsampling estimators for the distribution of extremes and, to fix ideas, we focus on the case of the normalized sample maximum. We first discuss sufficient conditions adapted to our weak dependence framework such that the normalized maximum converges in distribution. Then we discuss how to estimate the normalizing sequences and we derive the asymptotic properties of the subsampling estimators. A simulation study provides explicit comparisons of the various considered subsamplers in last section.

**ROC modeling for some of the life distributions** [C.7.6]

Sudesh Pundir

\[Pundir\]

55

\[Pundir\]
Statistics
Pondicherry University, India

Abstract. ROC model for the Biweibull, Birayleigh, Bilogistic and Bipareto distributions are fitted and their Area Under the ROC Curves (AUC) are found. The properties of all these distribution are studied and compared with Binormal and Biexponential distributions. These theoretical results are supported by one simulation exercise and real life data analysis.

Bayesian model-based methods for analyzing RNA-seq data [I.2.2]
Steve Qin
Biostatistics and Bioinformatics
Emory University

Abstract. RNA sequencing (RNA-seq) is a powerful new technology for mapping and quantifying transcriptome using ultra high-throughput next generation sequencing technologies. Using deep sequencing, gene expression levels can be quantified thus providing a digital measure of the presence and prevalence of all transcripts including novel ones. Although extremely promising, the massive amounts of data that are generated by RNA-seq, substantial biases, and uncertainty in short read alignment pose daunting challenges for data analysis. In particular, large base-specific variations and between-base correlations make naive approaches, such as averaging to normalizing RNA-seq data and quantifying gene expressions, ineffective. We propose to develop Poisson mixed effects models to characterize RNA-seq data. These models will accommodate the biases, variations, and correlations present in RNA-seq data so as to accurately estimate gene expression levels and to facilitate gene expression comparison and novel transcript structure or activities discovery. This is a joint work with Ming Hu and Michael Yu Zhu.

3-D image denoising by local smoothing and nonparametric regression [I.4.6]
Peihua Qiu
School of Statistics
University of Minnesota

Abstract. Three-dimensional (3-D) images are becoming increasingly popular in image applications, such as magnetic resonance imaging (MRI), functional MRI (fMRI), and other image applications. Observed 3-D images often contain noise that should be removed beforehand for improving the reliability of subsequent image analyses. In the literature, most existing image denoising methods are for 2-D images. Their direct extensions to 3-D cases generally can not handle 3-D images efficiently, because the structure of 3-D images is often substantially more complicated than that of 2-D images. For instance, edge locations are surfaces in 3-D cases, which are much more challenging to handle, compared to edge curves in 2-D cases. In this paper, we propose a novel 3-D image denoising procedure, based on nonparametric estimation of a 3-D jump surface from noisy data. One important feature of this method is its ability to preserve edges and major edge structures, such as intersections of two edge surfaces, pyramids, pointed corners, and so forth. Both theoretical arguments
and numerical studies show that it works well in various applications. Software and proofs are available online as supplemental material.

**Exact tests of Hardy-Weinberg equilibrium via algebraic statistics and Markov bases** [I.6.6]

M B Rao

*Environmental Health*
*Cincinnati*

Abstract. In the context of large contingency tables with sparse data, algebraic statistics methodology is successfully implemented to conduct an exact test of independence a la Fisher. In this presentation, the focus is on testing Hardy-Weinberg Equilibrium in multi-allelic markers a la Fisher. We demonstrate how a combination of algebraic statistics, Markov bases, and MCMC helps us achieve our objectives. The key idea is to look at the fiber of all possible genotype data sets with the same allele frequencies as those of the given genotype data set. Even for moderate sizes of alleles, the fiber could be humongous. We will construct a Markov base for the fiber so that the base in combination of the given data set will generate the entire fiber. Typically, Markov bases are much smaller in size compared with the size of the fiber. Armed with the given data genotype data set and a Markov base, an MCMC algorithm can be set up to estimate the p value under the null hypothesis.

**Quenched free energy and large deviations for random walks in random environments and random potentials** [I.2.1]

Firas Rassoul-Agha

*Mathematics*
*University of Utah*

Abstract. We prove a process-level large deviation principle for quenched random walk in random environment subject to a random potential. In particular, both quenched random walk in random environment and quenched polymers in a random potential are covered. The walk lives on a square lattice of arbitrary dimension and has an arbitrary finite set of admissible steps. The restriction needed is on the moment of the logarithm of the transition probability and the potential, in relation to the degree of mixing of the ergodic environment. The rate function is an entropy and two variational formulas are given for the free energy.

**On form-invariance of weighted distributions and their statistical manifolds** [C.7.7]

Makarand V. Ratnaparkhi

*Mathematics and Statistics*
*Wright State University*

Abstract. Let $f(x, \theta)$ be the p.d.f. of a non-negative r.v. $X$ with parameter where $\theta$ is a scalar or a vector of parameters. Let $w(x) > 0$ be a function of $x$ such that $E[w(X)] < \infty$. Then, $g(x, \theta) = w(x)f(x, \theta)/E[w(X)]$ is referred to as the weighted distribution of $X$. Furthermore,
if \( g(\cdot) \) has the same mathematical form as that of \( f(\cdot) \) then, it is said to be form-invariant. In this presentation, for certain well-known distributions that are used for data analysis, it is shown that the form-invariance of \( f(x, \theta) \) does not necessarily imply the form-invariance of the statistical manifold, say \( \mathcal{M}_\theta \), of \( f(x, \theta) \) with respect to the differential geometrical properties such as Amari’s \( \alpha \)-curvature tensor, Rao’s distance and the existence of sub-manifolds. Therefore, a due care becomes necessary in the statistical inference on the parameters of \( f(x, \theta) \). The results presented in this presentation can be extended to the bivariate distributions.

**Voter models and their scaling limits [I.6.7]**

K. Ravishankar

*Mathematics*

*SUNY-New Paltz*

Abstract. I will start with an overview of voter models and their connections to Ising and Potts models. I will then describe the scaling limits of the voter model, Ising and the Potts models which are respectively the Brownian web, Marked Brownian web, and branching Brownian net with killing. If time permits I will indicate some of the relevant ideas needed for proving the scaling limits.

