The ABSTRACTS

IISA Conference May 22-25, 2008

Department of Statistics
The University of Connecticut
Storrs

Abstracts Are Arranged According to
Last Names of Presenting Authors

Presenting Authors Are Identified with Asterisks (*)
In the Case of Multiple Authorship

Per Scheduled Program, the Time-Slots Are:

FriPMTimeSlot-1: Friday 1:15PM – 2:45PM
FriPMTimeSlot-2: Friday 3:15PM – 4:45PM
SatAMTimeSlot-1: Saturday 9:00AM – 10:30AM
SatPMTimeSlot-1: Saturday 1:30PM – 3:00PM
SatPMTimeSlot-2: Saturday 6:00PM – 7:30PM
SunAMTimeSlot-1: Sunday 8:00AM – 9:30AM
SunAMTimeSlot-2: Sunday 10:00AM – 11:30AM
SunAMTimeSlot-3: Sunday 11:30AM – 1:30PM

FriPM-TimeSlot-2: Session #40

Estimations of Population Parameters from Censored and Truncated Samples

AbouEl-Makarim A. Aboueissa
University of Southern Maine-Portland, USA.
(E-mail: aaboueissa@usm.maine.edu)

Abstract

In practice, environmental studies such as water quality, air quality and soil contamination studies, sample observations are often restricted in some range of possible population values. Scientists identify and measure concentrations of environmental pollutants, and establish safe risk levels. When concentrations of environmental pollutants fall below risk levels, no action is taken. However, when concentrations fall above these risk levels samples are collected. These risk levels are called truncated limits (TL) and data collected above these levels are said to be left-truncated. Data set for which all observations may be identified and counted, with some observations falling into the restricted interval of measurements and the remaining observations being fully measured, is said to be censored. This study is concerned with the problem of estimating parameters from truncated data, and from both censored and truncated data. The method of maximum likelihood is employed to develop estimators of parameters under consideration. Easily computed estimates of parameters obtained from normally or log-normally distributed samples which are truncated or censored and truncated are developed. A new computer algorithm for obtaining the maximum likelihood estimates of μ and σ from singly-truncated samples, and not requiring auxiliary tables, is provided.

SunAM-TimeSlot-3: Session #43

Random Design Space for Lung's Retention Model

[1*]M. Amo-Salas, [1]J. López-Fidalgo and [2]J. M. Rodríguez-Diaz
[1]University of Castilla-La Mancha, Spain; [2]University of Salamanca, Spain
([*]Paper Presenter; E-mail: mariano.amo@uclm.es)

Abstract

A model of lung retention of radioactive particles is considered in this work. When a leak of radiation happens in facilities with workers the accident is detected only at the end of the workers' shift when the filters are checked. Thus, the actual time when the leak happens in not known and therefore the design space of times to perform bioassays on the workers should be considered as a random set. Optimal designs for different possible times of the accident are computed and they are compared with the worst case, when the accident happens at the beginning of the shift. Moreover, different distributions of probability are assumed as distribution of the moment of the accident and designs computed for those situations.

Another research line is based on the study of the covariance matrix. The observations taken on the same worker are correlated. We can consider the covariance parameter as a parameter of interest (to be estimated). The "Virtual Noise" method is applied to compute the optimal designs for this model. Different covariance structures are used in this investigation.

SatPM-TimeSlot-1: Session #1

Intrinsic Dimensionality Estimation of High Dimension, Low Sample Size Data with Geometric Representation

Makoto Aoshima
University of Tsukuba-Ibaraki, Japan
(E-mail: aoshima@math.tsukuba.ac.jp)

Abstract

Datasets with more variables than observations are emerging in various areas of science, such as genetic microarrays, medical imaging and text recognition. Such data have surprising and often counter-intuitive geometric structures.

In their asymptotic study regarding increasing dimensionality with a fixed sample size, Hall et al. (2005, *J. R. Statist. Soc. B*, 67, 427-444) and Ahn et al. (2007, *Biometrika*, 94, 760-766) showed that, under some conditions, the geometrical structure of data becomes deterministic: each data vector is approximately located on the vertices of a regular simplex in a high-dimensional space.

We note that the local neighborhoods are corrupted and dominated by the high-dimensional noise. The process of denoising has the goal to project the data points onto the submanifold and one naturally has to know the intrinsic dimension (ID) of the submanifold as a parameter of the algorithm. However, in the presence of high-dimensional noise it is quite difficult to estimate ID correctly. In this talk, we consider this problem in a context of adaptive sample size determination.

Sat 11:15AM – 12:00Noon: Special Named Lecture-V

P. R. Krishnaiah Lecture

Flexible Augmentations of Multivariate Models

Barry C. Arnold
University of California-Riverside, USA
(E-mail: barry.arnold@ucr.edu)

Abstract

Although the classical multivariate model is aesthetically enticing, it is not unusual to encounter data sets which do not seem to fit the Gaussian model. A variety of flexible augmentations of the classical model have received attention in the post-Krishnaiah era. A full survey will not be attempted. Instead we will concentrate on a few of the directions in which model augmentation has been proposed. Specifically we will discuss: conditional specification, hidden truncation, contour specification and a generalized Rosenblatt construction.

SunAM-TimeSlot-2: Session #18

Discriminating Between Level Shifts and Unit Roots

[1*]Alexander Aue, [2]Lajos Horvath, [3]Marie Huskova, and [4]Shiqing Ling
[1]University of California-Davis, USA; [2]University of Utah-Salt Lake City, USA;
[3]Charles University-Prague, Czech Republic;
[4]University of Science and Technology-Hong Kong, China
([*]Paper Presenter; E-mail: alexaue@wald.ucdavis.edu)

Abstract

In this talk, we present several two-step testing procedures to detect and identify structural breaks such as level shifts and unit roots. In a first step, we test whether or not a given data set exhibits a structural break, while the second step is used to determine its precise form. The testing procedures are based on functionals of the partial sums of the observations and have limit distributions under level shifts, but tend to infinity in the case of unit root-type behavior. The results thereby extend procedures currently available in the literature that are simultaneously sensitive to both alternatives but do not distinguish between them. The theoretical findings are underlined by a simulation study and an application to returns of the German stock index DAX.

FriPM-TimeSlot-2: Session #36

SUMSRI (Summer Undergraduate Mathematical Sciences Research Institute) Projects in Multivariate Statistics

[1*]Andrea Austin and [2*]Christina McIntosh
[1]St. Michael's College-Vermont, USA;

[2]Spelman College-Georgia, USA
([*]Paper Presenter; E-mail: aaustin2@smcvt.edu)

Abstract

In this Paper we describe two projects that we worked on at SUMSRI, 2007, held at Miami University, Oxford, Ohio. In both projects, we applied Principal Components, Factor Analysis, and Discriminant Analysis to analyze innovative data sets.

FriPM-TimeSlot-l: Session #27

Bayesian Methods for Copy Number Data

Veera Baladandayuthapani
University of Texas M.D. Anderson Cancer Center, USA
(E-mail: veera@mdanderson.org)

Abstract

Array-based comparative genomic hybridization (array-CGH) provides a high-throughput, high-resolution method to measure relative changes in DNA copy number simultaneously at thousands of genomic loci. 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 Bayesian regression based approach to modeling these genomic profiles in the presence of multiple samples/arrays. The Bayesian model borrows strength across all arrays to call gains and losses at the population level as well as accounting for subject-specific deviations.

We illustrate our methods on a osteosarcoma metastasis experiment conducted at MD Anderson Cancer Center. We adopt a Bayesian functional regression based approach to characterize these profiles.

SatPM-TimeSlot-l: Session #34

Modeling Large Spatial and Spatiotemporal Datasets: Advancing Methods and Applications in Forestry, Ecology and Environmental Sciences

[1*]Sudipto Banerjee and [2]Andrew O. Finley
[1]University of Minnesota-Minneapolis, USA;
[2]Michigan State University-East Lansing, USA.
([*]Paper Presenter; E-mail: sudiptob@biostat.umn.edu)

Abstract

With accessibility to geocoded locations where scientific data are collected through Geographical Information Systems (GIS), investigators are increasingly turning to spatial process models for carrying out statistical inference. Over the last decade hierarchical models have become especially popular for spatial modeling, given their enhanced modeling flexibility. However, fitting hierarchical spatial models often involves expensive matrix decompositions whose computational complexity increases dramatically with the number of spatial locations. This renders them infeasible for large spatial data sets. In this talk we propose to use a predictive process that projects the original spatial process that projects process to a lower-dimensional subspace thereby reducing the computational burden. We show how the predictive process seamlessly adapts to settings with nonstationary processes, with richer and more complex space-varying regression models and with multivariate spatial models. A computationally feasible template that encompasses these diverse settings will be presented.

SatPM-TimeSlot-2: Session #12

Bayesian Joint Modeling of Multivariate Longitudinal Data with Dropout

[1*]Sanjib Basu and [2]Pulak Ghosh
[1]Northern Illinois University-De Kalb, USA;
[2]Georgia State University-Atlanta, USA
([*]Paper Presenter; E-mail: basu@niu.edu)

Abstract

We propose a Bayesian model for non-ignorable dropout where the dropout process is modeled jointly with the longitudinal observation process. In our application, the longitudinal process is multivariate involving multiple endpoints whereas the dropout process also involves some complexity. We describe a semiparametric joint model and develop Bayesian model comparison to compare different dropout models.

SatPM-TimeSlot-l: Session #38

General Specification Testing with Locally Misspecified Models

[1*]Anil K. Bera, [2]Gabriel Montes-Rojas, and [3]Walter Sosa-Escudero
[1]University of Illinois- Urbana Champaign, USA; [2]City University-London, UK;
[3]Universidad de San Andr´es-Victoria, Argentina
([*]Paper Presenter; E-mail: abera@ad.uiuc.edu)

Abstract

A well known result is that many of the tests used in econometrics such as the Rao score test, may not be robust to misspecified alternatives, that is, when the alternative model does not correspond to the underlying data generating process. Under this scenario, these tests spuriously reject the null hypothesis too often. We generalize this result to GMM based tests. We also extend the method proposed in Bera and Yoon (1993, *Econometric Theory*, 9) for constructing RS tests that are robust to local misspecifications to Newey-West GMM based tests. Finally, a further generalization for general estimating and testing functions is developed. This framework encompasses both the Bera-Yoon likelihood based results as well as its use in the GMM environment. A simple application of this model shows that the robust tests that maximize power under local alternatives for a given size should be constructed using the Neyman C(alpha) test statistics.

SunAM-TimeSlot-l: Session #30

Bayesian Downscaling of Outputs from Numerical Models

[1*]Veronica J. Berrocal and [2]Alan E. Gelfand
[1]U.S. Environmental Protection Agency, USA; [2]Duke University-Durham, USA
([*]Paper Presenter; E-mail: vjb2@stat.duke.edu)

Abstract

In many environmental problems data arise from two sources: numerical models and monitoring networks. The first source provides predictions at the level of grid cells, while the second source gives measurements at points. Accommodating the spatial misalignment between the two types of data is of fundamental importance for both improved predictions of exposure as well as for evaluation and calibration of the numerical model. In this paper we propose a simple method to downscale the output from numerical models to point level. The model, specified within a Bayesian framework, regresses the observed data on the numerical model output using spatially varying coefficients, modeled as correlated spatial Gaussian processes. Moreover, the model can be easily extended to a dynamic implementation. As an example, we have applied our method to ozone concentration data for the Eastern US and we have evaluated its predictive performance, comparing it with that of the Bayesian melding approach of Fuentes and Raftery (2005, *Biometrics*, 61, 36-45).

SunAM-TimeSlot-3: Session #19

Geometrical Domain of Spin-1/2 Probability Mass Function

[1*]Karthik Bharath, [2]Swarnamala Sirsi, and [3]A. R Usha Devi
[1]SUNY Binghamton-Binghamton, USA; [2]University of Mysore-Mysore, India
[3]Bangalore University-Bangalore, India
([*]Paper Presenter; E-mail: kbharat1@binghamton.edu)

Abstract

The quantum analogue of the classical characteristic function for a spin-1/2 statistical assembly of particles is considered and the probability mass function of the random vector associated with the assembly is derived. It is seen that the positive regions of Wigner and Margenau-Hill quasi-distributions for the three components of spin, correspond to a trivariate probability mass function at an expectation value level. We identify the domain of these positive regions as an Octahedron inscribed in the Bloch sphere with its vertices on the surface of the sphere. It is in this domain that a quantum characteristic function characterizing the quasi-distribution, admits the derived probability mass function in IR3. It is also observed that the classical variates X_1, $X_{2,}$ X_3 corresponding to the 3 spin operators σ_1, $\sigma_{2,}$ σ_3 in the domain, are independent if and only if the Bloch vector lies on any one of the axes.

SatPM-TimeSlot-1: Session #41

Stirling's Formula and its Extensions: Heuristic Approaches

[1]*Debanjan Bhattacharjee and [1]Nitis Mukhopadhyay
[1]University of Connecticut-Storrs, USA
(*Paper Presenter; E-mail: debanjan.bhattacharjee@gmail.com)

Abstract

Walsh (1995, *The Amer. Statistician*, 49, pp. 270-271) introduced a heuristic approach to arrive at Stirling's formula for *n*! when *n* is large by manipulating an appropriate Poisson probability. We show how similar heuristics may be applied to obtain Stirling-type approximations for interesting binomial and negative binomial coefficients as well as *kn*!, where *n* (large) and *k* are positive integers.

Next, we extend the approach to find Stirling-type approximations for interesting multinomial coefficients. These *direct* approaches avoid repetitive use of Stirling's approximation itself. We also show how such heuristics may validate Stirling's formula for a Gamma function that is evaluated at *ns* where *n* is a large positive integer but *s* is positive and arbitrary.

FriPM-TimeSlot-1: Session #3

Fractional Levy-Ornstein-Uhlenbeck Stochastic Volatility Models

Jaya Bishwal
University of North Carolina-Charlotte, USA
(E-mail: J.Bishwal@uncc.edu)

Abstract

We account for two stylized facts in financial markets: long memory and jumps. The stock price process has jumps and the stochastic volatility process has long memory and jumps. Long memory in volatility occurs when the volatility shocks decay slowly and this dates back to Engle (1993). Models driven by fractional Brownian motion contain long memory. Fractional Levy process is a generalization of fractional Brownian motion to include jumps. If a Levy process has finite activity, then the corresponding fractional Levy process has finite variation. By replacing the Brownian motion with fractional Levy process in the classical Ornstein-Uhlenbeck process, one obtains the fractional Levy-Ornstein-Uhlenbeck (FLOU) process. Thus the model generalizes Barndorff-Neilsen and Shephard stochastic volatility model. Estimation of the parameters in the FLOU process based on stock price data is a hard problem. Parameter estimation in short memory Heston model is hard. We study minimum contrast estimation of the mean reversion speed parameter of the volatility process using high frequency data and study the asymptotic behavior of estimators. We suggest quasi-maximum likelihood estimators for the parameters of a FIECOGARCH (fractionally integrated exponential continuous time generalized autoregressive conditional heteroscedastic) process based on equally spaced and randomly spaced observations. First, we study the finite activity case: compound Poisson FIECOGARCH process and study the asymptotic behavior of the quasi-maximum likelihood estimator. Then, we study the infinite activity case: Ornstein-Uhlenbeck-Gamma process, a pure jump FLOU process. Next, we show that resulting estimators are consistent and asymptotically normal.

SatPM-TimeSlot-1: Session #4

An Optimal Credible Set for a Class of Priors

Sudip Bose
George Washington University-Washington D.C., USA
(E-mail: sudip@gwu.edu)

Abstract

We present a robust procedure for the construction of Bayesian credible sets. Instead of assuming that there is a single prior known to us, we model uncertain prior beliefs by a class C of priors that constitute a 'neighborhood' of a starting prior. Among the class of credible sets with given credibility under the starting prior, we determine the one that maximizes the minimum probability of including the parameter of interest. This credible set procedure is in fact C-minimax with respect to the risk function probability of non-inclusion. We work with three of the most widely-used classes in Bayesian robustness studies. We demonstrate an optimality property of the maximum likelihood region.