**Functional factor analysis for periodic data [I.5.2]**

Surajit Ray

*Dept. of Mathematics and Statistics*

*Boston University*

Abstract. Functional data analysis techniques are used to analyze ecosystem data. The goals are (a) to identify and interpret the principal sources or variation driving the dynamics of ecosystem data (b) decompose these sources of variation into annual and extra annual components with a goal towards detection of ecosystem change. After preliminary smoothing we apply standard functional principal components to extract the principal sources of variation. Then we consider periodicity feature in the functional factor and extract this feature from the principal component factors by rotation. We demonstrate that this can be achieved efficiently using functional canonical correlations analysis with a space of periodic functions. This allows us to decompose functional variation into nearly-periodic and nearly a-periodic parts where a novel “elbow”

\[ \text{é} \text{riterion} \text{ is used to choose a trade-off. The stability of the extracted period components is demonstrated through a simulation study and a test is developed for the periodicity of the first } k \text{ principal components. We motivate and demonstrate this technique on a vegetation index data set.} \]

**A finite-dimensional construction of a max-stable process for spatial extremes [I.5.5]**

Brian Reich
Statistics
North Carolina State University

Abstract. From heat waves to hurricanes, often the environmental processes that are the most critical to understand probabilistically are extreme events. Such extremal processes manifestly exhibit spatial dependence. Max-stable processes are a class of asymptotically-justified models that are capable of representing spatial dependence among extreme values. While these models satisfy modeling requirements, they are limited in their utility because their corresponding joint likelihoods are unknown for more than a trivial number of spatial locations, preventing, in particular, Bayesian analyses. In this paper we propose an approximation to the Gaussian spatial max-stable process (sometimes called the “Smith process”) that, critically, is amenable to standard MCMC and inclusion in hierarchical models. We show that this model is max-stable and approximates the Smith process arbitrarily well. The proposed model also leads to a non-stationary extension, which we use to analyze the yearly maximum temperature in the southeast US for years 1983–2007.

Inferences for bivariate negative binomial models with applications to multiple sclerosis clinical trials [I.7.2]
Mallikarjuna Rettiganti

Pediatrics
University of Arkansas for Medical Sciences

Abstract. New magnetic resonance imaging (MRI) based brain lesion counts are being widely used in clinical trials in MS as the response variable of interest. We use a bivariate negative binomial distribution to model these count data from parallel group trials with a baseline correction (PGB) for relapsing remitting multiple sclerosis (RRMS) patients. In PGB trials, an additional baseline observation (X) is taken in addition to the counts seen in the follow-up period (Y). This is done to correct for any significant differences between the two groups even before the treatment is initiated.

Using the joint BNB distribution of (X,Y), we describe the associated likelihood ratio (LR), score, and Wald tests for testing the treatment effect. We obtain simulation based exact percentiles for the test statistics that maintain Type I error levels when the asymptotic tests fail to do so. We perform power analyses and sample size estimation using the simulated percentiles of the exact distribution of the test statistics. We also derive the distribution of the difference Y-X when (X,Y) is distributed according to a BNB distribution and show that the effect size for comparing between the groups based on the difference does not always increase monotonically with increasing follow-up period.

When compared to the nonparametric test, LR test based on the joint distribution results in 70-80% reduction in sample sizes and one of the Wald tests provides a further reduction of 25%.

Spectral analytic comparisons for data augmentation [I.5.2]
Vivekananda Roy

Department of statistics

Rettiganti 59 Roy
**Iowa State University**

Abstract. The data augmentation (DA) algorithm, though very useful in practice, often suffers from slow convergence. Hobert and Marchev (2008) recently introduced an alternative to DA algorithm, which we call sandwich DA (SDA) algorithm since it involves an extra move that is sandwiched between the two conditional draws of the DA algorithm. The SDA chain often converges much faster than the DA chain. In this paper we consider theoretical comparisons of DA and SDA algorithms. In particular, we prove that SDA is always as good as DA in terms of having smaller operator norm. If the Markov operator corresponding to the DA chain is compact and the extra move that is required in SDA is idempotent, which is often the case in practice, then the SDA is also compact and the spectrum of the SDA dominates that of the DA chain in the sense that all (ordered) eigenvalues of SDA are smaller than or equal to those corresponding eigenvalues of DA. We also present a necessary and sufficient condition that the extra move in SDA should satisfy for the operator norm of SDA to be strictly less than that of DA. We then consider some examples.

**A full-scale approximation of covariance functions for large spatial data sets** [I.1.4]

Huiyan Sang

*Statistics*

*Texas A&M University*

Abstract. Gaussian process models have been widely used in spatial statistics but face tremendous computational challenges for very large data sets. The model fitting and spatial prediction of such models typically require $O(n^3)$ operations for a data set of size $n$. Various approximations of the covariance functions have been introduced to reduce the computational cost. However, most existing approximations can not simultaneously capture both the large and small scale spatial dependence. A new approximation scheme is developed in this paper to provide a high quality approximation to the covariance function at both the large and small spatial scales. By combing the reduced rank representation and sparse matrix techniques, our approach allows for efficient computation for maximum likelihood estimation, spatial prediction and Bayesian inference. We illustrate the new approach with simulated and real data sets.

**Adaptive bh procedures controlling the FDR under positive dependence and improved power** [I.4.3]

Sanat K. Sarkar

*Statistics*

*Temple University*

Abstract. While a number of adaptive versions of the Benjamini-Hochberg (BH) method have been proposed in the literature with the idea of tightening its control over the false discovery rate (FDR) and thereby improving its power, most of these adaptive methods have been developed utilizing information about the number of true null hypotheses. This paper introduces some new adaptive BH methods that not only incorporate information about the
number of true null hypotheses but also utilize in a certain way the dependence structure. The ultimate control of the FDR by these new adaptive BH methods has been proved under some form of positive dependence among the underlying test statistics. Results of simulations studies indicating superior performance of the proposed adaptive BH methods over the existing ones under some dependence scenarios are presented.

**An empirical likelihood approach to goodness of fit testing** [I.3.1]
Anton Schick

*Mathematical Sciences*
*Binghamton University*

Abstract. Motivated by applications to goodness of fit testing, the empirical likelihood approach is generalized to allow for the number of constraints to grow with the sample size and for the constraints to use estimated criteria functions. The latter is needed to handle naturally occurring nuisance parameters. A central limit theorem is proved to deal with quadratic forms based on random vectors of increasing dimensions. This result is needed to prove the appropriate Wilks theorems. The proposed empirical likelihood based goodness of fit tests are asymptotically distribution free. For univariate observations, tests for a specified distribution, for a distribution of parametric form, and for a symmetric distribution are presented. For bivariate observations tests for independence, for spherical symmetry, and for equal marginals are developed.

**Loser’s hex: an evaluation of NFL team drafting performance** [I.6.3]
Michael E. Schuckers

*Mathematics, Computer Science and Statistics*
*St Lawrence University*

Abstract. Each year the National Football League, a professional American Football league, allocates newly eligible players among its 30 teams via a draft. This paper presents an evaluation of how well each of the league’s teams have done from 1991 to 2001. To arrive at our assessment, we derive a non-parametric regression based alternative to the PVC based upon past player performance. This assessment is done based upon data for each of the first 255 players selected in the National Football Leagues annual draft from 1991 to 2001. Using LOESS regression, we evaluate how each draft pick did relative to what would be expected at that selection number. Thus, we can evaluate if there are teams that systematically outperform or underperform give the draft selections that they made. We present results on individual players as well as results for each team.

**Nonparametric least squares estimation of multivariate convex regression function** [I.2.6]
Emilio Seijo

*Columbia University*
Abstract.

**Bayesian planning and inference of a progressively censored sample from linear failure rate distribution** [I.3.3]

Ananda Sen  
*Department of Family Medicine  
University of Michigan*

Abstract. In life testing and reliability studies, linear failure rate (LFR) distributions are useful in modeling the life length of a system or component when failures occur at random as well as from aging or wear-out. This two parameter model in the increasing failure rate class is a simple special case of the “polynomial hazard function” models and constitutes a generalization of the exponential distribution in a direction distinct from the gamma and Weibull. Motivation of the LFR model and its applications in engineering and biomedical context have been demonstrated aptly in the literature. In this talk we undertake the study of the planning and parameter inference for a LFR under a progressive censoring scheme. We propose a Bayesian framework which unifies both Type I and Type II censoring schemes. Under independent gamma prior specification for the parameters, the joint posterior turns out to be of the form of a finite mixture that makes simulation based computations for estimation and prediction easy. Due to the multimodal nature of the posterior, it is difficult to obtain a highest posterior density (HPD) credible region. Instead we appeal to some distributional feature to arrive at a joint credible set of the form of a trapezoid. Apart from inference, optimal sampling schemes under a dispersion criteria have been obtained through extensive numerical computation. Mortality data from a study of breast cancer patients has been used to implement our findings.

**Bootstrap in some non-standard problems** [I.5.3]

Bodhisattva Sen  
*Department of Statistics  
Columbia University*

Abstract. The talk will consider some issues with the consistency of different bootstrap methods in non-standard problems. The non-standard problems are characterized by non-\(\sqrt{n}\) rate of convergence and/or non-normal limit distributions. The talk will focus attention in problems characterized by two different rates of convergence, namely, \(n^{1/3}\) and \(n\). In the first problem we discuss the Grenander estimator, the nonparametric maximum likelihood estimator of an unknown non-increasing density function \(f\) on \([0, \infty)\). It is a prototypical example of a class of shape constrained estimators that converge at rate \(n^{1/3}\). In the second problem we consider a simple change point regression model where the estimate of the jump-discontinuity converges at rate \(n\). A comparison of the two examples sheds light on the (in)-consistency of different bootstrap procedures.

**Development and management of national health plans: health economics and statistical perspectives** [I.1.5]
Pranab K. Sen  
*Biostatistics*  
*University of North Carolina, Chapel Hill*  

Abstract. There is an innumerable number of strings, not all ascribable to causal factors nor all working in synergy or reconcilably, which undermine the complex of national health spectrum, its relation to global health, its diversity in the inclusion of people from all walks of life and society, its support by a matching national health economics plan, its impact on Quality of Life and its aim to avert eco-environmental imbalance to disasters. To fathom into this high-dimensional opaque simplex, there is a genuine need for implementing an interdisciplinary approach wherein health economics, socio-economics and geo-politics are to be properly blended with information and bio-technology, a battery of health sciences, and above all, statistical science in its total breadth and diversity in such interdisciplinary setups. Routine use of statistical packages as commonly perceived in data mining approaches may be highly inadequate to deleterious. Finite economic resources and infinite demand for health care are the major impasses. A composite narrative of this challenging task along with some basic statistical considerations are appraised. Conventional statistical approaches are of very limited utility in the development and management of such national health plans.

**Recent results on preferential attachment graphs** [I.3.4]  
Sunder Sethuraman  
*Mathematics*  
*Iowa State University*  

Abstract. We discuss recent work on the structure of the degree structure of a class of time-inhomogeneous preferential attachment graphs. In particular, large deviations, LLNs and other asymptotics will be presented.

**On complexity and correctness of Bethe approximation** [I.1.2]  
Devavrat Shah  
*EECS*  
*MIT*  

Abstract. Motivated to understand if Belief Propagation (BP) is indeed a computationally simple heuristic, we study the complexity of finding a fixed point of the BP described as the so called Bethe Approximation. For the #P complete problem of counting independent set, we establish that indeed it is possible to design a randomized polynomial time approximation scheme. We also characterize the error in approximation based on such a fixed point using the recently developed “loop series” characterization of Chertkov and Chernyak. As an interesting consequence, we establish that either the Bethe approximation is exceedingly good for a random 3-regular graph or the Shortest Cycle Cover Conjecture of Alon and Tarsi (1985) is false.

Based on joint work with Chandrasekaran, Chertkov, Gamarnik and Shin.

**Doubly-inflated Poisson (DIP) and related regression models** [C.7.5]  
Sethuraman  
*Sheth*
Manasi Sheth  
*Mathematics and Statistics*  
*Old Dominion University*  

Abstract. Usually over-dispersed count data consists of certain values occurring more frequently than allowed by the common parametric families of distributions. For dealing with zero-inflated count data with covariates, Lambert (Technometrics, 1992, pp. 1 – 14) proposed the zero-inflated Poisson (ZIP) regression model. In this talk, we introduce Doubly Inflated Poisson (DIP) models for count data situations where there is another inflated value \( k \neq 0 \). We present the distributional properties of these models. For the data consisting of un-grouped as well as grouped frequencies, with and without covariates, we discuss parameter estimation using maximum likelihood (ML) and method of moments. Asymptotic and small sample comparisons show that the ML estimators are far superior than the moment estimators. We illustrate the application of DIP models to analyze length of stay (LOS) in the hospital for patients with pancreatic disorder. Parameter estimation and analysis of count data with a negative binomial data-generating process will also be presented using a Doubly Inflated Negative Binomial (DINB) Model.

Nonparametric Shewhart-type synthetic control charts [I.6.5]  
D. T. Shirke  
*Department of Statistics*  
*Shivaji University, Kolhapur (MS) India 416 004*  

Abstract. Nonparametric control charts for detecting shifts in the location parameter or increase in the process variation are available in the literature. These charts are easy to use but their performance is inferior to their corresponding parametric charts. Many of the existing Shewhart-type nonparametric control charts can be improved in their performance by combining these charts with the Conforming Run Length (CRL) control chart (Bourke (1991)) leading to Synthetic Control Charts. In the present talk, I provide a unified approach for improving a Shewhart-type nonparametric control charts by combining them with a CRL chart. Design and implementation of these charts are discussed. Performance study of nonparametric synthetic charts based on sign statistic, signed-rank statistic and Mood statistic will be presented. Robustness study of these charts, their steady state performance and some extensions to multivariate process control will also be provided.