SunAM-TimeSlot-l: Session #33

On Selecting Among Binomial Populations

[1*]Elena M.Buzaianu and [2]Pinyuen Chen
[1]University of North Florida-Jacksonville, USA;
[2]Syracuse University-Syracuse, USA
(*Paper Presenter; E-mail: e.buzaianu@unf.edu)

Abstract

Thall, Simon, and Ellenberg (1988, *Biometrika*, 75, 303-310) proposed a two-stage selection and testing design in clinical trials for selecting the best experimental treatment and comparing it to a control. They dealt with a binomial setting where a patient's response may be characterized as either a success or a failure. A closed adaptive sequential procedure takes no more than n (predetermined) Bernoulli observations from each population and permits elimination of non-contending populations at each stage.

In this talk, we present closed adaptive sequential terminating rules, respectively, for both stages in Thall et al.'s two-stage design. We show that our adaptive sequential two-stage design achieves the same probability of a correct selection as the original two-stage design. Exact computations for small sample sizes and simulation studies for large sample sizes will be used to illustrate the savings over the original fixed-sample-size design. An example will be given to implement the proposed design.

SunAM-TimeSlot-3: Session #9

Statistical Considerations for Patient-Reported Outcomes

Joseph C. Cappelleri
Pfizer Inc, Global Research and Development-Connecticut, USA
(E-mail: joseph.c.cappelleri@pfizer.com)

Abstract

In February 2006 researchers at the Mayo Clinic convened with experts and Food and Drug Administration (FDA) personnel to facilitate discussion and dissemination of the FDA draft guidance on patient-reported outcomes (PROs) for use in medical product development to support labeling claims. Key topics were discussed at this Mayo Clinic/FDA meeting and eventually formalized and published in a special issue of *Value in Health* (2007, volume 10, supplement 2). Statistical considerations for PROs were among the key topics. This presentation therefore covers several areas of statistical import for consideration in the design, analysis, and interpretation of PROs for a regulatory label claim.

SatAM-TimeSlot-l: Session #37

Empirical Bayes Threshold Estimation in High Throughput Proteomics

Mark Carpenter
Auburn University-Alabama, USA
(E-mail: carpediem@auburn.edu)

Abstract

The literature is replete with theory and methods, as well as, software, to address estimation within general location-scale distributions. Most of the attention is spent on location estimation where either the distribution is symmetric with respect to location or a log-transform location/scale family is being studied (e.g., extreme value distribution). For left-truncated or positive support distributions, e.g., Weibull/exponential, the location parameter represents a threshold shift or truncation. In survival analysis this truncation parameter can be viewed as the minimum guarantee lifetime of an individual or group. In gene/protein expression analysis this parameter can be viewed as a minimum observable threshold value. In this paper, we examine threshold estimation with location mixtures of exponentials. Empirical Bayes estimators are derived, their performance studied and their application is assess using real data from a proteomic experiment.

SunAM-TimeSlot-l: Session #15

Impact of Dependence on the Stability of Model Selection in Supervised Classification for High-Throughput Data

*D. Causeur, M. Kloareg, and C. Friguet
IRMAR Applied Mathematics Department-Agrocampus Rennes, France
(*Paper Presenter; E-mail: david.causeur@agrocampus-rennes.fr)

Abstract

In the literature on statistical methodologies for high-throughput data, the impact of dependence between responses on properties of simultaneous inference has received increased scrutiny recently. This explains the need for a strong control of true proportion of false discoveries which is not guaranteed by the current methods for simultaneous inference. They are known to suffer from high instability in the presence of correlation between the response variables. Moreover, in some special microarray data, finding information on a correlation structure in high dimensional data is challenging while investigating complex regulation network in gene expression.

We propose a factor analysis model for the conditional covariance of the responses as a data reduction technique that identifies a part of shared variance. We show that a new multiple testing procedure that corrects each test statistics with respect to its contribution to the factor structure has desirable properties regarding both the power and the variance of the type-I error rate. Finally, we implement the present methodology to select the best sub-model for supervised classification. The gain with respect to usual selection procedures is illustrated via simulations and with a real microarray data example.

SatPM-TimeSlot-1: Session #41

Ranges of Measures of Association for Familial and Pedigree Binary Variables

[1] Yihao Deng and [2*] N. Rao Chaganty
[1] Indiana University Purdue University-Fort Wayne, USA;
[2]Old Dominion University-Virginia, USA
(*Paper Presenter; E-mail: rchagant@odu.edu)

Abstract

Analysis of familial and pedigree binary data plays an important role in genetic studies, linkage analysis, and epidemiologic research. Several measures are commonly employed to study the association between the familial and pedigree binary variables. These measures include correlations, odds ratios, kappa statistics and relative risks. In this talk we discuss permissible ranges of these measures of association. Understanding these ranges is the first step in developing efficient parameter estimation methods for statistical models for real life familial and pedigree binary data.

SatPM-TimeSlot-2: Session #26

Measurement Error Modeling for Noisy Point Patterns

*Avishek Chakraborty and Alan E Gelfand
Duke University-Durham, USA
(*Paper Presenter; E-mail: ac103@stat.duke.edu)

Abstract

We address the issue of observing a noisy point pattern. The unobserved true point process is modeled as a non-homogeneous Poisson process. For modeling the underlying intensity surface we are using a Gaussian mixture distribution. The noise that creeps in during the measurement procedure causes random displacement of the true locations. If our region of interest is bounded then this displacement may cause a true location within the boundary to be associated with an "observed" located outside of the region and thus missed. Also events outside the domain may “shift” inside and thus be recorded falsely. Depending on the variability in the measurement error as well as the number of locations close to boundary, this can cause a significant difference in number of locations between actual and recorded sets of data. Estimation of the intensity surface from the observed data can be significantly misleading and thus must be addressed in the model. We also allow the inclusion of training data to inform about the measurement error process as well as the inclusion of covariate information such as elevation to explain the intensity surface. We present the analysis of some simulated datasets to see how well our model performs.

SatAM-TimeSlot-l: Session #39

A Robust Test of Order Restrictions for the Variance Function of Non-Stationary Autoregressive Processes

[1*]Gabe Chandler and [2]Wolfgang Polonik
[1]Connecticut College-Connecticut; USA [2] University of California-Davis, USA
(*Paper Presenter; E-mail: gjcha2@conncoll.edu)

Abstract

The non-stationary autoregressive model in which the variance is allowed to change in time is a simple but useful model for many instances where homoscedasticity does not seem to hold. Such data arise in geology and finance, among other fields. Often times, the shape of the underlying variance function is informative. We present a robust test for the validity of order restrictions on the variance function, in particular the hypotheses of monotonicity or unimodality. The test is based on studying the location of large residuals on some set of interest, and not the size of the residuals themselves, thus robust

to heavy-tailed error distributions. The resulting test statistic is related to the Kolmogorov-Smirnov statistic from density estimation. The set of interest is chosen based on properties of the isotonic regression estimate of the variance function.

FriPM-TimeSlot-2: Session #23

Adaptive Trial Optimization, Re-Optimization, and Analysis

Mark Chang
Millennium Pharmaceuticals, Inc.-Massachusetts, USA
(E-mail: Mark.Chang@Statisticians.org)

Abstract

Adaptive trial design emphasizes on the integrated process, which view design, monitoring and analysis as an integrated process. A good adaptive trial should have solid design, monitoring, and analysis pieces in place at time of design the trial. This talk is a temp to address the challenging issues and suggest possible resolutions for the challenges. The determination of appropriate criteria for evaluation of adaptive design is critical. We'll outline several different criteria, and differentiate those good measures from bad measures. Unlike the classical design, an adaptive design allows for modifying trial aspects based on interim results of the trial or external information. The adaptations/modifications can be viewed as re-optimization using trial monitoring tools such as conditional power. We will discuss how to use these monitoring tools appropriately. For analysis, we often use bias-adjusted point estimate, adjusted p-values, and repeated confidence intervals. We will give a critical review on these report instruments, especially on philosophical views on the conditional versus unconditional p-values.

FriPM-TimeSlot-1: Session #3

A Resampling Based Study of Value at Risk

Snigdhansu Chatterjee
University of Minnesota-Minneapolis, USA
(E-mail: chatterjee@stat.umn.edu)

Abstract

The value-at-risk, corresponding to a particular probability and a period of time, is a measure of the maximum loss that may be incurred during that period with the stated probability. Properties of value-at-risk estimates are often obtained using historical simulations, or by using standard statistical (parametric, semiparametric or nonparametric) models on the nature of returns. We study the sensitivity of value-at-risk measures to distributional assumptions. We show that a resampling based approach leads to a high order accurate computation of some properties of value-at-risk measures.

SunAM-TimeSlot-2: Session #17

Variable Window Scan Statistics

[1*]Jie Chen and [2]Joseph Glaz
[1]University of Massachusetts-Boston, USA;
[2]University of Connecticut-Storrs, USA
(*Paper Presenter; E-mail: jie.chen@umb.edu)

Abstract

In this article approximations for the distributions of two-dimensional maximum scan score-type statistic and minimum p-value scan statistic are derived for independent and identically distributed Poisson distribution when the total number of events is known. Numerical results are presented to compare the power of these variable window type scan statistics with fixed single window scan statistics.

SatPM-TimeSlot-1: Session #4

Bayes Factor versus Other Model Selection Criteria for the Selection of Constrained Models

[1*]Ming-Hui Chen and [2] Sungduk Kim
[1] University of Connecticut- Storrs, USA;
[2]National Institute of Child Health and Human Development, USA
(*Paper Presenter; E-mail: mhchen@stat.uconn.edu)

Abstract

Four Bayesian criteria, including L measure, Deviance Information Criterion (DIC), Logarithm of the Pseudo-marginal Likelihood (LPML), and marginal likelihood (Bayes factor), are considered for comparing the inequality constrained ANOVA models. A class of priors for constrained ANOVA models based on the conjugate prior of Chen and Ibrahim (2003) is constructed and the properties of these priors are examined. Computational issues for various Bayesian criteria will be addressed for constrained models. The Dissociative Identity Disorder (DID) data is presented to illustrate and investigate how each of these four criteria selects the "best" model under various priors.

FriPM-TimeSlot-1: Session #3

The Determinants of Operational Losses

[1*] Anna Chernobai, [2] Philippe Jorion, and [3] Fan Yu

[1] Syracuse University-Syracuse, USA; [2] Michigan State University-East Lansing, USA;

[3] University of California-Irvine, USA

(*Paper Presenter; E-mail: annac@syr.edu)

Abstract

We examine the microeconomic and macroeconomic determinants of operational losses in financial institutions. Using 24 years of U.S. public operational loss data from 1980 to 2003, we demonstrate that the firm-specific environment is a key determinant of operational risk; firm- specific characteristics such as size, leverage, volatility, book-to-market, profitability, and the number of employees are all highly significant in our models. In contrast, while there is some evidence that operational losses are more frequent and more severe during economic downturns, overall the macroeconomic environment appears less important. We further test the doubly-stochastic Poisson assumption with respect to the arrivals of operational losses, given the estimated arrival intensities. Despite the traditional view that operational risk is unsystematic, we find evidence of clustering of operational risk events at the industry level in excess of what is predicted by the stochastic frequency estimates.

SunAM-TimeSlot-l: Session #33

Sequential Risk-Efficient Estimation for the Ratio of Two Binomial Proportions

Hokwon Cho
University of Nevada-Las Vegas, USA
(E-mail: cho@unlv.nevada.edu)

Abstract

A risk-efficient sequential point estimator is considered for the ratio of two independent binomial proportions based on maximum likelihood estimation under squared error loss and cost proportional to the observations. It is assumed that the cost per observation is constant. First-order asymptotic expansions are obtained for large-sample properties of the proposed procedure. Performance of the procedure is studied through the criteria of risk efficiency and regret analysis. Monte Carlo simulation is carried out to obtain the expected sample size that minimizes the risk and to examine its finite sample behavior. An example is provided to illustrate its use.

SatPM-TimeSlot-l: Session #25

A Bayesian Prediction for Undecided Voters

[1]Balgobin Nandram and [2*]Jai Won Choi
[1]Worcester Polytechnic Institute-Worcester,USA;
[2]Medical College of Georgia-Augusta, USA
(*Paper Presenter; E-mail: jchoi@mcg.edu)

Abstract

Data from election polls are typically presented in two-way categorical tables, and there are many polls before the election in November. For example, in the Buckeye State Poll in 1998 for governor there are three Polls, January, April and October; the first category represents the candidates (e.g., Fisher, Talft, other and undecided) and the second category represents the current status of the voters (likely to vote, not likely to vote and undecided). There is a substantial number of undecided voters. We use a Bayesian prediction to allocate the undecided voters to the three candidates. A Bayesian method permits modeling different patterns of missingness under ignorability and nonignorability assumptions, and a multinomial-Dirichlet model is used to estimate the cell probabilities which can help to predict the winner. We propose a time-dependent nonignorable nonresponse model for the three tables. As competitors we also consider two other models, an ignorable and a nonignorable nonresponse models, which assume a common stochastic process over time. Markov chain Monte Carlo methods are used to fit the models. We also construct a parameter that can be used to predict the winner among the candidates.

SunAM-TimeSlot-l: Session #35

A Review and Some New Ideas for Objective Priors in the Non-Parametric Context

B. Clarke
University of British Columbia-Vancouver, Canada
(E-mail: riffraff@stat.ubc.ca)

Abstract

Several Existing methodologies for deriving Objective priors outside the finite dimensional parameter context will be reviewed. Then two alternatives will be presented: the natural generalizations of the Reference Prior method for an increasing number of parameters of parameters and a variation of Rissanen's Prior. This is a joint work with S. Ghoshal.

SatPM-TimeSlot-2: Session #6

Bayesian Meta-Analysis Models for Microarray Studies

[1*]Erin M. Conlon, [2]Joon Jin Song, [3]Jun S. Liu
[1]University of Massachusetts-Amherst, USA;
[2]University of Arkansas-Fayetteville, USA;
[3]Harvard University-Cambridge, USA
(*Paper Presenter; E-mail: econlon@mathstat.umass.edu)

Abstract

Biologists often conduct multiple but different cDNA microarray studies that all target the same biological system or pathway. Pooling information across studies can help more accurately identify true target genes. Here, we introduce a Bayesian hierarchical model to combine cDNA microarray data across multiple independent studies to identify differentially expressed genes. Each study has several sources of variation, that is, replicate slides within repeated identical experiments. Our model produces the gene-specific posterior probability of differential expression, which is the basis for inference. In simulations combining two and five independent studies, we found that the meta-analysis model outperformed individual analyses using several comparison measures. We further illustrate our methods using biological data from the model organism *Bacillus subtilis*.

SunAM-TimeSlot-3: Session #43

An Efficient Algorithm to Mine Prodigious Frequent Patterns

[1*]Madhavi Dabbiru and [2]Mogalla Shashi
[1]Nagarjuna University-Vijayawada, India; [2]Andhra University-Visakhapatnam, India
(*Paper Presenter; E-mail: madhavicris@yahoo.co.in)

Abstract

Mining frequent itemsets is the core operation of Association Rule mining (Data mining algorithms). Abundant numbers of pattern mining algorithms are brought forth that are both effective and efficient. The growth of bioinformatics has resulted in datasets with high-dimensional characteristics. The existing 'closed', 'maximal' (Bayardo, 1998) frequent pattern algorithms such as LCM2 (Uno et al., 2004) and TFP (J. Wang et al., 2005) encounter challenges at mining rather large patterns, called prodigious frequent patterns, in the presence of an enormous number of frequent patterns. A new mining methodology called Pattern-Merger is presented here to efficiently find a good approximation to prodigious patterns. With Pattern-Merger, a prodigious pattern is discovered by merging its small fragments in one step, whereas the incremental pattern growth mining strategies such as those adopted in Apriori (Agrawal, 1994) and FP-growth (Han et al., 2000) examine a large number of midsized patterns. This property distinguishes Pattern-Merger from existing frequent pattern mining approaches and draws a new mining methodology. Further, we are investigating a method to deduce bounds of the obtained prodigious pattern if supports of most of its subsets (Yang et al., 2005, *IEEE*) are known.