Optimum designs in linear mixture model with synergistic effects [I.6.4]  
Bikas K. Sinha  
*Bayesian and Interdisciplinary Research Unit (BIRU)*  
*Indian Statistical Institute*  

Abstract. In a mixture experiment, the response depends on the relative proportions of the components in it. Examples of mixture experiment can be found in industries manufacturing products like paint, glass, polymers, alloys, etc., where the quality of the product depends on its composition. Scheffe (1958, 1963) first introduced canonical models of different de-

*Shirke*  

64  

*Sinha*
degrees to represent the response function in a mixture experiment and gave designs suitable for estimation of the regression coefficients. He introduced the Simplex Lattice Designs and Simplex Centroid Designs in such situations. An overview of different mixture models with the study of estimability of the underlying parameters is given in Sinha et al (2009). Optimality of mixture designs for the estimation of parameters of the response function was considered by Kiefer (1961), Farrel et al. (1967), Atwood (1969), Galil and Kiefer (1977), Liu and Neudecker (1995), among others. Draper and Pukelsheim (1999) established the optimality of Weighted Centroid Designs with respect to Partial Loewner Ordering (PLO) for two and three component mixtures. Optimum designs for the estimation of some specific non-linear functions of the parameters have also been studied. (See Pal and Mandal, 2006, 2007, 2008, Mandal and Pal, 2008, Mandal et al., 2008a and 2008b.) The occurrence of a component exhibiting linear blending might be expected in mixture experiments where some of the components serve as diluents. The diluents have the effect of diluting the mixture in the sense that as their proportions in the mixture increase the effects of the remaining components, in terms of their proportions, diminish. Thus, the mean response changes linearly as a diluent varies between 0 and 1. Examples of such experiments are found in biology, pharmaceuticals, etc. Becker (1968) proposed three additive and homogeneous mixture models of degree one, involving synergism. (see also Cornell and Gorman, 1978). These models are found to give good fit to many experimental data. See, for example, Snee (1973), Becker (1978), Sovolev and Chemleva (1976). In this talk, we consider a three-component mixture model with synergism between some or all components, and attempt to find optimum designs for estimation of the model parameters, using D- and A- optimality criteria. Kiefer’s ’Equivalence Theorem’ has been exploited in the characterization of optimal designs.

**The effect of noise on the Fitzhugh-Nagumo neuronal model** [Poster]
Charles Eugene Smith

*Statistics*

*North Carolina State Univ.*

Abstract. A nonlinear system of differential equations known as the Fitzhugh-Nagumo (FN) is used to describe the physiological state of a nerve membrane. Two different kinds of noise are added to the FN model to investigate the effect of noise on the membrane. They are Gaussian white noise and Poisson noise. Gaussian white noise represents many small synaptic inputs and Poisson noise represents a few large synaptic inputs. The non-oscillatory region before and after the bifurcation region is used to distinguish between Wiener vs. Poisson inputs by a hypothesis test about the mean number of level crossings. The null hypothesis is the expected level crossings of the equilibrium state by a time sampled linearized FN set of differential equations with Wiener input. The test performs well in rejecting non Wiener inputs in simulation studies, both in the linearized and nonlinear F-N model.

**S distributions** [C.7.6]
Ahmad Reza Soltani

*Statistics and Operations Research*

*Kuwait University*

*Smith* 65  *Soltani*
Abstract. We introduce and study a new class of Cauchy type parameterized distribution functions, called S-distributions, that assume same tail behavior as the Cauchy distribution on large values, but provide variety of patterns on small and mid size values. We introduce a new conjugate property as in the line of the Rossberg conjugate property for distributions functions, in terms of the Laplace transform, and show that S-distributions satisfy our conjugate property. We also present certain characterizations involving S-distributions and the generalized Linnik distributions. Data analysis is also performed on certain real data sets using S-distributions, where effectiveness of S-distributions in modeling phenomena with high variations is brought into light.

Fixed-width confidence interval based on a minimum Hellinger distance estimator [I.3.5]
T. N. Sriram

Abstract. In the context of discrete data, a sequential fixed-width confidence interval for an unknown parameter is constructed using the minimum Hellinger distance estimator (MHD) as the center of the interval. It is shown that our sequential MHD estimator is asymptotically consistent and efficient, when the assumed parametric model is correct. This shows that the MHD is a viable alternative to the maximum likelihood estimator (MLE), even when the sample size is random. Monte Carlo simulations show that the performance of our sequential procedure based on MHD, measured in terms of average sample size and the coverage probability, are as good as those based on MLE, when the assumed Poisson model is correct. However, when the samples come from a gross-error contaminated Poisson model, our numerical results show that the performance of the MLE is severely affected, while the procedure based on MHD continues to perform well. This shows that the MHD is robust against gross-error contamination, even for random sample sizes.

Towards statistical modeling of shapes of elastic functions, curves and surfaces [I.5.2]
Anuj Srivastava

Abstract. Shape analysis of 2D and 3D objects, both static and dynamic, is important in many applications, ranging from graphics, bioinformatics, and military to medical. Restricting to the boundaries of these objects, we are interested in tools for analyzing shapes of curves and surfaces. In particular we seek probability models that can capture typical variability within shape classes of interest, and can be used as priors in Bayesian shape estimation, clustering, classification, and tracking. Although parameterized curves and surfaces are efficient in representing objects boundaries, compared to say level sets or embedding in larger spaces, there is an important need to be invariant to parameterizations. This is often accomplished using representations and metrics such that the actions of the re-
parameterization groups are by isometries. For functions and curves, I will demonstrate an extension of the square-root representations, used historically for studying probability density functions under the Fisher-Rao Riemannian metric, that facilitates a desired elastic Riemannian framework. The resulting spherical manifolds are convenient for computing geodesics, distances, sample means and covariances. For surfaces, I will describe some preliminary work in parameterization-invariant shape analysis. I will use examples from computer vision, protein structure analysis, face recognition, and medical diagnosis to illustrate these ideas.

**Testing the structure of the covariance matrix with fewer observations than the dimension** [I.5.4]

Muni S. Srivastava

*Statistics*

*University of Toronto*

Abstract. We consider two hypothesis testing problems with N independent observations on a single m-vector, when m \( \leq N \), and the N observations on the random m-vector are independently and identically distributed as multivariate normal with mean vector \( \mathbf{b} \) and covariance matrix \( \mathbf{C} \), both unknown. In the first problem, the m-vector is partitioned into two subvectors of dimensions \( m_1 \) and \( m_2 \), respectively, and we propose a test for the independence of the two subvectors. The asymptotic distribution of the test statistic under the hypothesis is shown to be standard normal, and the power examined by simulations. The proposed test performs better than the likelihood ratio test even when \( m \) is much smaller than \( N \). The second problem is that of testing that the covariance matrix \( \mathbf{C} \) is of the intraclass correlation structure. The attained significance level and simulated power are obtained, and compare favourably with the likelihood ratio test.

**Modeling and forecasting malaria and dengue hemorrhagic fever incidence and prevalence in northern Thailand** [C.1.1]

Wattanavadee Sriwattanapongse

*Department of Statistics*

*Chiangmai University*

Abstract. Malaria and dengue hemorrhagic fever (DHF) are infectious diseases prevalent in many tropical countries, including Thailand. Thailand is located geographically in a tropical zone and the transmission of malaria and DHF is common, particularly in the upper northern region of the country. The objective of this study is to identify the patterns of hospital-diagnosed Malaria and DHF incidences using the previous monthly or quarterly periods in the upper northern region of Thailand. We use additive plus multiplicative regression models to describe these patterns. The models can be used to forecast malaria and DHF incidences, thus predicting where epidemics are likely to occur. This information can be used to prevent disease outbreaks occurring. Graphical displays showing district and period effects are presented. The results of this study show that historical malaria and DHF incidence rates can be used to provide a useful model for forecasting future epidemics. The
graphical display shows the improvement of risk prediction brought about by model. The model, even if based purely on statistical data analysis, can provide a useful basis for allocation of resources for disease prevention.

**Identifiability of large phylogenetic mixture models** [I.6.1]

Seth Sullivant

*Mathematics\nNorth Carolina State University*

Abstract. Phylogenetic models are hidden variable graphical models used to infer the ancestral relationships between a given collection of taxa. Phylogenetic mixture models are statistical models of character evolution allowing for heterogeneity. Each of the classes in some unknown partition of the characters may evolve by different processes, or even along different trees. The fundamental question of whether parameters of such a model are identifiable is difficult to address, due to the complexity of the parameterization. We analyze mixture models on large trees, with many mixture components, showing that both numerical and tree parameters are indeed identifiable in these models when all trees are the same. We also explore the extent to which our algebraic techniques can be employed to extend the result to mixtures on different trees. This is joint work with John Rhodes (University of Alaska–Fairbanks).

**Joint modeling of multiscale longitudinal and time to event data: an application to prospective pregnancy studies** [I.7.4]

Rajeshwari Sundaram

*Biostatistics and Bioinformatics Branch\nNICHD, NIH*

Abstract. Prospective pregnancy studies usually collect data on multiple time scales, namely on daily as well as cycle (in units of months) level, with the time to event of interest also measured on cycle level. Some longitudinal processes of interest in such studies are the intercourse behavior, stress levels of subjects, measured on daily level within a menstrual cycle, as well as the length of menstrual cycle which is subject to inter, as well as intra-subject variability. The time to event of interest in such studies is time to pregnancy, which is measured in number of menstrual cycles needed to conceive. We propose joint models using a shared parameter approach that allow us to study various associations of interest between these longitudinal processes and time to event. The proposed methods are investigated through extensive simulation studies and are illustrated on two prospective pregnancy studies (New York State Angler Prospective Pregnancy Cohort Study and Oxford Conception Study). We also show how the joint models can be used to predict subject-specific time to events allowing us to assess the longitudinal process as a marker of the time to event. We show how our method can be used to obtain some (new) biological findings of interest.

**Using ascertainment for targeted resequencing to increase power to identify causal variants** [I.5.1]

Sullivant

Swartz
Michael Swartz

Division of Biostatistics
University of Texas School of Public Health

Abstract. Researchers continue to use genome-wide association studies (GWAS) to find the genetic markers associated with disease. Recent studies have added to the typical two-stage analysis a third stage that uses targeted resequencing on a randomly selected subset of the cases to detect the causal single-nucleotide polymorphism (SNP). We propose a design for targeted resequencing that increases the power to detect the causal variant. The design features an ascertainment scheme wherein only those cases with the presence of a risk allele are selected for targeted resequencing. We simulated a disease with a single causal SNP to evaluate our method versus a targeted resequencing design using randomly selected individuals. The simulation studies showed that ascertaining individuals for the targeted resequencing can substantially increase the power to detect a causal SNP, without increasing the false-positive rate.

Duckworth-Lewis and twenty20 cricket [I.6.3]

Tim Swartz

Statistics and Actuarial Science
Simon Fraser University

Abstract. Originally designed for one-day cricket, this paper considers the use of the Duckworth-Lewis method as an approach to resetting targets in interrupted Twenty20 cricket matches. The Duckworth-Lewis table is reviewed and an alternative resource table is presented. The alternative table is constructed using observed scoring rates from international Twenty20 matches. A desideratum of a resource table is monotonicity in both the rows and columns corresponding to wickets and overs respectively. Consequently, a Gibbs sampling scheme related to isotonic regression is applied to the observed scoring rates to provide a nonparametric resource table. Taking into account the more aggressive batting style of Twenty20 compared to one-day cricket, the resultant resource table is seen to possess sensible features. A discussion is provided concerning the use of the Duckworth-Lewis method applied to Twenty20.

On Pearson’s correlation [I.5.4]

Gabor J Szekely

DMS
NSF

Abstract. Properties of the following version of Pearson’s correlation will be discussed. Take all pairwise distances between sample elements from a random variable X, and do the same for the sample elements from a random variable Y. Center the resulting distance matrices so that each has column and row means equal to 0. Finally, average the entries of the matrix which holds componentwise products of the two centered distance matrices. This is the definition of the sample distance covariance of X and Y. The population version equals zero iff X and Y are independent, whatever be the underlying distributions and whatever be the
dimension of X and Y.

**Mixture gatekeeping procedures with clinical trial applications** [I.4.3]

Ajit C. Tamhane

*Department of Industrial Engineering and Management Sciences*
*Northwestern University*

Abstract. Gatekeeping procedures address the problems of testing hierarchically ordered and logically related null hypotheses that arise in clinical trials involving multiple endpoints, multiple doses, noninferiority-superiority tests, subgroup analyses etc. In this talk we will review previous work and then introduce a very general and powerful method for constructing gatekeeping procedures based on the closure principle, called the mixture method. This method is capable of handling arbitrary logical restrictions among the hypotheses. A clinical trial example will be presented to illustrate the method.

**Penalized empirical likelihood and growing dimensional general estimating equations** [I.3.1]

Cheng Yong Tang

*Statistics and Applied Probability*
*National University of Singapore*

Abstract. When a parametric likelihood function is not specified for a model, estimating equations (EEs) provide an instrument for statistical inference. Qin and Lawless (1994) illustrated that empirical likelihood (EL) makes optimal use of the EEs in inferences for fixed (low) dimensional unknown parameters. In this paper, we study EL for general EEs with growing (high) dimensionality and propose an EE-based penalized empirical likelihood (PEL) approach for parameter estimation and variable selection. Theoretically, we quantify the asymptotic properties of EL and PEL. We further show that the PEL has the Oracle property. Namely, with probability tending to one, PEL identifies the true sparse model. In addition, the efficiency of the estimated nonzero coefficients is optimal. The performance of the proposed PEL is illustrated via four simulated applications and a data analysis.

**Pairwise maximum likelihood procedure for estimating shift in two-sample location problem** [C.7.7]

Feridun Tasdan

*Mathematics*
*Western Illinois University*

Abstract. This study is about maximum likelihood estimation of shift parameter using pairwise differences in two-sample location problem. The proposed procedure uses pairwise differences of data to find a distribution function of the differences. Hence, using the distribution function of the pairwise differences, one can find a likelihood function with respect to the shift parameter. By using the likelihood function, a MLE of shift parameter will be found. Moreover, it will be shown that the proposed procedure is equivalent to the regular
likelihood ratio test. An asymptotic level hypothesis test and confidence interval will be investigated for the proposed estimator. The study ends with a bootstrap simulation study which is performed to show the efficiency results of the proposed estimator.

**Toxicity profiling of engineered nanomaterials** [I.4.2]
Donatello Telesca

*Biostatistics*
*UCLA*

Abstract. In-vitro high throughput screening (HTS) assays for the assessment of engineered nanomaterials provide an opportunity to learn how these particles interact at the cellular level, particularly in relation to injury pathways. These types of assays are often characterized by small sample sizes, high measurement error and high dimensionality as multiple cytotoxicity outcomes are measured across an array of doses and durations of exposure. We propose a probability model for toxicity profiling of engineered nanomaterials. A hierarchical framework is used to account for the multivariate nature of the data by modeling dependence between outcomes and thereby combining information across cytotoxicity pathways. In this framework we are able to provide a flexible surface-response model that provides inference on various classical risk assessment parameters. We discuss applications of this model to data on eight nano particles evaluated in relation to four cytotoxicity parameters.