SatAM-TimeSlot-l: Session #39

On Some New Bayesian Model Diagnostics for Generalized Linear Models

*Sourish Das and Dipak K. Dey
University of Connecticut-Storrs, USA
(*Paper Presenter; E-mail: sdas@stat.uconn.edu)

Abstract

In this paper, we develop Bayesian model diagnostic techniques for Generalized Linear Models (GLM) using prior elicitation on the canonical parameters. We consider elicitation of prior on canonical parameters, since it is much easier for expert, instead of parameters of interest, that is, the regression parameters. We broaden the horizon of simple diagnostic techniques of linear model, like Cook's distance under GLM framework. This new Generalized Cook's distance is invariant to the choice of prior. We also develop various other simple model diagnostic methods like detecting outliers using Highest Residual Density Interval. We present the performance of this new method on a real data set to investigate the efficacy of a new treatment for Osteoarthritis of knee. Hence it was important to identify if there exist a group of patients that are outliers or have strong influence on the final analysis of the study. This study was performed during the summer of 2006 where pain score was collected after 90 days of treatment. Biomarker information on TNFα cytokines was also recorded as covariates information in the study.

FriPM-TimeSlot-2: Session #13

Support Vector Machines: A Useful Tool for Classification

Shibasish Dasgupta

University of Florida-Gainesville, USA
(E-mail: dasgupta@stat.ufl.edu)

Abstract

Support Vector Machines (SVMs) has emerged as a very useful technique for classification. SVMs perform pattern recognition between two point classes by finding a decision surface determined by certain points of the training set, termed Support Vectors (SV). This surface, which in some feature space of possibly infinite dimension can be regarded as a hyperplane, is obtained from the solution of a quadratic programming that depends on a regularization parameter. This presentation is mainly about the application of support vector machines in classification problems though I will give an outline about the theoretical perspective behind SVM in the context of two-class classification problem. I try to give a brief description of the classification problem and explain in detail, the various techniques applied to explore the best possible method for classification for the specific problem.

SatAM-TimeSlot-1: Session #7

Reconstruction of Genetic Association Network

*Susmita Datta, Vasyl Pihur, and Somnath Datta
University of Louisville-Kentucky, USA
(*Paper Presenter; E-mail: susmita.datta@louisville.edu)

Abstract

Detecting genetic association/interaction network can potentially provide vast amounts of information about essential processes inside the cell. A complete picture of gene–gene associations/interactions would open new horizons for biologists, ranging from pure appreciation to successful manipulation of biological pathways for therapeutic purposes. Therefore, identification of important biological complexes whose members (genes and their products proteins) interact with each other is of prime importance. Although, there exists numerous experimental methods for the determination of genetic network they are costly and labor intensive and often not reproducible. Computational techniques, such as the one proposed in this work, provide an initial alternative solution that can be used as a screening tool before more expensive experimental techniques are attempted. Here, we introduce a novel computational method based on the partial least squares (PLS) regression technique for reconstruction of genetic networks from microarray data.

FriPM-TimeSlot-2: Session #36

Starting and Running an REU for Minorities and Women

*Dennis E. Davenport and Bonita Porter
Miami University-Ohio, USA
(*Paper Presenter; E-mail: davenpde@muohio.edu)

Abstract

The decreasing number of US citizens with advanced degrees in the mathematical sciences is a growing concern. Also of concern is the small number of advanced degrees in the sciences going to African Americans, Latinos, and women. Several Research Experience for Undergraduates (REU) programs have been developed to address these issues. In this presentation we describe the design of our ongoing program, the Summer Undergraduate Mathematical Sciences Research Institute (SUMSRI) in the Department of Mathematics and Statistics at Miami University.

SatPM-TimeSlot-l: Session #34

Jeffreys Priors for CAR Models

Victor De Oliveira
University of Texas-San Antonio, USA
(E-mail: victor.deoliveira@utsa.edu)

Abstract

Conditionally autoregressive (CAR) models have been extensively used for the analysis of spatial data in diverse areas, such as demography, economy, epidemiology and geography, both as models for (part of) the prior and the likelihood. In the latter case, the most common inferential method has been maximum likelihood, and the Bayesian approach has not been used much. This work proposes default (automatic) Bayesian analyses of CAR models. Two versions of Jeffreys prior, the independence Jeffreys and Jeffreys-rule priors, are derived for the parameters of CAR models and properties of the priors and resulting posterior distributions are obtained. The two priors and their respective posteriors are compared based on simulated data. Finally, frequentist properties of inferences based on maximum likelihood are compared with those based on the Jeffreys priors and the uniform prior.

FriPM-TimeSlot-2: Session #40

Discriminating Between Log-Normal and Weibull Distribution Under Type-I and Type-II Censoring

*Arabin Kumar Dey and Debasis Kundu
Indian Institute of Technology-Kanpur, India
(*Paper Presenter; E-mail: arabin@iitk.ac.in)

Abstract

In this paper, we propose a methodology for discriminating between log-normal and Weibull distributions in type-I and type-II censoring case. We use the ratio of maximized likelihoods in discriminating between the two distribution functions. We obtain the asymptotic distributions of the logarithm of the ratio of maximized likelihoods. It is used to determine the probability of correct selection between the two distributions. In both censoring case, we perform some simulation studies to observe how the asymptotic results work for different sample sizes. Two real data sets have been analyzed for illustrative purposes.

FriPM-TimeSlot-l: Session #27

Model Selection and Diagnostics to Identify Genetic Markers for Single-Nucleotide Polymorphisms

Dipak K. Dey
University of Connecticut-Storrs, USA
(E-mail: dipak.dey@uconn.edu)

Abstract

The distribution of genetic variation among populations is conveniently measured by Wright's F_ST which is a scaled variance taking on values in [0,1]. For certain types of genetic markers, and for single-nucleotide polymorphisms (SNPs), in particular, it is reasonable to presume that the genotype at most loci detected by those markers are selectively neutral. For such loci, the distribution of genetic variation among populations is determined by the size of local populations, the pattern and rate of migration among those populations, and the rate of mutation. Because the demographic parameters (population size and migration rates) are common across all loci, locus-specific estimates of F_ST will depart from a common mean only for loci with unusually high or low rates of mutation and for loci that are closely associated with genomic regions having a substantial effect on fitness. Thus, loci showing significantly more variation than background are likely to mark genomic regions subject to diversifying selection among the sample populations, while those showing significantly less variation than background are likely to mark genomic regions subject to stabilizing selection across the sample populations. We propose several Bayesian hierarchical models to estimate locus-specific effects on F_ST, and we apply these models to single nucleotide polymorphism data from the HapMap project.

SatAM-TimeSlot-1: Session #14

Hierarchical Models for the Estimation of Manatee Abundance from Aerial Surveys

Robert M. Dorazio
University of Florida-Gainesville, USA
(E-mail: bdorazio@ufl.edu)

Abstract

Predictions of manatee abundance as a function of habitat characteristics are needed for making conservation decisions for this threatened species. The spatial distribution of manatees in southwest Florida varies seasonally in response to changes in salinity, water temperature, food availability, and other factors related to habitat. Consequently, aerial surveys were developed to estimate abundance of manatees in spatially referenced sample units using a combination of sampling protocols. Groups of manatees were detected using double-observers, and the number of manatees in each group were detected by repeated circling to yield a sequence of "removal" counts. Thus, both kinds of counts (i.e., those of groups and those of manatees within groups) were spatially referenced. A hierarchical modeling framework is developed to estimate maps of manatee abundance while accounting for the imperfect detectability of groups and of individuals within groups. A critical component of these models is the functional dependence between the probability of detecting a group and the group's size, which is unknown, but estimable.

SatPM-TimeSlot-2: Session #2

Modeling Space-Time Data using Stochastic Differential Equations

[1*]Jason A. Duan, [2]Alan E. Gelfand, and [3]C. F. Sirmans
[1]Yale University-New Haven, USA; [2]Duke University-Durham, USA;
[3]University of Connecticut-Storrs, USA
(*Paper Presenter; E-mail: jd522@som.yale.edu)

Abstract

This project demonstrates the use and value of stochastic differential equations for modeling space-time data in two common settings. The first consists of point-referenced or geostatistical data where observations are collected at fixed locations and times. The second considers random point pattern data where the emergence of locations and times is random. For both cases, we employ stochastic differential equations to describe a latent process within a hierarchical model for the data. A motivating problem for the second setting is to model urban development through observed locations and times of new home construction; this gives rise to a space-time point pattern. We show that a spatio-temporal Cox process whose intensity is driven by a stochastic logistic equation is a viable mechanistic model that affords meaningful interpretation for the results of statistical inference. Other applications of stochastic logistic differential equations with space-time varying parameters include modeling population growth and product diffusion, which motivate our first, point-referenced data application. We propose a method to discretize both time and space in order to fit the model. We demonstrate the inference for the geostatistical model through a simulated dataset. Then, we fit the Cox process model to a real dataset taken from the greater Dallas metropolitan area.

SunAM-TimeSlot-3: Session #9

Effect of Torcetrapib on the Progression of Coronary Atherosclerosis

[1]Steven E. Nissen, [2]Jean-Claude Tardif, [1]Stephen J. Nicholls, [3]James H. Revkin,
[3]Charles L. Shear, [3*]William T. Duggan, [4]Witold Ruzyllo, [5]William B. Bachinsky,

[6]Gregory P. Lasala, and [1]E. Murat Tuzcu
[1]Cleveland Clinic-Ohio, USA; [2]Montreal Heart Institute-Montreal, Canada; [3]Pfizer Inc.-New London, USA; [4]Instytut Kardiologii-Warsaw, Poland, [5]Pinnacle Health at Harrisburg Hospital-Pennsylvania, USA, [6]Tchefuncte Cardiovascular Associates-Louisiana, USA
(*Paper Presenter; E-mail: william.t.duggan@pfizer.com)

Abstract

Levels of high-density lipoprotein (HDL) cholesterol are inversely related to cardiovascular risk. Torcetrapib is a cholesteryl ester transfer protein (CETP) inhibitor which increases HDL levels.

A total of 1188 subjects were randomized to receive either atorvastatin monotherapy or atorvastatin plus 60 mg of torcetrapib daily. After 2 years, progression of coronary disease was assessed by intravascular ultrasonography in 910 subjects.

The primary endpoint (change in percent atheroma volume) was not significant between the treatment groups (p = 0.72). Torcetrapib was associated with an increase in HDL, a decrease in LDL, and an increase in blood pressure. The lack of efficacy may be related to the inhibition of CETP or to some molecular toxicity.

SatAM-TimeSlot-l: Session #21

Subsampling the Mean of Spatial Lattice Data with Varying Expected Values

Magnus Ekström
Swedish University of Agricultural Sciences-Umeå, Sweden
(E-mail: Magnus.Ekstrom@sekon.slu.se)

Abstract

Subsampling methods have been suggested in the literature to nonparametrically estimate the variance of statistics computed from spatial data. Usually stationary data are required. However, in empirical applications, the assumption of stationarity often must be rejected.

When a statistic g is computed on some spatially indexed data observed in some region A of R^2, then the subsampling variance estimator of g uses 'replicates' of g computed on subshapes of A. Thus, if the statistic of interest is the sample mean over A, then the subsampling estimator uses the sample means computed on (overlapping) subshapes of A as 'replicates', i.e. the estimator is a normalized sample variance of the subshape means. Although this estimator can handle considerable heteroscedasticity, it is sensitive to variation in the expected values. Unless the subshape means have (essentially) the same expected value, the subsampling estimator of variance will fail, as it cannot distinguish the variation in the expected values from random variation.

Therefore, a modified subsampling estimator of the variance of (functions of) sample means is introduced. It is established, under weak moment and mixing conditions, that the modified estimator can handle heteroscedastic data that also possess a spatial trend (e.g., a periodic trend, a smooth trend, directional components or any combination of these). An example with applications to forestry, using satellite data, is also discussed.

SatPM-TimeSlot-2: Session #2

Gaussian Multiscale Spatio-Temporal Models

[1*]Marco A. R. Ferreira, [1]Scott Holan, and [2]Adelmo I. Bertolde
[1]University of Missouri–Columbia, USA; [2]Federal University of Espirito Santo, Brazil
(*Paper Presenter; E-mail: ferreiram@mssouri.edu)

Abstract

We develop a new class of multiscale spatio-temporal models for Gaussian areal data. Our framework decomposes the spatio-temporal observations and underlying process into several scales of resolution. Under this decomposition the model evolves the multiscale coefficients through time with structural state-space equations. The multiscale decomposition considered here, which includes wavelet decompositions as particular case, is able to accommodate irregular grids and heteroscedastic errors. Our multiscale spatio-temporal framework has several salient attributes. First, the multiscale decomposition leads to an extremely efficient divide-and-conquer estimation algorithm. Second, the multiscale coefficients have an interpretation of their own; thus, the multiscale spatio-temporal framework may offer new insight on understudied multiscale aspects of spatio-temporal observations. Finally, deterministic relationships between different resolution levels are automatically respected for both the observations, the latent process, and the estimated latent process. We illustrate the use of our multiscale framework with an analysis of a spatio-temporal dataset on agriculture production in the state of Espirito Santo, Brazil.

SunAM-TimeSlot-3: Session #42

Robust Semi-Parametric Estimators for Longitudinal Data

Daniel Gervini
University of Wisconsin-Milwaukee, USA
(E-mail: gervini@uwm.edu)

Abstract

Longitudinal data can sometimes be seen as discrete observations of a continuous-time stochastic process. In that case, it is of interest to estimate the mean and the covariance functions, or the principal components, of the process. The most commonly used estimators are smoothed versions of the sample mean and the sample covariance, which are very sensitive to outliers. Two types of outliers may be present in longitudinal datasets: isolated atypical observations within trajectories, or atypical individuals with entirely outlying trajectories.

In this talk we will present semi-parametric models based on the Student's t distribution and show that the resulting maximum likelihood estimators are resistant to both types of outliers. The estimators can be easily computed via the EM algorithm. As an example of application, we will analyze a dataset of CD4-count trajectories of AIDS patients.

SatPM-TimeSlot-2: Session #12

Joint Modeling of Longitudinal Data and Informative Dropout in the Presence of Multiple Changepoints with an Application to HIV-AIDS

[1]Pulak Ghosh, [2*]Kaushik Ghosh, and [3]Ram C. Tiwari
[1]Georgia State University-Georgia, USA; [2]University of Nevada Las Vegas-Nevada, USA;
[3]National Cancer Institute-Maryland, USA
(*Paper Presenter; E-mail: kaushik.ghosh@unlv.edu)

Abstract

In longitudinal studies of patients with the Human Immunodeficiency Virus (HIV), objectives of interest often include modeling of individual-level trajectories of HIV Ribonucleic Acid (RNA) as a function of time. Empirical evidence suggests that individual trajectories often possess multiple points of rapid change, which may vary from subject to subject --- both in number and in location. Presence of such changepoints makes the modeling of individual viral RNA levels difficult, since usual methods become unsuitable.

In this talk, we present a new robust multiple-changepoint model for longitudinal trajectories. The proposed method uses a joint model to incorporate information from the longitudinal data as well as from informative dropouts, which are common in such studies. A Dirichlet process prior is used to model the distribution of the changepoints. The Dirichlet process leads to a natural clustering, and thus, sharing of information among subjects with similar trajectories. A fully Bayesian approach for model fitting and prediction is implemented using the Gibbs sampler on the ACTG 398 clinical trial data.