**Likelihood ratio test based methods for signal detection in longitudinal drug safety databases** [I.2.3]
Ram C. Tiwari

*Office of Biostatistics*
*Food and Drug Administration*

Abstract. Purpose of this talk is to extend a newly developed signal detection method, Likelihood Ratio Test (LRT), for signal generation in spontaneous adverse event reporting databases such as the FDAs Adverse Events Reporting System (AERS) database, to the sequential LRT; first, for the analysis of AERS data over time, and second, for the analysis of exposure-based safety data. We evaluate the performance of the tests using simulated data and show that they control type-I error, even in the situations where the interest may be to look at several drugs (or several adverse events (AEs)) such as a drug class (or biologically/medically related AEs) simultaneously. These methods are applied to a longitudinal dataset from the AERS and to an exposure-based dataset.

**Generalised Gumble distributions and minification processes** [C.7.7]
Sister Alice Thomas

*Statistics*
*caltcut*

density and hazard rate functions are derived and their shape properties are considered. Expressions for the nth moments are given. The unknown parameters of the distribution are estimated by using the maximum likelihood estimation and some numerical results of the estimation are given. Autoregressive time series models of order 1 as well as k are developed with minification structure, having these stationary marginal distributions. Various characterizations are obtained.

**A functional central limit theorem for empirical processes under a strong mixing condition [C.7.6]**

Cristina Tone

*Mathematics*  
*Louisville*

Abstract. We introduce a functional central limit theorem for empirical processes endowed with real values from a strictly stationary random field that satisfies an interlaced mixing condition. We proceed by first obtaining the limit theorem for the case where the random variables of the strictly stationary random field are uniformly distributed on the interval [0, 1]. We then generalize the result to the case where the absolutely continuous marginal distribution function is not longer uniform. In this case we show that the empirical process endowed with values from the stationary random field, due to the interlaced strong mixing condition, doesn’t converge in distribution to a Brownian bridge, but to a continuous Gaussian process with mean zero and the covariance given by the limit of the covariance of the empirical process.

**Some nonparametric precedence-type tests based on progressively censored samples and evaluation of power [I.3.3]**

Ram C. Tripathi

*Management Science and Statistics*  
*University of Texas at San Antonio*

Abstract. Comparison of two populations is quite common in reliability and survival studies. For example, while comparing treatment versus control, one may be interested in assessing whether the observations corresponding to the treatment population have longer life than those from the control population. In such studies, it may be desired to make a reliable decision early in the experiment based on the first few failures, that results in savings of some units which would have been destroyed otherwise. In this paper, we consider progressively censored independent random samples from two populations with cdfs F(x) and G(y). We derive the joint distribution of first m placement statistics for the sample from the population with cdf F(x). We consider the following two statistics: (i) Wilcoxon rank-sum precedence-type, and (ii) a statistic based on Kaplan-Meier estimator of the cumulative distribution function. The exact null distributions of these statistics are derived. These distributions are used to generate critical values and the corresponding level of significance for some combinations of sample sizes and censoring schemes. We also present an outline for deriving their non-null distributions under Lehmann alternatives. A power study of the proposed tests is
carried out under Lehmann alternatives as well as under location-shift alternatives through Monte Carlo simulation. Examples are presented to illustrate computations of the proposed test. Finally, concluding remarks and recommendations are made regarding their choice.

**Targeted maximum likelihood based super learning: assessing effects in RCT and observational studies** [I.1.1]

Mark van der Laan

*Biostatistics*

*UC Berkeley*

Abstract. Current statistical practice to assess an effect of an intervention or exposure on an outcome of interest often involves either maximum likelihood estimation for a priori specified regression model, or, manual and/or data adaptive interventions to fine tune a choice of model. In both cases, the point estimates and the estimate of the signal to noise ratio will be potentially severely biased, causing an epidemic of false claims based on data analyses.

In this talk we present targeted maximum likelihood based estimators of a causal effect defined in nonparametric models, that take away the need for specifying regression models. Two fundamental concepts underlying this methodology are Super Learning, the very aggressive use of cross-validation to select optimal combinations of many model fits, and subsequent targeted maximum likelihood estimation to target the fit towards the causal effect of interest.

We illustrate this method in observational studies for assessing the effect of an intervention on adherence to drug regimen in HIV infected patients, and for discovery of mutations in the HIV virus that cause resistance to a particular drug regimen. We also illustrate the performance on FDA approved clinical trials, simulated data imitating postmarket safety analysis, and the analysis of single nucleotide polymorphisms.

**Large deviations** [Plenary III]

Srinivasa Varadhan

*Mathematics*

*New York University*

Abstract. We will explore the large deviation properties of Random graphs where between each pair of vertices an edge is present with probability p and absent with probability 1-p. This happens independently for each possible edge. We are interested in studying the large deviation properties of of the number of triangles or similar count of any finite graph in the random graph, calculation of the rate as well as determination the most likely source of the large deviation.

**A method of construction a multiple increment-decrement life table of HIV population** [C.7.5]

Manoj Kumar Varshney

*Department of Statistics*

*University of Delhi, India*
Abstract. This paper deals with the size of the HIV population over the time that starts with a cohort of fresh HIV and subsequently altered because of fresh HIV births, HIV deaths and transmission from HIV to AIDS; While the birth is an incremental force, death and transmission to AIDS is a decremental force in nature. Our objective is to estimate the size of HIV population over the time taking into consideration of all the incremental and decremental factors. For the estimation of HIV births and deaths use of Kendall’s (1948) generalized birth and death process has been taken as a model with the birth and death intensities following Weibull distribution. For the estimation of births and deaths, data have been simulated on the basis of information of US official statistics (UNAIDS/WHO-2002 and UNAIDS-2006). The probabilities for transmission from HIV to AIDS have been extracted from the results of Munoz and Xu (1996) by further smoothing the rates for the intermediate years in the pattern of log-normal distribution. Finally a life table has been constructed to exhibit the expectation of life in the state of HIV from 0-20 years.

Model selection and estimation in generalized additive models [Poster]
Dong Wang

Department of Statistics
North Carolina State University

Abstract. Generalized additive models, proposed by Hastie and Tibshirani (1990), allow the distribution of the responses to be any member from the family of exponential distributions. Compared to linear models, generalized additive models assume an additive structure of non-parametric smooth functions of covariates of interest. The above two properties make this model family more proper and general in real life. Thus, the model selection and estimation is of interest. We use smoothing splines to estimate the nonparametric functions. Taking advantage of the linear mixed model representation of a smoothing spline, we propose a unified EM algorithm to perform the model selection and estimation simultaneously. In this process, we treat the smoothing parameters as extra variance parameters of mixed effects. In addition, to make the algorithm more efficient, we propose a way to reduce the dimension of large matrix while keeping most information in the data via the eigenvalue-eigenvector decomposition approach.