SunAM-TimeSlot-l: Session #35

Objective Priors: A Selective Review

Malay Ghosh
University of Florida-Gainesville, USA
(E-mail: ghosh@stat.ufl.edu)

Abstract

Bayesian methods are increasingly applied in these days in the theory and practice of statistics. Any Bayesian inference depends on a likelihood and a prior. Ideally one would like to elicit a prior from related sources of information or past data. However, in its absence, Bayesian methods need to rely on some "objective" or "default prior", and the resulting posterior inference can still be quite valuable.

Not surprisingly, over the years, the catalog of objective priors also has become prohibitively large, and one has to set some specific criteria for the selection of such priors. Our aim is to review some of theses criteria, compare their performance, and illustrate them with some simple examples. While for very large sample sizes, it does not possibly matter what objective prior one uses, the selection of such a prior one uses, the selection of such prior does influence inference for small or moderate sample sizes. While without any nuisance parameters, Jeffreys' general rule prior, namely the positive square root of the determinant of the Fisher information matrix meets most of the desired selection criteria, it is possible to find some suitable alternate priors as well even in such cases.

SatPM-TimeSlot-2: Session #12

Bayesian Analysis for Longitudinal Semicontinuous Data

Pulak Ghosh
Georgia State University-Atlanta, USA
(E-mail: pulakghosh@gmail.com)

Abstract

In many biomedical applications, researchers encounter semicontinuous data whereby data are either continuous or zero. When the data are collected over time the observations may be correlated. Analysis of these kinds of longitudinal semi-continuous data is challenging due to the presence of strong skewness in the data. In this paper, we develop a flexible class of zero-inflated models in a longitudinal setting. We use a Bayesian approach to analyze longitudinal data from a acupuncture clinical trial in which we compare the effects of active acupuncture, sham acupuncture and standard medical care on chemotherapy-induced nausea in patients being treated for advanced stage breast cancer. A spline model is introduced into the linear predictor of the model to explore the possibility of nonlinear treatment effect. We also account for possible serial correlation between successive observations using Brownian motion. Thus, the approach taken in this paper provides for a more flexible modeling framework. We illustrate the Bayesian methodology with the acupuncture clinical trial data.

SunAM-TimeSlot-l: Session #31

Dimension Augmenting Vector Machine: A New General Classifier System for Large p Small n Problem

[1]Samiran Ghosh, [2]Dipak K. Dey, and [2]Yazhen Wang
[1]Indiana University Purdue University-Indianapolis, USA;
[2]University of Connecticut-Storrs, USA
(E-mail: samiran@math.iupui.edu)

Abstract

Support vector machine (SVM) and other reproducing kernel Hilbert space (RKHS) based classifier systems are drawing much attention recently due to its robustness and generalization capability. All of these approaches construct classifiers based on training sample in a high dimensional space by

using all available dimensions. SVM achieves huge data compression by selecting only few observations which are lying in the boundary of the classifier function. However when the number of observations are not very large (small *n*) but the number of dimensions are very large (large *p*), then it is not necessary that all available dimensions are carrying equal information in the classification context. Selection of only a useful fraction of the available dimensions will result in huge data compression. In this talk we will present and algorithmic approach by means of which such an optimal set of dimensions could be selected. In short we reverse the traditional idea of sequential data point selection to sequential dimension selection. To achieve this we have modified the solution proposed by Zhu and Hastie (2005, *J. Comp. Graphical Statist*., 14, 185-205) in the context of Import Vector Machine (IVM), to select an optimal sub dimensional model to build the final classifier.

SunAM-TimeSlot-1: Session #15

Optimal Designs for Model Identification and Discrimination

Subir Ghosh
University of California-Riverside, USA
(E-mail: subir.ghosh@ucr.edu)

Abstract

We can never be sure about a model that will describe a data best before actually collecting the data from an experiment. We can specify a class of possible models so that one of them will describe the data better over the other models within the class. In designing an experiment, we want to determine a design that will best identify all models and will discriminate between models within the class. We present such designs for fractional factorial experiments where the models within the class have the common parameters as the general mean and main effects, and the uncommon parameters as different two factor interactions. The common parameters between two models may also include some two factor interactions. We present optimum designs within the class of such designs with respect to six optimality criterion functions.

FriPM-TimeSlot-1: Session #8

Convergence of Sequential Bayesian D-Optimal Designs in Phase-I Clinical Trials

Subhashis Ghoshal
North Carolina State University-Raleigh, USA
(E-mail: ghoshal@stat.ncsu.edu)

Abstract

In clinical trials, the probability of responding to a treatment increases with the dose level, but at the same time chance of toxicity increases too. Due to ethical constraints on toxicity, a dose has to be within tolerable limits. The probability of a toxic reaction is often modeled by a cumulative distribution function with unknown location parameter alpha and scale parameter beta. In phase-I of a clinical trial, gathering maximum information so that tolerance limits of toxicity can be efficiently estimated is extremely important. This leads to the concept of the D-optimal design which maximizes the log-determinant of the Fisher information matrix, but because of its dependence on (alpha,beta), the design is not usable in practice. Often patients arrive sequentially so that a dose level can be assigned depending on the outcomes of previous treatments with the goal of maximizing the information about (alpha,beta). To this end, a prior is put on (alpha,beta) and a dose level is sequentially chosen maximizing the posterior expectation of the log-determinant of a sequential version of the Fisher information matrix. We show that the posterior for (alpha,beta) is consistent and this sequential Bayesian D-optimal design converges to the theoretical D-optimal design with increasing sample size. This is based on a joint work with Anindya Roy and William Rosenberger.

Fri 11:00AM – 11:45AM: Special Named Lecture-I

R. R. Bahadur Lecture

Uniform Limit Theorems for Wavelet Density Estimators

*Evarist Giné and Richard Nickl
University of Connecticut-Storrs, USA
(*Paper Presenter; E-mail: gine@math.uconn.edu)

Abstract

Given a bounded density, it is first shown that the linear wavelet estimator with compactly supported wavelets, introduced by Doukhan and Le ́on (1990) and Kerkyacharian and Picard (1992) satisfies a law of the logarithm analogous to the classical law of the logarithm for kernel density estimators, that it attains the minimax rate in sup-norm loss on Holder balls of densities, and that it satisfies the plug-in property in the sense that the cdf of this estimator approaches the true cdf in the sup norm at rate one over root n. These results are then used for proving that the Donoho-Johnstone-Kerkyacharian-Picard (1996) hard thresholding wavelet estimator is rate adaptive in sup norm loss to the unknown smoothness of the density and satisfies as well the plug in property. For this last result we do not have to assume that the density has compact support, but only that it has some integrability. It is also shown that another adaptive procedure consisting in applying Lepski's method based on the linear wavelet estimator produces a new adaptive estimator that has the same properties with respect to the sup norm as the thresholded one, with two advantages: one can use not only compactly supported wavelets, but also splines, and no integrability of the density is required.

SunAM-TimeSlot-2: Session #17

Variable Window Bayesian Scan Statistics

Joseph Glaz
University of Connecticut-Storrs, USA
(E-mail: joseph.glaz@uconn.edu)

Abstract

In this lecture two-dimensional variable window Bayesian scan statistics will be discussed. Algorithms used to implement testing procedures based on these statistics will be presented along with some numerical results. These testing procedures have interesting applications in the area of spatial statistics. Recent advances and open problems in this area will be discussed as well.

SunAM-TimeSlot-2: Session #18

Change Detection in the Distribution of Data Described by Time Series

Edit Gombay
University of Alberta-Edmonton, Canada
(E-mail: egombay@ualberta.ca)

Abstract

There is an extensive literature on detecting change in the parameters of autoregressive time series. The reason for this is that the AR(p) model allows writing the likelihood function in a simple form, and from this the maximum likelihood estimators are readily derived. This is not the case for moving average models, or for the more general model of linear processes.

Detecting change in the auto-covariance structure of linear processes will be considered in this talk. The application will be change-detection in the variance of moving average (MA) and autoregressive moving average (ARMA) models. To detect such changes is very important in economics, finance, environmental monitoring, etc, as they are signs of disturbances. The presented methods are very powerful, and solve a problem that has had only theoretical but no practical solution so far.

SatPM-TimeSlot-2: Session #26

Hybrid Dirichlet Process Models for the Analysis of Spatial Data

Michele Guindani
University of New Mexico-Albuquerque, USA
(E-mail: michele at stat.unm.edu)

Abstract

Recent Bayesian modeling of univariate spatial data has considered mixed effect models, where a residual stationary (homogeneous) Gaussian effect is assumed. Arguably, one might prefer the flexibility of a nonstationary, non-Gaussian specification. In a nonparametric setting, this can be accommodated by mixture of Dirichlet process (DP) models. The DP is an example of a species sampling prior, which are typically used to describe diversity of different ecological groups of species under different environmental conditions.

However, a limitation of the mixture of DP models is that the latent factor driving species sampling is globally defined and may fail to account for spatial heterogeneity. In this work, we introduce a novel class of prior distributions, the hybrid Dirichlet Processes (hDP), which generalize the DP and overcome this limitation. In a spatial setting, the hDP are defined as mixtures of Gaussian random fields with spatially varying weights. A crucial feature of this specification is the possibility to model local speciation and hybrid clustering. We illustrate the procedure by means of a simulated example and an application to the analysis of hippocampal atrophy in brains of patients affected by Alzheimer's disease. This is joint work with Alan Gelfand and Sonia Petrone.

SunAM-TimeSlot-l: Session #31

Detecting Patterns of Gene Regulation through Joint Bayesian Modeling of Genomic Sequence and ChIP-chip Data

[1]Jonathan A. L. Gelfond, [2*]Mayetri Gupta, and [3]Joseph G. Ibrahim
[1]University of Texas Health Science Center-San Antonio, USA; [2]Boston University-Boston,USA;
[3]University of North Carolina-Chapel Hill, USA
(*Paper Presenter; E-mail: gupta@bu.edu)

Abstract

We propose a unified framework for the analysis of Chromatin (Ch) Immunoprecipitation (IP) microarray (ChIP-chip) data for detecting transcription factor binding sites (TFBSs) or motifs. ChIP-chip assays focus the genome-wide search for TFBSs by isolating a sample of DNA fragments with TFBSs and applying this sample to a microarray with probes corresponding to tiled segments across the genome. Present methods use a two-step approach: (i) analyze tiling array data to estimate IP enrichment peaks then (ii) analyze the corresponding genomic sequences for regulatory motifs, independently of intensity information. The two-step approach leads to potentially high numbers of false positives and false negatives, and the array design often induces spatial correlations and gaps in the data which are ignored, leading to possible biases. Joint modeling of both sources of correlated data in a unified Bayesian framework has the potential to remove many such biases and provide an efficient and precise analysis method. Our proposed model integrates peak finding and motif discovery through a unified Bayesian hidden Markov model (HMM) framework that accommodates the inherent uncertainty in both measurements. A Markov Chain Monte Carlo algorithm is formulated, adapting recursive techniques used for HMMs. In simulations and applications to yeast data, the proposed methodology has highly improved sensitivity and specificity compared to currently available two-stage procedures.

SatPM-TimeSlot-1: Session #41

Distribution of Linear Function of Correlated Ordered Variables

*Pushpa L.Gupta and Ramesh C.Gupta
University of Maine-Orono, USA
(*Paper Presenter; E-mail: Pushpa.Gupta@umit.maine.edu)

Abstract

In this paper, we have considered the problem of finding the distribution of a linear combination of the minimum and the maximum for a general bivariate distribution. The general results are used to obtain the required distribution in the case of bivariate normal, bivariate exponential of Arnold and Strauss, absolutely continuous bivariate exponential distribution of Block and Basu, bivariate exponential distribution of Raftery, Freund's bivariate exponential distribution and Gumbel's bivariate exponential distribution. The distributions of minimum and maximum are obtained as special cases.

SunAM-TimeSlot-3: Session #19

General Frailty Model and Stochastic Orderings

[1*]Ramesh C. Gupta and [2]Rameshwar D. Gupta
[1] University of Maine-Orono, USA;
[2]University of New Brunswick-St. john, Canada
(*Paper Presenter; E-mail: Ramesh.Gupta@umit.maine.edu)

Abstract

In this paper, we propose a general frailty model and develop its properties including some results for stochastic comparisons. More specifically, our main result lies in seeing how the well-known stochastic orderings between distributions of two frailties translate into the orderings between the corresponding survival functions. The results are used to obtain the properties of the classical multiplicative frailty model and the additive frailty model. Several of the results, in the literature, are obtained as special cases.

FriPM-TimeSlot-2: Session #23

A Statistician's Perspective of Pharmacokinetic/Pharmacodynamic Modeling

Alan Hartford
Merck Research Laboratories, Inc.-North Wales, Pennsylvania, USA
(E-mail: alan_hartford@merck.com)

Abstract

Pharmacokinetic/Pharmacodynamic Modeling and Simulation has been practiced since at least as early as the 1970's. Systems of differential equations are used to model PK behavior of a drug and to model how the pharmacokinetics are intertwined with the pharmacodynamics. The solutions to the system of differential equations are generally nonlinear in the PK/PD parameters. So when intra- and inter-subject variability is included in models, sophisticated maximum likelihood algorithms are required. Although this research employs complex statistical methods, few statisticians have focused their attention in this area. This talk will explain how PK/PD models are set-up and how they are fitted in practice using these maximum likelihood approaches with the hopes of sparking more interest from statisticians for this field.

SatAM-TimeSlot-1: Session #14

Modeling Age and Nest-Specific Survival Using a Hierarchical Bayesian Approach

[1] Jing Cao and [2*] Chong He

[1] Southern Methodist University-Dallas, USA; [2] University of Missouri-Columbia, USA
(*Paper Presenter; E-mail: hezh@missouri.edu)

Abstract

Recent studies have shown that grassland birds are declining more rapidly than any other group of terrestrial birds. Current methods of estimating age-specific bird nest survival rates require knowing the ages of nests or assuming homogeneous nests in terms of nest survival rates or treating the hazard function as a piecewise step function. We propose a Bayesian hierarchical model with nest-specific covariates to estimate age-specific daily survival probabilities without the above requirements. The model provides a smooth estimate of the nest survival curve and identifies the factors that influence the nest survival curve when the nest age is unknown. Biologists will be able to develop effective conservation and land management strategies with this information. The model can handle irregular visiting schedules and it has the least restrictive assumptions compared to existing methods. The typical features of nest survival data, truncation and censoring, are accounted for by the likelihood function and the latent variables. An intrinsic auto-regressive prior is employed for the nest age effect. This nonparametric prior provides a more flexible alternative to the parametric assumptions. The Bayesian computation is efficient because the full conditional posterior distributions either have closed forms or are log-concave. We use the method to analyze a Missouri dickcissel data set and find that (1) nest survival is not homogeneous during the nesting period, and it reaches its lowest at the transition from incubation to nestling; and (2) nest survival is influenced by vegetation coverage and vegetation height in the study area.

SunAM-TimeSlot-2: Session #11

False Discovery Rates for Discrete Data

Joseph F Heyse
Merck Research Laboratories-North Wales, USA
(E-mail: joseph_heyse@merck.com)

Abstract

Almost all multiple comparison and multiple endpoint procedures applied in clinical trials are designed to control the Family Wise Error Rate (FWER) at a prespecified level of alpha = 0.05. Benjamini and Hochberg (1995) argued that in certain settings, requiring strict control of the FWER is often too conservative. They suggested controlling the False Discovery Rate (FDR), defined as the expected proportion of true (null) hypotheses that are incorrectly rejected. When one or more of the hypotheses being tested uses a discrete data endpoint then it is possible to further increase the power of both FWER and FDR controlling procedures. Methods proposed by Tarone (1990) and Gilbert (2005) have increased power by using the discreteness in the data to reduce the effective number of endpoints considered for the multiplicity adjustment. A modified fully discrete FDR sequential procedure is introduced that uses the exact distribution of potential outcomes. The potential gains in power are estimated using simulation. Application of FDR in the setting of clinical safety data analysis is reviewed and other potential uses of the proposed method are discussed.