Hierarchical Bayesian analysis of directional data using the projected normal distribution [Poster]
Fangpo Wang

Department of Statistical Science
Duke University

Abstract. The projected normal distribution is an under-utilized model for explaining directional data. It provides flexibility, e.g., bimodality, asymmetry, and convenient regression specification. We develop fully Bayesian hierarchical models for circular data. We show how they can be fit using suitable latent variables and MCMC methods. We show how posterior estimation of analytical quantities such as mean direction and concentration can be implemented as well as a regression setting. Simulated and real data examples are provided for
Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases [I.5.1]

Jian Wang

Epidemiology
MD Anderson Cancer Center

Abstract. Genetic association studies for binary diseases are designed as case-control studies: the cases are those affected with the primary disease and the controls are free of the disease. At the time of case-control collection, information about secondary phenotypes is also collected. Association studies of secondary phenotype and genetic variants have received a great deal of interest recently. To study the secondary phenotypes, investigators use standard regression approaches, where individuals with secondary phenotypes are coded as cases and those without secondary phenotypes are coded as controls. However, using the secondary phenotype as an outcome variable in a case-control study might lead to a biased estimate of odds ratios (ORs) for genetic variants. The secondary phenotype is associated with the primary disease; therefore, individuals with and without the secondary phenotype are not sampled following the principles of a case-control study. In this article, we demonstrate that such analyses will lead to a biased estimate of OR and propose new approaches to provide more accurate OR estimates of genetic variants associated with the secondary phenotype for both unmatched and frequency-matched (with respect to the secondary phenotype) case-control studies. We also propose a bootstrapping method to estimate the empirical confidence intervals for the corrected ORs. Using simulation studies and analysis of lung cancer data for single-nucleotide polymorphism associated with smoking quantity, we compared our new approaches to standard logistic regression and to an extended version of the inverse-probability-of-sampling-weighted regression. The proposed approaches provide more accurate estimation of the true OR.

Shape restricted regression with multivariate Bernstein polynomials [Poster]

Jiangdian Wang

Statistics
North Carolina State University

Abstract. The estimation of shape-restricted regression curves is challenging for multivariate predictors, especially for functions with compact support. Most of the currently available statistical estimation methods for shape restricted regression functions are generally computationally very intensive. Some of the existing methods are applicable to only nonparametric additive models ignoring the interaction terms. This article considers a suitable class of multivariate polynomials with unknown order and proposes a sieved estimator obtained from a nested sequence of shape-restricted multivariate Bernstein polynomials. Our proposed method is shown to be computationally attractive and universally consistent under some mild regularity conditions. Our method is also flexible in the sense that it can be easily
adapted to accommodate many popular multivariate shape restrictions, such as nonnegativity, isotonicity, convexity and concavity. The proposed estimator is evaluated by Monte Carlo simulation studies.

**A multivariate spatial zero-inflated poisson model for fish species richness estimation in Gulf of Maine [I.5.5]**

Xia Wang

*National Institute of Statistical Sciences*
*National Institute of Statistical Sciences*

Abstract. Gulf of Maine is one of the world most varied and productive marine ecosystem. Its complex geology and oceanography provide excellent and unique hosts for many ocean creatures. An important aspect for the marine management is to study the pattern of fish distribution and to select the marine protected areas based on different biological criteria. Biological diversity is one of those important factors to consider. Kuo et al. (2010) studies the fish species richness in this area using NEFSC data. A uniform grid network is drawn over the area of interest at different spatial scales. The number of species caught in each grid cell is used to estimate fish species richness. This approach, however, does not provide any inference on those grids not surveyed. Also, it tends to underestimate the fish species richness. As zero caught may not be a true indicator for the absence of the fish from a given grid. The false zero may be caused by incomplete coverage and imperfect detection. Also, we expect that correlation exists among species and modeling co-presence of multiple species would allow stronger information borrowing. In this paper, we use a multivariate spatial zero-inflated Poisson model to draw statistical inference on the presence or absence of a specific fish species in the Gulf of Maine. Particular, we study how the ocean geographical conditional are related to the presence, absence and abundance of the fish. Also, with the long time series of the survey data (1963-2008), we discuss how the the fish species evolves along the time.

**Sequences of regressions and their independences [I.6.1]**

Nanny Wermuth

*Mathematical Statistics*
*Chalmers/Gothenburg University*

Abstract. Univariate and multivariate regressions for categorical or continuous responses are well understood and widely used. It is less well known that recursive sequences of such regressions can form an important subclass of graphical Markov models, one type of so-called chain graphs. We give a number of reasons for the importance of this model class and also point to related open problems.

**Level set estimation from projection measurements [I.4.4]**

Rebecca Willett

*Electrical and Computer Engineering*
*Duke University*

Wang 76 Willett
Abstract. In this talk, I will describe a new methodology and associated theoretical analysis for rapid and accurate extraction of level sets from projection measurements. The identification of the boundaries of such sets is an important theoretical problem with applications in digital elevation maps, medical imaging, and astronomy. However, in many settings we do not directly observe the function of interest, but rather indirect measurements such as tomographic projections, coded aperture measurements, or pseudo-random projections associated with compressed sensing. Estimating a level set of the function underlying such data can be accomplished without an intermediate function reconstruction step, leading to significantly faster estimation methods. In addition, a combination of coherence measures and concentration inequalities can be used to characterize the performance of the proposed method as a function of the measurement model.

An early step in assessing an estimate’s accuracy in the presence of competing estimates [I.1.5]

Tommy Wright

Center for Statistical Research & Methodology
US Census Bureau

Abstract. Frequently, a statistical agency finds it useful to compare a particular estimate for a given parameter the agency has produced with other independently produced estimates which seek to estimate the same or similarly defined parameter. Potential advantages of this practice are well documented with a history [e.g., Eckler and Pritzker (1951); Morgenstern (1963); Martin (1981)], but there are limitations. Tsao and Wright (1983) proposed a simple tool, called the maximum ratio, which serves as a measure of closeness among comparable estimates and as the basis of a test for determining a minimum distance that at least one of the estimates is from the unknown true value of the parameter. In this talk, we consider a possible role for this simple tool in assessing the goodness of decennial census counts by comparing them with estimates from demographic analysis as well as estimates from a national sample survey. We give other examples. We illustrate geometrically why the test works and consider theory and development for extension to vector-valued parameters.

Inverse-probability weighted estimating equations for analyzing secondary phenotypes in case-control genetic association studies [C.7.5]

Chuanhua (Julia) Xing

Biostatistics and Bioinformatics
Duke University

Abstract. The case-control study is an approach to assess association between genetic exposures and disease by measuring statistical differences between cases (diseased) and controls (disease-free). There is considerable interest in measuring other phenotypes than the primary one. Studying multiple traits from the same set of subjects can also maximally gain the association mapping between genotypes and phenotypes, in addition to the economical consideration. Many previous statistical methods may be biased in failing to properly consider the different sampling rates in selecting cases and controls. We propose a general
approach for estimating and testing the population effect of a genetic variant on a secondary phenotype. Our approach is based on inversely-weighting estimating equations within a generalized method of moments (GMM) framework using a conditional probability of an individual being a case or a control as the corresponding weight. Our method is substantially more robust to model misspecification, and out-performs a likelihood-based analysis, both in terms of validity and power.