SunAM-TimeSlot-2: Session #18

Break Detection in the Covariance Structure of Multivariate Time Series Models

[1]Alexander Aue, [2*]Siegfried Hörmann, [2]Lajos Horváth, [2]Matthew Reimherr
[1]University of California-Davis, USA;
[2]University of Utah-Salt Lake City, USA
(*Paper Presenter; E-mail: alexaue@wald.ucdavis.edu)

Abstract

Many standard methods in time series analysis work under the assumption that the data investigated have been sampled from some weakly stationary process. Especially, this is the crucial

requirement in order that we can produce meaningful estimates and reliable forecasts. It is therefore of considerable interest to detect structural breaks in the data before any other statistical analysis is carried out.

The purpose of this talk is to present methods to identify structural breaks in the volatilities and cross-volatilities of multivariate time series data. Our results apply to several important linear and non-linear models, containing multivariate ARMA and different multivariate GARCH models.

FriPM-TimeSlot-2: Session #20

Middle-Censoring for Circular Data

[1*]S. Rao Jammalamadaka and [2]Mangalam Vasudevan
[1]University of California-Santa Barbara, USA; [2]Universiti Brunei-Darussalam, Brunei.
(*Paper Presenter; E-mail: rao@pstat.ucsb.edu)

Abstract

A generalized censoring scheme in the survival analysis context was introduced by Jammalamadaka and Mangalam (2003, *J. Nonparametric Statist*., 253-265). In this talk we discuss how such a censoring scheme applies to circular data and in particular when the original data is assumed to come from a parametric model such as the von Mises. Maximum likelihood estimation of the parameters as well as their large sample properties are considered under this censoring scheme. We also consider nonparametric estimation of the circular probability distribution under such a general censoring scheme and use Monte Carlo methods to investigate its effects on the estimation of the mean direction and concentration.

SatAM-TimeSlot-1: Session #14

Bayesian Inference for Marine Mammal Telemetry Data: A Continuous-Time Approach

[1*]Devin S. Johnson, [1]Joshua M. London, [2]Mary-Anne Lea,
[1]John W. Durban, and [1]Carey E. Kuhn
[1]National Marine Mammal Laboratory-Seattle, USA;
[2]University of Tasmania-Hobart, Australia
(*Paper Presenter; E-mail: devin.johnson@noaa.gov)

Abstract

Marine mammal telemetry data is obtained by determining the location of an animal in space at several points in time. The observed locations are function of two important factors, true location and measurement error. In order to handle the fact that most telemetry data is collected opportunistically on an irregular basis we consider a movement model stochastic process in continuous time. Using this approach likelihood can be formed using the raw data instead of aggregated, thinned, or interpolated data. Bayesian methods are explored for making inference on animal locations as well as other movement quantities of interest, such as travel speed and habitat use. The Bayesian inference paradigm allows for inclusion of parameter uncertainty in location estimation, as well as, propagation of uncertainty through nonlinear relationships in the movement parameters of interest. Several model extensions which we are investigating will be discussed. These include change-point models, for making inference on behavioral shifts, and a model for detection of measurement error outliers.

SatAM-TimeSlot-l: Session #28

Bayesian Semiparametric Modeling of fMRI Data at the Population Level

Timothy D. Johnson
University of Michigan-Ann Arbor, USA
(E-mail: tdjtdj@umich.edu)

Abstract

The aim of this work is to develop a spatial semiparametric Bayesian model for multi-subject fMRI data. While there has been much work on univariate modeling of each voxel for single- and multi-subject data, and some work on spatial modeling for single-subject data, there has been no work on spatial models that explicitly account for inter-subject variability in activation location. The data are fitted with a Bayesian semiparametric hierarchical spatial model. At the first level we model "population centers" that mark activation centers. At the second level subjects' "individual centers" are associated with population centers. At the third level individual activation regions are fitted with "individual components" around individual centers. While most previous work uses Gaussian mixtures for the activation shape, at the fourth level we instead use Gaussian mixtures for the probability that a voxel belongs to an activated region. Our approach incorporates the unknown number of mixture components into the model as a parameter whose posterior distribution is estimated by reversible jump Markov chain Monte Carlo (RJMCMC) at levels three and four. A mixture of Dirichlet process priors (MDP) is used to nonparametrically model the distribution of individual components about the population centers at level 2. We demonstrate our method with an fMRI study of resolving proactive.

FriPM-TimeSlot-l: Session #22

Precedence Tests for Progressively Censored Samples

[1] N. Balakrishnan, [2] Ram C. Tripathi, and [2*] Nandini Kannan
[1]McMaster University-Hamilton, Canada; [2]University of Texas-San Antonio, USA
(*Paper Presenter; E-mail: Nandini.kannan@utsa.edu)

Abstract

In many problems in reliability and survival analysis, the researcher is often interested in the comparison of two or more populations. For example, while comparing a treatment group with a control group, one may be interested in determining whether the observations in the treatment group have a longer lifetime than those from the control group, i.e., whether the treatment is effective or not. In these kinds of experiments, a decision based on early failures would be of great value. In this talk, we consider independent progressively Type-II censored random samples from two populations with cumulative distribution function's (cdf) F and G respectively, and discuss a precedence test for testing the equality of the two distributions based on placements. The exact null distribution of the test statistic is derived. We also provide the rejection regions for fixed levels of significance and various sample sizes and different progressive censoring schemes. We also consider the problem of determining the power under specific alternatives.

SatPM-TimeSlot-3: Session #16

Multi-objective Optimal Experimental Designs for Event-Related fMRI Studies

*Ming-Hung Kao, Abhyuday Mandal, Nicole Lazar, and John Stufken
University of Georgia-Athens, USA
(*Paper Presenter; E-mail: jasonkao@uga.edu)

Abstract

Well planned experimental designs are crucial to successfully achieving statistical goals under psychological restrictions in functional magnetic resonance imaging (fMRI) studies. With sophisticated allocations of stimuli, researchers can gather valuable fMRI time series and acquire precise information about human brain activities. However, due to the nature of fMRI experiments, the underlying design space is very large and irregular. This makes it difficult to find an optimal design that simultaneously accomplishes various goals of a study and fulfills the scientific restrictions. In this work, the design measurements evaluating the "goodness" of a design with respect to statistical aims and psychological constraints are defined and a multi-objective design criterion is created by combining these measurements. We also propose a modified genetic algorithm to efficiently search over the design space. Taking advantage of well-known fMRI designs, this algorithm, which is also based on a more rigorous model formulation and consistent objective functions, significantly outperforms previous search algorithms in terms of achieved efficiency, computation time and convergence rate.

SatPM-TimeSlot-1: Session #1

Two-stage Equivalence Tests that Control Both the Size and Power

Kazuyoshi Yata
University of Tsukuba-Ibaraki, Japan
(E-mail: yata@math.tsukuba.ac.jp)

Abstract

We consider equivalence tests that control both a size and power for a different of means from two normally distributed populations with unknown variances. We give equivalence test procedures in two-stage sampling for both the cases that two variances are equal and two variances are unequal. The test procedures control both a size and power with the second--order accuracy. We also consider some related problems in clinical trials. Actual examples are given to illustrate how it should be done by using the methodologies.

SatPM-TimeSlot-3: Session #16

Recent Development in Nonlinear Statistical Modeling and Model Selection

Sadanori Konishi
Faculty of Mathematics, Kyushu University, Japan
(E-mail: konishi@math.kyushu-u.ac.jp)

Abstract

Statistical modeling is a critical tool in scientific data analysis. Models are used to understand phenomena with uncertainty, to determine the structure of complex systems, and to control such systems as well as to make reliable predictions in various natural and social science fields. Considerable effort has been made in establishing practical methods of modeling complex structures of stochastic phenomena. We first give a systematic account of some recent developments in nonlinear statistical modeling and also model selection. Second nonlinear modeling techniques are applied to the analysis of data with complex structure and/or high-dimensional data, including the functional data. Our modeling strategies are used to express discrete observations in the form of a function, and then draw information from a collection of functional data.

FriPM-TimeSlot-l: Session #22

Bayesian Inference and Life Testing Plan for Weibull Distribution in Presence of Progressive Censoring

Debasis Kundu
Indian Institute of Technology-Kanpur, India
(E-mail: kundu@iitk.ac.in)

Abstract

This paper deals with the Bayesian inference of unknown parameters of the progressively censored Weibull distribution. It is well known that for a Weibull distribution, while computing the Bayes estimates, the continuous conjugate joint prior distribution of the shape and scale parameters, does not exist. In this paper it is assumed that the shape parameter has a log-concave prior density function and for the given shape parameter the scale parameter has a conjugate prior distribution. As expected, when the shape parameter is unknown, the closed form expressions of the Bayes estimators cannot be obtained.

We use Lindley's approximation to compute the Bayes estimates and the Gibbs sampling procedure to calculate the credible intervals. For given priors, we also provide a methodology to compare two different censoring schemes and hence to find the optimal Bayesian censoring scheme. Monte Carlo simulations are performed to observe the behavior of the proposed methods and one data analysis is performed for illustrative purposes.

SatPM-TimeSlot-l: Session #25

Hierarchical Bayes Modeling of Small Area Proportions from Complex Survey Data

[1]Benmei Liu, [2*]Partha Lahiri, and [1]Graham Kalton
[1]Westat-Rockville, USA; [2]University of Maryland-College Park, USA
(*Paper Presenter; E-mail: plahiri@survey.umd.edu)

Abstract

When a Hierarchical Bayes area level model is used to produce estimates of proportions of units with a given characteristic for small areas, it is commonly assumed that the survey weighted proportion for each sampled small area has a normal distribution with known sampling variance. However, the assumptions of known sampling variances and normality are problematic when the small area sample size is small or when the true proportion is near zero or one. In an effort to overcome these problems, we propose an alternative modeling of the survey weighted proportion based on the beta distribution. We compare the results obtained from this alternative modeling with those obtained from a few commonly used modeling approaches using a Monte Carlo simulation study in which samples are generated from fixed finite population using both equal probability of selection (epsem) and non-epsem sampling designs.

SunAM-TimeSlot-l: Session #32

Network Modeling with Applications to Hedge Fund Returns

Yoonjung Lee
Harvard University-Cambridge, USA
(E-mail: ylee@stat.harvard.edu)

Abstract

Systemic risk describes the likelihood of the collapse of a financial system, such as a market crash or a breakdown of the banking system. In the wake of the recent growth of hedge funds, understanding joint dynamics of hedge fund returns becomes central in gauging the systemic risk. Embedding a social network model in a hierarchical Bayesian modeling framework, the paper explores an alternative way of characterizing the risk of hedge fund returns whose non-linear dependence dynamics are difficult to capture within a standard econometric time series model. The proposed approach allows us to map hedge fund returns on an easy-to-interpret structure, to identify a few funds that may be more influential than other funds, and to reduce a high-dimensional volatility estimation problem into a low-dimensional one.

SatPM-TimeSlot-1: Session #44

A Multivariate Spatial Model for Prediction of Storm Outages

[1*]Hongfei Li and [1]Jonathan R. M. Hosking

[1] IBM T. J. Watson Research Center-Yorktown Heights, USA
(*Paper Presenter; E-mail: liho@us.ibm.com)

Abstract

We have developed a Bayesian hierarchical model for multivariate spatial data, with the aim of predicting the number of outages suffered by customers of an electric utility as a result of severe weather events such as high winds, thunderstorms, and tornadoes. Specific attention is given to building models that incorporate covariates as well as spatial correlation in a multivariate context. Bayesian inference is implemented by means of a MCMC algorithm.

SatPM-TimeSlot-2: Session #6

Gene Set Enrichment Analysis for Non-Monotone Association and Multiple Experimental Categories

[1] *Rongheng Lin, [2] Shuangshuang Dai, [3] Richard D. Irwin, [3] Alexandra N. Heinloth,
[4] Gary A. Boorman, and [3] Leping Li
[1]University of Massachusetts, USA; [2]Alpha-Gamma Technologies, Inc., USA;
[3]National Institute of Environmental Health Science, NIH, USA; [4]Covance Inc., USA
(*Paper Presenter; E-mail: rlin@schoolph.umass.edu)

Abstract

Recently, microarray data analyses using functional pathway information, e.g., gene set enrichment analysis (GSEA) (Subramanian et al., 2005, *PNAS*, 102: 15545–15550) and significance analysis of function and expression (SAFE) (Barry et al., Bioinformatics 2005, 21(9): 1943–1949), have gained recognition as a way to identify biological pathways/processes associated with a phenotypic endpoint). In these analyses, a local statistic is used to assess the association between the expression level of a gene and the value of a phenotypic endpoint. Then these gene-specific local statistics are combined to evaluate association for pre-selected sets of genes. Commonly used local statistics include t-statistics for binary phenotypes and correlation coefficients that assume a linear or monotone relationship between a continuous phenotype and gene expression level. Methods applicable to continuous non-monotone relationships are needed. Furthermore, for multiple experimental categories, methods that combine multiple GSEA/SAFE analyses are needed.

For continuous or ordinal phenotypic outcome, we propose to use as the local statistic the coefficient of multiple determination (i.e., the square of multiple correlation coefficient) from fitting natural cubic spline models to the phenotype-expression relationship. Next, we incorporate this association measure into the GSEA/SAFE framework to identify significant gene sets. Furthermore, we describe a procedure for inference across multiple GSEA/SAFE analyses. We illustrate our approach using gene expression and liver injury data from liver and blood samples from rats treated with eight hepatotoxicants under multiple time and dose combinations.

SatPM-TimeSlot-2: Session #5

Information Conversion and Bayesian Effective Sample Size

Xiaodong Lin
University of Cincinnati-Cincinnati, USA
(E-mail: linxd@math.uc.edu)

Abstract

In high cost experiments, it is imperative that we develop new efficient techniques to utilize all relevant data from possible resources to improve prediction accuracy and inference. One typical example is the case where we have two data sets generated from two experiments to estimate the same parameters. In order to create a unified larger dataset, it is critical to accurately convert the data from one likelihood to the other. We proposed a novel method in Bayesian framework to compute the effective sample in this context. The key technique in this procedure is to minimize the K-L distance between the two posterior distributions of the concerned parameters given the two datasets respectively. This procedure has been successfully applied to combine two gene expression level datasets, one from Sun Yat-Sen cancer center in Taipei, and the other from Duke University Medical Center. The combined data enables us to derive a more accurate parameter estimation than other methods.

FriPM-TimeSlot-2: Session #13

Sparse Discriminant and Classification Methods for Genetic Pathways

*Lingsong Zhang and Xihong Lin
Harvard University-Boston, USA
(*Paper Presenter; E-mail: ZHANG@hsph.harvard.edu)

Abstract

An increasing challenge in analysis of genomic data is how to interpret and gain biological insight of profiles of thousands of genes. There is an increasing interest in analysis of genomic data by incorporating prior biological knowledge using gene sets and genomic pathways, which consist of groups of biologically similar genes. Such approaches allow one to study the joint effects of a group of genes. Existing methods include over-representation analysis, gene set enrichment analysis, principal component analysis, global test, and kernel machine. However, these pathway analysis methods do not provide a selection of important genes in the pathway and the analysis can be dominated by the noises of non-informative genes. We propose sparse linear discriminant analysis (SLDA) and sparse distance weighted discriminant analysis (SDWD) for genetic pathway data, which allow us to study the joint effects of genes within a pathway while selecting important genes that drive the differences. We provide an efficient path algorithm to obtain the solution. We illustrate these methods by application to a type II diabetes data set and a metal fuse exposure data set.

SatPM-TimeSlot-2: Session #6

Generalized Spline Mixed-effects Models with Applications in AIDS Clinical Trials

[1*]Anna Liu and [2]Hua Liang
[1]University of Massachusetts-Amherst, USA;
[2]University of Rochester Medical Center-Rochester, USA
(*Paper Presenter; E-mail: anna@math.umass.edu)

Abstract

This paper is concerned with effective estimation and testing of the population and subject-specific curves for longitudinal data using generalized semiparametric mixed-effects models. Spline based estimation methods for the nonparametric functions have gained popularity because of their flexibility and relatively easy implementation due to their connection with linear mixed-effects models. For generalized linear mixed-effects models, full likelihood analyses have been hindered by high-dimensional integrations. The contribution of the paper is two-fold: we apply the spherical-radial integral approximation method proposed by Monahan and Genz (1997, *J. Amer. Statist. Assoc.*, 92, 664-674), to generalized semiparametric mixed-effects models with population and individual nonparametric random effects, and we develop the likelihood ratio test for hypotheses on parametric and nonparametric components. Both estimation and tests are compared with the Laplace approximation based ones and it is found that the former has improved performances, especially on estimating the variance components and hypothesis testing with a relatively large number of random effects. The favorable performance of the spherical-radial approximation and its inexpensive computational cost should make it a useful tool for practitioners. The proposed methodological work is motivated by modeling of virologic and immunologic responses in an AIDS clinical trial.

FriPM-TimeSlot-2: Session #13

Estimation and Testing for the Effect of a Genetic Pathway on a Disease Outcome Using Logistic Kernel Machine Regression via Logistic Mixed Models

[1*] Dawei Liu, [2]Debashis Ghosh, and [3] Xihong Lin

[1] Brown University-Providence, USA; [2] Pennsylvania State University-State College, USA;
[3]Harvard School of Public Health-Cambridge, USA
(*Paper Presenter; E-mail: daweiliu@stat.brown.edu)

Abstract

Growing interest on biological pathways has called for new statistical methods for modeling and testing the multi-dimensional pathway effect. In this talk, we propose a semiparametric logistic regression model for binary outcomes, where the clinical effects are modeled parametrically and the genetic pathway effect is modeled nonparametrically using kernel machines. The nonparametric function of a genetic pathway allows for the possibility that genes within the same pathway are likely to interact with each other and relate to the clinical outcome in a complicated way. We show that the kernel machine estimate can be formulated using a generalized linear mixed model. Estimation hence can proceed within the generalized linear mixed model framework using standard mixed model software. A score test based on a Gaussian process approximation is developed to test for the genetic pathway effect. The methods are illustrated using a prostate cancer data set and evaluated using simulations.

SatPM-TimeSlot-2: Session #2

On Mixture of Kalman Filtering and Learning

[1*]Hedibert F. Lopes, [1]Carlos Carvalho, [2]Michael Johannes and [1]Nicholas Polson
[1]University of Chicago-Chicago, USA; [2]Columbia University-New York, USA
(*Paper Presenter; E-mail: hlopes@chicagogsb.edu)

Abstract

In this paper we present a novel particle filtering and learning strategy for a wide class of state space models that can be represented as mixture of dynamic linear models (DLMs). These methods provide samples from the joint posterior distribution of states and parameters, in a sequential fashion, avoiding the burden of “hard to converge” MCMC samplers. Our methodology provides an extension to the mixture of Kalman filters (Liu and Chen, 2000) and naturally incorporates nonlinearities in the state dynamics. We use conditional sufficient statistics for parameter learning and we extend this approach to state filtering whenever possible. We provide two applications. First, a dynamic factor switching model which illustrates the efficiency gains over traditional methods. Second, we analyze a nonlinear model that has been extensively considered in the pure filtering literature, where we also add sequential parameter learning.

SatPM-TimeSlot-3: Session #16

Optimal Designs for Models with Potential Censoring

[1*]J. López-Fidalgo, [2]María Jesús Rivas-López, and [2]Sandra Garcet--Rodríguez
[1]University of Castilla-La Mancha; [2]University of Salamanca
(*Paper Presenter; E-mail: jesus.lopezfidalgo@uclm.es)

Abstract

This presentation deals with optimal design theory for models with potential censoring either on independent or dependent variables. On the one hand optimal approximate designs when an independent variable might be censored is considered. The problem is which design should be applied to obtain an optimal approximate design when the censored distribution function is assumed known in advance. The approach for finite and continuous design spaces deserves different attention. In both cases equivalent theorems and algorithms are provided in order to calculate optimal designs. Some examples illustrate this approach for D-optimality.

On the other hand a development of the optimal design theory is carried out for a particular Cox Regression problem. The failure time is modeled according to a probability distribution depending on some explanatory variables through a linear model. At the end of the study some units will have not failed and thus their time records will be censored. In order to deal with this problem from an experimental design point of view it will be necessary to assume a probability distribution of the time of debut of an experimental unit in the study. Then an optimal conditional design will be computed at the beginning of the study for any possible given time of debut. Thus, every time a new unit enters the study there is an experimental design to be applied. A particular and simple case is used throughout the presentation in order to illustrate the procedure.

FriPM-TimeSlot-2: Session #23

Data Mining Issues in Drug Development

David Madigan
Columbia University-New York, USA
(E-mail: madigan@stat.columbia.edu)

Abstract

Data mining methods play an increasingly important role in drug safety. Prior to approval, clinical trial data can provide important insights into potentially unforeseen drug side effects. Following approval, important data sources include spontaneous report databases, claims databases, and medical record systems. Challenging statistical issues arise in all of these contexts.

This talk will review the general area focusing especially on flaws in the current standards for assessing drug safety during the approval process and on recent developments in statistical methodology for monitoring post-approval safety in longitudinal medical claims databases.

SunAM-TimeSlot-3: Session #42

Efficient Estimation of Population-Level Summaries in General Semiparametric Regression Models

[1*]Arnab Maity, [1]Yanyuan Ma, and [1]Raymond J. Carroll
[1]Department of Statistics, Texas A&M University-College Station, USA
(*Paper Presenter; E-mail: amaity@stat.tamu.edu)

Abstract

This article considers a wide class of semiparametric regression models in which interest focuses on population-level quantities that combine both the parametric and the nonparametric parts of the model. Special cases in this approach include generalized partially linear models, generalized partially linear single-index models, structural measurement error models, and many others. For estimating the parametric part of the model efficiently, profile likelihood kernel estimation methods are well established in the literature. Here our focus is on estimating general population-level quantities that combine the parametric and nonparametric parts of the model (e.g., population mean, probabilities). We place this problem in a general context, provide a general kernel-based methodology, and derive the asymptotic distributions of estimates of these population-level quantities, showing that in many cases the estimates are semiparametric efficient. For estimating the population mean with no missing data, we show that the sample mean is semiparametric efficient for canonical exponential families, but not in general. We apply the methods to a problem in nutritional epidemiology, where estimating the distribution of usual intake is of primary interest and semiparametric methods are not available. Extensions to the case of missing response data are also discussed.

SunAM-TimeSlot-l: Session #30

Generalized Convolution Models for Nonstationary Multivariate Spatial Processes

[1*]Anandamayee Majumdar, [2]Debashis Paul, and [3]Dianne Bautista
[1]Arizona State University-Tempe, USA; [2]University of California-Davis, USA;

[3]Ohio State University-Columbus, USA
(*Paper Presenter; E-mail: ananda@math.asu.edu)

Abstract

We propose a general constructive method for specifying flexible classes of nonstationary stochastic models for multivariate spatial data. The method is based upon convolutions of spatially varying covariance functions and produces mathematically valid covariance structures. This method generalizes the convolution approach suggested by Majumdar and Gelfand (2007, *Math. Geology*, 39, 225-245) to extend multivariate spatial covariance functions to the nonstationary case. A Bayesian method for estimation of the parameters in the covariance model based on a Gibbs sampler is proposed and carried out on simulated data.

FriPM-TimeSlot-l: Session #27

Bayesian Semiparametric Modeling of MPSS Data: Gene Expression Analysis of Bovine *Salmonella* Infection

Bani K. Mallick
Texas A&M University-College Station, USA
(E-mail: bmallick@stat.tamu.edu)

Abstract

Expression profiling techniques based on transcript counting offer a powerful alternative to the conventional microarray technology and address some of its shortcomings. Among the counting-based technologies, massively parallel signature sequencing (MPSS) has some advantages over competitors such as serial analysis of gene expression (SAGE) or direct sequencing of cDNA and is ideal for building complex relational databases for gene expression. In the existing literature, MPSS data (or some transformation thereof) have been modeled by continuous densities asymptotically or empirically. Here we use zeo-inflated Poisson (ZIP) distribution to develop Bayesian models. We adopt two Bayesian hierarchical models---one parametric and the other, semiparametric with a Dirichlet process prior. The deviance information criterion (DIC) is used for model choice. The goal is to identify differentially expressed signatures and the inference on differential expression is based on the posteriors of the 'treatment effect' parameters. To this end, symmetrized Kullback-Leibler (KL) divergences with bootstrapped cut-off values are used, as well as the Kruskall-Wallis test for equality of medians. Among the genes associated with the differentially expressed signatures identified by our semiparametric model, there are several important GO categories that are consistent with the existing biological knowledge about the host response to *Salmonella* infection. We conclude with a summary of the biological significance of our discoveries.

SatAM-TimeSlot-l: Session #21

Aggregation Effect and Forecasting Temporal Aggregates of Seasonal Long Memory Processes

*Kasing Man and Anna Valeva
Western Illinois University, USA
(*Paper Presenter; E-mail: ks-man@wiu.edu)

Abstract

We study the effect of time aggregation on seasonal long memory process, with d and D being the regular and seasonal long memory parameters and S the seasonal period. Upon time aggregation and as the level of aggregation tends to infinity, we argue that the Man and Tiao (2006, *Inter. J. of Forecasting*, 22, 267-281) effect applies and the aggregates can be closely approximated by an ARFIMA(0,d+D, e) structure, where e is the nearest integer just greater than d+D. In addition, when the seasonal period S is long enough and as the aggregation level approaches S, the aggregate process will preserve both long memory parameters and develop an approximate MA(d*) structure in this intermediate step. The d* is the nearest integer just greater than d. Next, we turn to forecasting and our aim is to predict the aggregate of a seasonal long memory time series. We study the efficiency loss in using the approximate aggregate model relative to using the underlying disaggregate model.

SunAM-TimeSlot-3: Session #42

How Many People Do You Know? Efficiently Estimating Personal Network Size

[1]*Tyler H. McCormick, [2]Matthew J. Salganik, and [1]Tian Zheng
[1]Columbia University-New York, USA; [2]Princeton University-New Jersey, USA
(*Paper Presenter; E-mail: tyler@stat.columbia.edu)

Abstract

In this paper we develop a method to estimate both individual social network size (i.e., degree) and the distribution of network sizes in a population by asking respondents how many people they know in specific subpopulations (e.g., people named Kevin). Building on the scale-up method of Killworth et al. (1998) and other previous attempts to estimate individual network size, we first propose a latent non-random mixing model which resolves three known problems with previous approaches. As a byproduct, our method also provides estimates of the rate of social mixing between population groups. We demonstrate the model using a sample of 1,370 adults originally collected by McCarty et al. (2001). Based on insights developed during the statistical modeling, we conclude by offering practical guidelines for the design of future surveys in this area. Most importantly, we show that if the specific subpopulations are chosen wisely, complete statistical procedures are no longer required for estimation.

Sat 11:15AM – 12Noon: Special Named Lecture-III

D. Basu Lecture

A Noninformative Bayesian Approach to Finite Population Sampling Using Auxiliary Variables

Glen Meeden
University of Minnesota-Minneapolis, USA
(E-mail: glen@stat.umn.edu)

Abstract

In finite population sampling prior information is often available in the form of partial knowledge about an auxiliary variable, for example its mean may be known. In such cases, the ratio estimator and the regression estimator are often used for estimating the population mean of the characteristic of interest. The Polya posterior has been developed as a noninformative Bayesian approach to survey sampling. It is appropriate when little or no prior information about the population is available. Here we show that it can be extended to incorporate types of partial prior information about auxiliary variables. We will see that it typically yields procedures with good frequentist properties even in some problems where standard frequentist methods are difficult to apply. Moreover one does not need to select a model which explictly relates the characteristic of interest to the auxiliary variables.

SunAM-TimeSlot-1: Session #15

Optimal Design for Nonlinear Regression Models

Viatcheslav B. Melas
St. Petersburg State University, Russia
(E-mail: v.melas@pochta.tvoe.tv)

Abstract

The talk concerns constructing and studying optimal experimental designs for nonlinear regression models. Nonlinear (in parameters) regression models are widely applied in experimental studying of problems in biology, chemistry and many other fields. Optimal design in such problems is the way to improve results under the limited budget. The main feature of such models consists of dependence of asymptotic covariance matrix on true values of model parameters. The standard ways to overcome this difficulty are locally optimal, maximin and Bayesian approaches. Constructing and studying the corresponding optimal designs proves to be a very complicated problem. Usually such problems are tried to solve by merely numerical techniques. Here we propose a development of the functional approach described in the book Melas (*Functional approach to optimal experimental design.* Lecture Notes in Statistics, vol. 184, Springer: 2006). The main idea of this approach consists of representing support points of locally optimal, maximin efficient and Bayesian designs by Taylor series. The approach is demonstrated by a few examples including the Michaelis–Menten model and its generalizations. The efficiencies of the three types of optimal designs are compared.

SatPM-TimeSlot-1: Session #44

Inference on Multi-Resolution Spatial-Temporal Process with Application

Wanli Min
IBM T. J. Watson Research Center, Yorktown Heights, USA
(E-mail: wanlimin@us.ibm.com)

Abstract

We consider a spatial stochastic process where the observed data consist of two sets of measurements: one is measured dense in time but spatially sparse, the other is measured occasionally over time at with dense spatial coverage. The goal is to predict the dependent observable at any spatial location and time. We establish an inference procedure based on mean-field approximation and kriging with drift taking into account the effect of exogenous environmental variables on the observables. The kriging model has a functional form derived from solving a partial differential equation with Dirichlet boundary conditions. The obtained model makes for a convenient tool to infer quantitatively the effect of physical variables on the process, from which an optimum setting of physical variables can be found subject to the certain administrative constraints on the process. We also propose an algorithm of sequential design to add new sampling points to existing layout. Application to a real dataset is reported in the end.

SunAM-TimeSlot-3: Session #9

On a Multivariate Bayesian Prediction Problem with Applications

Robb Muirhead
Pfizer Global Statistics-Connecticut, USA
(E-mail: robb.j.muirhead@pfizer.com)

Abstract

In this talk, we introduce a novel vague prior distribution for the parameters of a multivariate normal distribution, and discuss the resulting Bayesian predictive distribution. The prior is motivated by an “importance ordering” imposed on the variables; and the predictive distribution is motivated by a classification problem. The predictive distribution is shown to be superior, in a number of frequentist settings, to that corresponding to the standard (Jeffreys) prior. Possible clinical applications are discussed. This is joint work with Morris L. Eaton (University of Minnesota) and Eve Pickering (Pfizer).

FriPM-TimeSlot-1: Session #10

Semi Parametric Modeling and Inference of Gene Dependence Using Copulas.

[1*]Nitai D. Mukhopadhyay and [2]Sarat C. Dass
[1]Virginia Commonwealth University-Richmond, USA;
[2]Michigan State University-East Lansing, USA
(*Paper Presenter; E-mail: ndmukhopadhy@vcu.edu)

Abstract

Certain dependence patterns in gene expression data cannot be captured by standard statistical methods. For example, statistical methods designed for capturing linear trends fail to detect a multitude of complex non-linear patterns. Dependence with a temporal lag has been studied with time series tools in Mukhopadhyay and Chatterjee (2007). However, another type of dependence, namely mixture dependence, can be hard to identify with either approach. For pairs of genes that display complex dependence patterns in a population, we propose a copula mixture model that identifies sub-populations with homogeneous component dependence patterns. The dependence structures are modeled via a mixture of Gaussian copulas with minimal assumptions on the marginal distributions of the individual genes. A reversible jump MCMC algorithm is developed and used to estimate the population structure. Both simulated and real gene expression data will be used in evaluating the applicability of the method.