A Monte Carlo test of linkage disequilibrium for single nucleotide polymorphisms [I.5.1]
Hongyan Xu

Biostatistics
Georgia Health Sciences University

Abstract. Genetic association studies, especially genome-wide studies, make use of linkage disequilibrium (LD) information between single nucleotide polymorphisms (SNPs). LD is also used for studying genome structure and has been valuable for evolutionary studies. The strength of LD is commonly measured by r-square, a statistic closely related to the Pearson’s chi-square statistic. However, the computation and testing of linkage disequilibrium using r-square requires known haplotype counts of the SNP pair, which can be a problem for most population-based studies where the haplotype phase is unknown. Most statistical genetic packages use likelihood-based methods to infer haplotypes. However, the variability of haplotype estimation needs to be accounted for in the test for linkage disequilibrium. We develop a Monte Carlo based test for LD based on the null distribution of the r-square statistic. Simulations show that our test has similar power as a previous test based on a composite statistic of linkage disequilibrium. Unlike the test based on the composite statistic, our test is based on r-square and can be reported together with r-square. Simulation studies show that it offers better power than existing methods. The new test provides an alternative test for LD and has been implemented as a R program for ease of use. It also provides a general framework to account for other haplotype inference methods in LD testing.

Statistical methods for analyzing tissue microarray images – algorithmic scoring and co-training [I.7.3]
Donghui Yan

Biostatistics and Biomathematics
Fred Hutchinson Cancer Research Center

Abstract. Recent advances in tissue microarray technology have allowed immunohistochemistry to become a powerful medium-to-high throughput analysis tool, particularly for the validation of diagnostic and prognostic biomarkers. However, as study size grows, the manual evaluation of these assays becomes a prohibitive limitation; it vastly reduces throughput and greatly increases variability and expense. We propose an algorithm – Tissue Array Co-Occurrence Matrix Analysis (TACOMA) – for quantifying cellular phenotypes based on textural regularity summarized by local inter-pixel relationships. The algorithm can be easily trained for any staining pattern, is absent of sensitive tuning parameters and has the
ability to report salient pixels in an image that contribute to its score. Pathologists’ input via informative training patches is an important aspect of the algorithm that allows the training for any specific marker or cell type. With co-training, TACOMA can be trained with a radically small training sample (e.g., with size 30). We give theoretical insights into the success of co-training via thinning of the feature set in a high dimensional setting when there is “sufficient” redundancy among the features. TACOMA is flexible, transparent and provides a scoring process that can be evaluated with clarity and confidence. In a study based on an estrogen receptor (ER) marker, we show that TACOMA is comparable to, or outperforms, pathologists’ performance in terms of accuracy and repeatability.

**Two dimensional stochastic singular control via Dynkin game and Dirichlet form** [I.6.7]

Yipeng Yang

*Dept. of Mathematics*

*University of Missouri-Columbia*

Abstract. A two dimensional stochastic singular control problem is considered. Through Dynkin game and Dirichlet form, we showed that under the optimal control policy, the controlled process is a reflected diffusion within certain boundaries, hence a Skorokhod problem. The existence, uniqueness and conditions of the optimal solution are discussed.

**Bayesian modeling for collaborative filtering at Yahoo!** [I.4.5]

Liang Zhang

*Yahoo! Labs*

*Yahoo! Inc.*

Abstract. We consider the problem of algorithmically recommending items to users who visit Yahoo!. We first introduce a generalized matrix factorization model (GMF) that provides state-of-the-art performance for users and items with large enough past data, and also good performance for new users and new items through a flexible regression model prior on their features. The flexibility of this model allows one to leverage a large body of work on regression modeling, variable selection and model interpretation. In the second part of this talk we introduce a novel multi-level hierarchical model that we refer to as User Profile Model with Graphical Lasso (UPG). This model includes a regression model upon user and item features, and a user-item affinity random effect vector that is assumed to be drawn from a prior with a precision matrix that measures residual partial associations among items. To ensure better estimates of precision matrix in high-dimensions, the matrix elements are constrained through a Lasso penalty. This model is fitted through a penalized-quasi likelihood procedure coupled with a scalable EM algorithm, and to incorporate the Lasso penalty, we take recourse to the Graphical Lasso algorithm for covariance selection in the M-step. Both GMF and UPG have been tested and compared with various kinds of baseline models on both the benchmark data set MovieLens and several Yahoo! data sets.

**Testing the structural stability of temporally dependent functional**
observations and applications to climate data [Poster]
Xianyang Zhang

Statistics
University of Illinois at Urbana-Champaign

Abstract. In this paper, we develop a self-normalization (SN) based test to test the structural stability of temporally dependent functional observations. Testing for a change point in the mean of functional data has been studied in Berkes, Gabrys, Horvath and Kokoszka (2009, JRSSB), but their test was developed under independence assumption. In many applications, functional observations are expected to be dependent, especially when the data is collected over time. Building on the SN-based change point test proposed in Shao and Zhang (2010, JASA) for a univariate time series, we extend the SN-based test to the functional setup by testing the constant mean of the finite dimensional eigenvectors after performing functional principle component analysis. Some asymptotic theory are derived under both the null and local alternatives. Through theory and extensive simulations, our SN-based test statistic proposed in the functional setting is shown to inherit some nice properties in the univariate setup, e.g., the test is asymptotically distribution free and its power is monotonic. Furthermore, we extend the SN-based test to test for a change point in the dependence structure of functional observations. Illustrations using two climate data sets are also presented.

High order Laplace approximation for statistical inference of dependent data [I.5.6]
Zhengyuan Zhu

Department of Statistics
Iowa State University

Abstract. Laplace approximation has been used in many statistical applications to approximate integrals. In this talk we present two applications which use high order Laplace approximation to derive theoretical results and develop efficient algorithms respectively. The first application is the asymptotic comparison of predictive densities for dependent observations. We study Bayesian predictive densities based on different priors and compare them with frequentist plug-in type predictive densities when the predicted variables are dependent on the observations. Average Kullback-Leibler divergence to the true predictive density is used to measure the performance of different inference procedures. The notion of second order KL REML dominance is introduced, and an explicit condition for a prior to be second-order KL REML dominant is given using an asymptotic expansion. The second application is approximate inference for spatial generalized linear mixed models (SGLMM). Estimation and prediction in SGLMM are often hampered by intractable high dimensional integrals. We provide a framework to solve this intractability using asymptotic expansions when the number of random effects is large, and propose algorithms for the estimation and prediction of SGLMM based on modified Laplace approximation which are accurate and computationally fast. The algorithms are evaluated through simulation and applied to a rhizoctonia root rot data.