FriPM-TimeSlot-1: Session #29

Editorial Forum - A Panel Discussion

Nitis Mukhopadhyay
University of Connecticut-Storrs, USA
(E-mail: nitis.mukhopadhyay@uconn.edu)

Abstract

As an organizer and chair of this invited session I have gathered an illustrious group of past and present editors and associate editors of some of the leading international journals. The panel includes Makoto Aoshima (University of Tsukuba, Japan), Subir Ghosh (University of California-Riverside, USA), Joseph Glaz (University of Connecticut-Storrs, USA), and Linda Young (University of Florida-Gainesville, USA). I will moderate and take part in the discussion as needed.

The panel will discuss some of the important issues surrounding the process of publication of research papers in refereed international journals and address some of the details inherent in the process itself. The panel will devote a significant part of the session's allocated time to address questions from the floor. This session should be of interest to all conference participants of all ages.

SatPM-TimeSlot-1: Session #1

On Fixed-Width Confidence Intervals

Nitis Mukhopadhyay
University of Connecticut-Storrs, USA
(E-mail: nitis.mukhopadhyay@uconn.edu)

Abstract

Suppose one has a random sample of fixed size n from a normal distribution having some unknown mean and unknown standard deviation. Note that the random length of a customary 100(1-p)% Student's t confidence interval for the population mean may be "large" with positive probability even though a population standard deviation is "small". That is, a customary confidence interval may be (and many times, is) useless in practice.

Thus, one may want to have a confidence interval for the mean with (i) at least 100(1-p)% probability coverage and (ii) a preassigned length 2d (>0). Unfortunately, no fixed-sample-size procedure will deliver both (i) and (ii).

However, properly designed sampling strategies can solve the problem. Stein's (1945,1949) groundbreaking two-stage sampling design solves this problem exactly. We will introduce that methodology. There are more efficient sampling designs than Stein's. We will introduce some of them in the same context.

We will explore important ideas and concepts surrounding this uniquely fundamental problem in statistical inference. With the help of data analysis, we will emphasize practical aspects of some important developments since Stein's breakthrough contribution.

SatPM-TimeSlot-l: Session #25

A Bayesian Benchmarking for Small Areas

Balgobin Nandram
Worcester Polytechnic Institute-Worcester, USA
(E-mail: balnan@wpi.edu)

Abstract

When the finite population totals are estimated for individual areas, they do not necessarily add up to the known total for all areas. Benchmarking is a technique used to ensure that the totals for all areas match the grand total, and is desirable to practitioners of survey sampling. Benchmarking shifts the small area estimators to accommodate the benchmark constraint. In doing so, it can provide increased precision to the small area estimators of the finite population means or totals. We use a Bayesian approach to show how to benchmark the finite population means of small areas. We illustrate our method to estimate body mass index using data in the third National Health and Nutrition Examination Survey. Several properties of the benchmarked small area estimators are obtained using a simulation study.

SunAM-TimeSlot-2: Session #11

Statistical Challenges in Late Stage Drug Development

Kannan Natarajan
Novartis Pharmaceutical Corporation-Florham Park, USA
(E-mail: kannan.natarajan@novartis.com)

Abstract

A major challenge in drug development is faster and efficient development to get much needed drugs to our patients, especially with life threatening diseases like cancer, quicker and with good quality data. The talk will focus on current statistical issues in 3 main areas of development: assessment of optimal biologic dose with proof of concept, confirmation of benefit vs. risk, and life cycle management.

In the dose assessment phase, the use of Bayesian or adaptive methods have gained traction, particularly to utilize prior knowledge from either pre-clinical dosing and/or dosing information from patients in previous cohorts. However, most of the current statistical methods are focused on either safety or efficacy and not on the early assessment of both risk and benefit. The confirmatory phase often involves a Phase-II followed by a Phase-III. Issues at this stage range from the appropriate choice of control drug (if any), integrating information from Phase-II and Phase-III, and safety signal assessment. Post marketing surveillance for further characterization of safety has issues ranging from data collection to appropriate methods for integrating this data with that of controlled trials. Examples from oncology drug development will be provided with the current statistical methods, highlighting the issues and unmet need where appropriate.

SunAM-TimeSlot-2: Session #11

A Simulation Tool for Design of Adaptive Dose Finding Trials

[1*]Nitin Patel, [2]Jim Bolognese, [2]Inna Perevoszkaya
[1]Cytel Inc.-Cambridge, USA; [2] Merck & Co., Inc.-Rahway, USA
(*Paper Presenter; E-mail: nitin@cytel.com)

Abstract

This talk will describe CytelSim, a simulation tool for designing adaptive Ph 2 dose finding trials. The software was developed by Cytel over two years in collaboration with Merck & Co. and includes both frequentist and Bayesian methods. CytelSim is used to assess the performance of adaptive designs and to compare them with standard designs for a number of different scenarios. This talk will discuss two case studies to illustrate its use and will also include a brief demonstration of the software.

SunAM-TimeSlot-3: Session #19

Estimating Reliability in Proportional Odds Ratio Models

[1]Ramesh C. Gupta and [2*]Cheng Peng
[1]University of Maine-Orono, USA; [2]University of Southern Maine- Portland, USA
(*Paper Presenter; E-mail: cpeng@usm.maine.edu)

Abstract

In this talk, we are mainly interested in inference on the reliability coefficient, R = P(X < Y), in proportional odds models based on the new family of tilted survival functions introduced by Marshal and Olkin (1997, *Biometrika*, 84, 641-652). We also present some results on stochastic comparison between the survival distribution functions. Asymptotic and various parametric and nonparametric bootstrap confidence intervals of R and the parameter determining the stochastic ordering of X and Y will be detailed. The performance of the confidence intervals will be assessed through simulations. We also present a numerical example to illustrate the implementation of the procedure as well. Some discussions on choosing appropriate confidence intervals for practical purpose will also provided.

SatAM-TimeSlot-l: Session #37

Performance of the Empirical Bayes Estimator for Fixed Parameters

Poduri S. R. S. Rao
University of Rochester-Rochester, USA
(E-mail: raos@math.rochester.edu)

Abstract

When a set of parameters is considered to be a random sample from a normal distribution, the Stein-type estimator for an individual parameter is the same as the Empirical Bayes Estimator. James and Stein (1961) showed that the average of the mean square errors (MSEs) of the individual estimators over all the parameters is smaller than the average of the variances of the corresponding unbiased estimators. C. R. Rao and Schinozaki (1975) derived the bias and MSE of an individual estimator of this type when all the parameters are considered to be fixed, and numerically evaluated the conditions for its MSE to be smaller than the variance of the unbiased estimator. In this paper, we examine analytically and numerically the conditions to be satisfied by the dispersion of the parameter for the above requirement, and also for its squared bias to be smaller than a specified percentage of the MSE. Approximations to these conditions are also considered and evaluated. This investigation is further extended to the cases of the difference of two parameters and a contrast of the parameters.

SunAM-TimeSlot-2: Session #17

Martingale Methods for Patterns and Scan Statistics

[1*]Vladimir Pozdnyakov and [2]J. Michael Steele
[1]University of Connecticut-Storrs, USA;
[2]University of Pennsylvania-Philadelphia, USA
(*Paper Presenter; E-mail: Vladimir.Pozdnyakov@uconn.edu)

Abstract

We show how martingale techniques (both old and new) can be used to obtain otherwise hard-to-get information for the moments and distributions of waiting times for patterns in independent or Markov sequences. In particular, we show how these methods provide moments and distribution approximations for certain scan statistics. Each general problem that is considered is also illustrated with a concrete example confirming the computational tractability of the method.

FriPM-TimeSlot-2: Session #20

Maximum Spacing Estimation for Dependent Variables

*Bo Ranneby and Jun Yu
Swedish University of Agricultural Sciences–Umea, Sweden
(*Paper Presenter; E-mail: Bo.Ranneby@sekon.slu.se)

Abstract

First the motivation and idea, as described in Ranneby (1984, *Scand. J. Statist.*, 11, 93-112), behind the maximum spacing method is presented. In that paper consistency is proved for independent and identically distributed univariate stochastic variables. In Ranneby et al. (2005, *J. Statist. Plan. Inf.,* 129, 427-446) these results were extended to multivariate observations.

In the present paper different possibilities for extensions to dependent variables are discussed. For some of the possibilities the asymptotic properties are examined. Numerical illustrations are given for autoregressive and moving average processes.

SunAM-TimeSlot-3: Session #43

A Discrete Probability Problem in Chemical Bonding

[1]Fu-Chih Cheng, [2*]M. Bhaskara Rao, and [3]Subramanyam Kasala
[1]North Dakota State University-Fargo, USA; [2]University of Cincinnati-Cincinnati, USA;
[3]University of North Carolina-Wilmington, USA
(*Paper Presenter; E-mail: Marepalli.rao@uc.edu)

Abstract

There are n cards serially numbered from 1 to n. The cards are shuffled and placed in a line one after the other on top of a table with faces up. The numbers on the faces are read from left to right. If there are consecutive numbers in increasing order of magnitude the corresponding cards are merged into one. After the merger, the cards are numbered serially from one to whatever the number of cards we now have. The cards are shuffled and placed in a line one after another on top of the table with faces up. The process continues until we have only one card left. The question of interest is what the expected number of shuffles is in order to have only one card left at the end.

SatPM-TimeSlot-2: Session #26

Modeling Mercury Deposition Through Latent Space-Time Processes

[1*]Ana G. Rappold, [2]Alan E. Gelfand, and [1]David M. Holland
[1]Environmental Protection Agency-Research Triangle Park, USA;
[2]Duke University-Durham, USA
(*Paper Presenter; E-mail: rappold.ana@epa.gov)

Abstract

This paper provides a space-time process model for total wet mercury deposition. Key methodological features introduced include direct modeling of deposition rather than of expected deposition, the utilization of precipitation information (there is no deposition without precipitation) without having to construct a precipitation model, and the handling of point masses at 0 in the distributions of both precipitation and deposition. The result is a specification that enables spatial interpolation and temporal prediction of deposition as well as aggregation in space or time to see patterns and trends in deposition.

We use weekly deposition monitoring data from the NADP/MDN (National Atmospheric Deposition Program/Mercury Deposition Network) for 2003 restricted to the eastern U.S. and Canada. Our spatio-temporal hierarchical model allows us to interpolate to arbitrary locations and, hence, to an arbitrary grid,

enabling weekly deposition surfaces (with associated uncertainties) for this region. It also allows us to aggregate weekly depositions at coarser, quarterly and annual, temporal levels.

SunAM-TimeSlot-l: Session #31

Bayesian Variable Selection for Spatially-Varying Coefficient Regression: Application to Physical Activity in Prenatal Women

Brian Reich
North Carolina State University-Raleigh, USA
(E-mail: reich@stat.ncsu.edu)

Abstract

Physical activity has well-documented health benefits for cardiovascular fitness and weight control. For pregnant women, the American College of Obstetricians and Gynecologists currently recommends 30 minutes of moderate exercise on most, if not all, days; however, very few pregnant women achieve this level of activity. Epidemiologists, policy makers, and city planners are interested in whether characteristics of the physical environment in which women live and work have influence on physical activity levels during pregnancy and in the postpartum period. In this paper we study the associations between physical activity and several factors including personal characteristics, meteorological/air quality variables, and neighborhood characteristics in pregnant women in four counties of North Carolina. We simultaneously analyze six types of physical activity and investigate cross-dependencies between these activity types. Exploratory analysis suggests that the associations are different in different regions. Therefore we use a multivariate regression model with spatially-varying regression coefficients. This model includes a regression parameter for each covariate at each spatial location. For our data with many predictors, some form of dimension reduction is clearly needed. We introduce a Bayesian variable selection procedure to identify subsets of important variables. Our stochastic search algorithm determines the probabilities that each covariate's effect is null, non-null but constant across space, and spatially-varying.

SatPM-TimeSlot-l: Session #34

Objective Bayesian Analysis for a Spatial Model with Nugget Effects

[1*]Cuirong Ren, [2]Dongchu Sun, and [2]Zhuoqiong He
[1]South Dakota State University-Brookings, USA;
[2]University of Missouri-Columbia, USA
(*Paper Presenter; E-mail: cuirong.ren@sdstate.edu)

Abstract

We often need to consider geostatistical data with nugget effects. In this paper, we have systematically studied the Jeffreys priors and various reference priors, derived by both "`exact" and asymptotic marginalization. Interestingly, not all Jeffreys and reference priors yield proper posterior distributions. We have found the conditions under which the corresponding posteriors are proper. Finally, we conduct simulation study to compare the objective priors by frequentist coverage probabilities of the one-sided credible intervals.

FriPM-TimeSlot-1: Session #8

Sample Size Determination for Hierarchical Longitudinal Designs with Differential Attrition Rates

[1,2*]Anindya Roy, [2] Dulal K. Bhaumik, [2]Subhash Aryal, and [2]Robert D. Gibbons
[1]University of Maryland-Baltimore, USA; [2]Center for Health Statistics-UIC, USA
(*Paper Presenter; E-mail: anindya@math.umbc.edu)

Abstract

We consider the problem of sample size determination for three-level mixed-effects linear regression models for the analysis of clustered longitudinal data. Three-level designs are used in many areas, but in particular, multi-center randomized longitudinal clinical trials in medical or health-related research. In this case, level 1 represents measurement occasion, level 2 represents subject, and level 3 represents center. The model we consider involves random-effects of the time trends at both the subject-level and the center-level. In the most common case, we have two random-effects (constant and a single trend), at both subject and center-levels.

The approach presented here is general with respect to sampling proportions, number of groups, and attrition rates over time. In addition, we also develop a cost model, as an aid in selecting the most parsimonious of several possible competing models (i.e., different combinations of centers, subjects within centers, and measurement occasions). We derive sample size requirements (i.e., power characteristics) for a test of treatment by time interaction(s) for designs based on either subject-level or cluster-level randomization. The general methodology is illustrated using two characteristic examples.

SatPM-TimeSlot-l: Session #38

Testing the Equality of the Means of Several Groups of Count Data in the Presence of Unequal Dispersions

Krishna K. Saha
Central Connecticut State University-New Britain, USA
(E-mail: sahakrk@ccsu.edu)

Abstract

Extra-dispersion (overdispersion or underdispersion) is a common phenomenon in the biostatistical field when count data exhibit extra-variation relatively a Poisson model. This arises when the data are grouped or when the assumption of independence is violated. In this paper, procedure for testing the equality of the means of several groups of count data when extra-dispersions among the treatment groups are unequal is developed based on the adjusted count data using the concept of the design effect and size effect proposed by Rao and Scott (1999, *Statistics in Medicine*, 18, 1373 – 1385). We also obtain the score type test statistics based on the quasi-likelihoods using the mean-variance structure of the negative binomial model, and study the properties and performance characteristics of these statistics. The simulation results indicate that the statistic based on the adjusted count data, which has a very simple form and does not require the estimates of the extra-dispersion parameters, holds the best performance characteristics over the other statistics. Illustrative application of the proposed test is also presented.

SatPM-TimeSlot-2: Session #5

Uniform Asymptotics for Robust Estimators

Matias Salibian-Barrera
University of British Columbia-Vancouver, Canada
(E-mail: matias@stat.ubc.ca)

Abstract

Most asymptotic results for robust estimators rely on regularity conditions that apply to fixed distribution functions, and that are generally difficult to verify in practice. In the robustness framework, where the distribution of the data remains largely unspecified, results that hold uniformly over a set of plausible distribution functions are of both theoretical and practical interest. Furthermore, it is also desirable to be able to determine the size of this set of distribution functions where the uniform properties hold. In this talk, I will briefly review existing uniform asymptotic results for robust estimators, and present some new results for S- and MM-estimators for linear regression.

SunAM-TimeSlot-l: Session #30

Bayesian Analysis of Microscale Spatial Variations

*Huiyan Sang and Alan E. Gelfand
Duke University-Durham, USA
(*Paper Presenter; E-mail: hs37@duke.edu)

Abstract

Microscale spatial variability is defined as the variability at distances smaller than the smallest interlocation distance in the spatial data set. In practice, little work has been done to study microscale spatial process. In this talk, I am going to present a Bayesian spatial modeling approach to account for both large scale and microscale spatial patterns. Since microscale spatial variation study often involves data observed at high resolution, I will discuss several potential solutions to tackle computational difficulty of 'large n' problem.

Fri 11:00AM – 11:45AM: Special Named Lecture-II

P. V. Sukhatme Lecture

Some New Results on Controlling False Discoveries in Multiple Testing

Sanat K. Sarkar
Temple University-Philadelphia, USA
(E-mail: sanat@temple.edu)

Abstract

Simultaneous testing of a large number of hypotheses has become a common statistical problem in many modern scientific investigations, such as in genomics and brain imaging. Often in these applications, as thousands of hypotheses are being tested, some false rejections, at most k-1, for some fixed k, can be tolerated. So, for traditional procedures that control at least one false discovery, the ability to detect false null hypotheses can potentially be improved by generalizing them to procedures that control at least k false discoveries. The concepts of FWER (familywise error rate) and FDR (false discovery rate) have recently been generalized to the k-FWER and k-FDR respectively. This talk will present the developments of these new measures and multiple testing methods that control them.

FriPM-TimeSlot-l: Session #22

Properties of Graphical Estimators from Q-Q Plots

Ananda Sen
University of Michigan-Ann Arbor, USA
(E-mail: anandas@umich.edu)

Abstract

Probability plots are popular graphical tools used by reliability engineers and other practitioners for assessing parametric distributional assumptions. They are particularly well suited for location-scale families or those that can be transformed to such families. When the plot indicates conformity to the assumed family, it is reasonable to estimate the underlying location and scale parameters from the least-squares line fitted through the plot. In this talk I shall present the properties of such (generalized) least-squares estimators with multiply right-censored data and compare their performance with those of maximum likelihood estimators. Large-sample results on consistency, asymptotic normality, and asymptotic variance expressions are obtained. Small-sample properties are studied through simulation for selected distributions and censoring patterns.

SunAM-TimeSlot-l: Session #32

Jumps and Microstructure Noise in Stock Price Volatility:

An FDA Approach

[1*]Rituparna Sen, [1]Hans-Georg Muller and [2]Ulrich Stadtmuller
[1]University of California-Davis, USA; [2]Universitat Ulm, Germany
(*Paper Presenter; E-mail: rsen@wald.ucdavis.edu)

Abstract

An important component in the Black-Scholes model for stock price process is volatility. It is necessary to estimate volatility in many practical applications like option pricing, portfolio selection and risk management. Now-a-days stock price data is available at very high frequency and the most common estimator of volatility using such data is the realized variance. However in the presence of microstructure noise, realized variance diverges to infinity. I shall describe a principal component analysis of functional data approach to handle this problem. I shall then introduce the concept of jumps in asset prices and show some real data examples on how the FDA approach can be used to detect days on which the asset price has jumps and to measure the size of jumps. Thus we can separate the jump component from the integrated volatility in the quadratic variation process. This separation leads to better prediction of integrated volatility. We develop the theory and present simulation as well as real data examples.

SatAM-TimeSlot-1: Session #7

Some Issues in the Analysis of Array CGH Data

[1*]Venkatraman E. Seshan and [2]Adam Olshen
[1]Columbia University-New York, USA; [2]MSKCC-New York, USA
(*Paper Presenter; E-mail: ves2111@columbia.edu)

Abstract

DNA sequence copy number is the number of copies of DNA at a region of a genome. The development of malignant tumors and their progression often involve alterations in DNA copy number. We will present the motivation for the Circular Binary Segmentation algorithm we developed (Olshen et al., *Biostatistics*, 2004) to segment the genome into regions of equal copy number. We will also present refinements to the algorithm to handle the large arrays that are being used more commonly now (Venkatraman & Olshen, *Bioinformatics*, 2007). We will present extensions to the problem such as parental copy numbers and the application to tumor data.

Fri 9:15AM – 10:15AM

Plenary Lecture-I

New Insights into Dirichlet Priors and Partition Based Priors with Applications to Reliability and Censoring

Jayaram Sethuraman
Florida State University-Tallahassee, USA
(E-mail: sethu@stat.fsu.edu)

Abstract

Several proofs that the mean of a random probability measure is finite with probability one when the random probability measure has a Dirichlet distribution if and only if $\log(1 + |x|)$ is integrable with respect to the parameter of the Dirichlet distribution, have been given in the literature invoking results from several areas of probability. We will show that this result follows easily from our constructive definition of a Dirichlet measure.

We define a new class of prior distributions, called Partition Based prior distributions, that can be used in nonparametric problems. We show applications of these priors in problems relating to failure models in reliability and in problems relating to general censoring.

FriPM-TimeSlot-1: Session #10

A Combined Test for Genetic Association that Incorporates Information about Hardy Weinberg Disequilibrium in Cases

Sanjay Shete
University of Texas M. D. Anderson Cancer Center-Houston, USA
(E-mail: sshete@mdanderson.org)

Abstract

To assess genetic association between single nucleotide polymorphisms (SNPs) and diseases status, typically either logistic regression model or general linear model is employed for testing associations. However, deviation from Hardy-Weinberg equilibrium could be another approach for genetic association study. The Hardy-Weinberg equilibrium (HWE) is one of important principles in population genetics. Deviation from Hardy-Weinberg equilibrium, among case (patients) group, may provide additional evidence for association between SNPs and diseases. Our purpose here is to combine evidence from Hardy-Weinberg departures in case subjects and standard regression approaches that use case and control subjects, so that a more powerful statistical test would be developed for genetic association study. In this paper, we propose two new approaches to combine such information. One approach is using the mean based tail strength measure and another is the median based tail strength measure to integrate logistic regression and Hardy-Weinberg equilibrium test, to study the association between a binary disease outcome and a SNP on the basis of case-control data. For both mean based and median based tail strength measures, we derived exact formulas to compute p-values. We also developed an approach to obtain empirical p-values values using a re-sampling procedure. Results from simulation studies and real diseases studies demonstrate that the new approach is more powerful than the traditional logistic regression model. And the type I error probabilities of our approach were also well controlled.

SatAM-TimeSlot-l: Session #37

Role of Information Measures in Bayesian Sequential Estimation

Ramkaran Singh
University of Lucknow, India
(E-mail: ramkarans@hotmail.com)

Abstract

Due to the independence of stopping and terminal estimation rules in sequential Bayesian decision theory, the problem of Bayesian sequential estimation reduces to the problem of finding (i) the optimal stopping rule, and (ii) the optimal fixed size sample Bayes estimator for a given loss function. The optimal stopping boundaries are obtained using Bellman's principle of optimality. However, this backward recursion procedure involves enormous computing. To circumvent this difficulty, many authors have suggested asymptotic stopping rules. Some sub-optimal stopping rules have also been proposed. It has been found that performance of sub-optimal stopping rules is quite satisfactory in majority of the cases. In the present paper, we propose to obtain stopping boundaries using various information measures. We have used two types of information measures to obtain the stopping boundaries. The first type uses the conjugate prior and Kullback-Leibler information measure to obtain stopping boundaries and in the second method stopping boundaries are based on a measure of information which depends on the expected change in the incurred losses due to the sample information. We have obtained the stopping boundaries and terminal estimation rule for the shape parameter of the Pareto distribution.

SatAM-TimeSlot-l: Session #28

Association Models for Clustered Data with Binary and Continuous Responses

[1*]Debajyoti Sinha, [1]Lanjia Lin, [2]Stuart R. Lipsitz, and [3]Dipankar Bandyopadhyay
[1]Florida State University-Tallahassee, USA; [2]Harvard Medical School-Boston, USA;

[3]Medical University of South Carolina-Charleston, USA
(*Paper Presenter; E-mail: sinhad@stat.fsu.edu)

Abstract

In this article, we propose a joint model for clustered bivariate responses data containing both binary and continuous outcomes for each subject within each cluster. We use cluster-specific random effects to address the within-subject as well as within-cluster associations. The marginal functional form of the binary response probability integrated over all the random effects preserves the logistic form. Finally, estimators of parameters of our model applied to data from development toxicity study are obtained using the maximum likelihood and Bayesian methods. We illustrate the advantages of our new model and comparisons between different estimation approaches via the analysis of a toxicity study.

SatAM-TimeSlot-l: Session #28

Semiparametric Bayesian Analysis of Nutritional Epidemiology Data in Presence of Measurement Error

[1]Samiran Sinha, [1]Raymond J. Carroll, [1]Bani K. Mallick, and [2]Victor Kipnis
[1]Texas A&M University-College Station, USA;
[2]National Cancer Institute-Bethesda, USA
(E-mail: sinha@stat.tamu.edu)

Abstract

The present paper proposes a semiparametric Bayesian method for handling measurement error in nutritional epidemiological data. The goal of this research is to estimate the nonparametric form of association between a disease and exposure variable while the true values of the exposure are never observed. Motivated by a nutritional epidemiological data we consider the setting where a surrogate covariate is recorded in the primary data, and a calibration data contain information on the surrogate variable and repeated measurements of an instrumental variable of the true exposure. We develop a completely flexible Bayesian method with feasible computational tool. We apply the proposed method on the motivating data set from NIH-AARP diet and health study. Finally, a small-scale simulation study is performed to assess the performance of the proposed method.

SunAM-TimeSlot-l: Session #33

On Approximate Optimality of the Unbalanced Sequential Procedure for the Partition Problem

[1*] Tumulesh K. S. Solanky and [2] Yuefeng Wu
[1] University of New Orleans-Louisiana, USA;
[2] North Carolina State University-Raleigh, USA
(*Paper Presenter; E-mail: tsolanky@uno.edu)

Abstract

We consider the problem of partitioning a set of normal populations with respect to a control population into two disjoint subsets according to their unknown means. Taking c observations from the control population instead of the usual *vector-at-a-time* approach, we consider an approximate optimality of the purely sequential procedure. Using the Monte Carlo simulation techniques, the small and moderate sample size performance of the proposed procedures is studied and the derived optimality is verified.

FriPM-TimeSlot-2: Session #24

Model Checking in Partial Linear Regression Models with Berkson Measurement Errors

[1]Hira L. Koul and [2*]Weixing Song

[1]Michigan State University-East Lansing, USA;
[2]Kansas State University-Manhattan, USA
(*Paper Presenter; E-mail: weixing@ksu.edu)

Abstract

This talk will discuss the problem of fitting a parametric model to the nonparametric component in partially linear regression models when both covariates in the parametric and nonparametric parts are subject to the Berkson measurement error. The proposed test is based on the supremum of a martingale transform of a certain marked empirical process of calibrated residuals. Asymptotic null distribution of this transformed process is shown to be the same as that of standard Brownian motion. Consistency of this sequence of tests at some fixed alternatives and asymptotic power under some local nonparametric alternatives are also discussed. Simulation studies are conducted to assess the performance of the proposed test. A Monte Carlo study shows the dominance of the power of the proposed test over some of the existing tests at the chosen alternatives.

FriPM-TimeSlot-2: Session #24

Statistical Analysis of Survival-Sacrifice Data

Brandon H. Greene and *Winfried Stute
University of Giessen, Germany
(*Paper Presenter; E-mail: Winfried.Stute@math.uni-giessen.de)

Abstract

When analyzing lifetime data a major difficulty is caused by the fact that due to time limitations placed on a study or follow-up-losses, information may only be available in incomplete form. Undoubtedly, the best-studied example is the right-censorship case handled by Kaplan and Meier. A more complicated situation arises when, for example in an animal experiment, the goal is not to analyze the lifetime, but the time elapsed from the exposition to some risk until the onset of disease. However, the onset of the disease remains unobservable unless the animal is sacrificed. In this work we discuss various statistical issues related with such Survival-Sacrifice data.

SunAM-TimeSlot-l: Session #35

The Formal Definition of Reference Priors

[1]James O. Berger, [2]Jose M. Bernardo, and [3*]Dongchu Sun
[1] Duke University-Durham, USA; [2] Universitat de Valencia-Valencia, Spain;
[3] University of Missouri-Columbia, USA
(*Paper Presenter; E-mail: sund@missouri.edu)

Abstract

Reference analysis produces objective Bayesian inference, in the sense that inferential statements depend only on the assumed model and the available data, and the prior distribution used to make an inference is least informative in a certain information-theoretic sense. Reference priors have been rigorously defined in specific contexts and heuristically defined in general, but a rigorous general definition has been lacking. We produce a rigorous general definition here, and then show how an explicit expression for the reference prior can be obtained under very weak regularity conditions. The explicit expression can be used to derive new reference priors both analytically and numerically.

FriPM-TimeSlot-1: Session #8

Statistical Modeling of Human Conception

*Rajeshwari Sundaram, Sung Duk Kim, Kirsten J. Lum, and Germaine Buck Louis
Eunice Kennedy Shriver National Institute of Child Health and Human Development, USA
(*Paper Presenter; E-mail: sundaramr2@mail.nih.gov)

Abstract

Reproductive scientists are interested in constructing biologically valid statistical models for predicting the probability of conception. Two differing approaches for studying the underlying probability of conception have been investigated extensively in literature. In one approach various models for the underlying distribution of time to conception (a discrete survival time) has been proposed by Weinberg et al. (1986) and Scheike et al. (1997). Alternatively, Dunson et al. (1999,2001,2002,2005) have extensively studied models for conception in a cycle using day-specific conception probabilities. In both approaches, an important predictor is the longitudinally observed couple-specific behavior, which also has common risk factors with the underlying event of interest, namely conception. In this talk, we propose joint modeling of the longitudinal human behavior and probability of conception for both modeling approaches mentioned above. This work is motivated by the LIFE study being conducted by NIH.

FriPM-TimeSlot-2: Session #24

Generalized Information Criterion

[1*]Masanobu Taniguchi and [2]Junichi Hirukawa
[1]Waseda University, Japan; [2]Niigata University, Japan
(*Paper Presenter; E-mail: taniguchi@waseda.jp)

Abstract

Here we propose a generalized AIC (GAIC), which includes the usual AIC as a special case, for general class of stochastic models (that is, i.i.d., non-i.i.d., time series models, etc.). Then we derive the asymptotic distribution of selected order by GAIC, and show that the selected order is inconsistent, that is, it does not converge the true one in probability. In actual statistical analysis, it is natural to suppose that the true model would be contiguously perturbed. Under this setting we derive the asymptotic distribution of the selected order by GAIC. Then, in this situation, we will elucidate the essential features of GAIC, BIC and H-Q criteria. Also numerical studies will be given to confirm the results.

FriPM-TimeSlot-2: Session #20

Characteristic Function Estimation of Non-Gaussian Ornstein-Uhlenbeck Processes

[1*]Emanuele Taufer, and [2]Nikolai N. Leonenko
[1]University of Trento, Italy; [2]University of Cardiff, UK
(*Paper Presenter; E-mail: emanuele.taufer@unitn.it)

Abstract

Continuous non-Gaussian stationary processes of the Ornstein-Uhlenbeck (OU)-type with self-decomposable marginal distributions are becoming increasingly popular given their flexibility in modelling stylized features of financial series such as asymmetry, heavy tails and jumps.

Efficient estimation of OU processes in the non-Gaussian case maybe quite cumbersome to implement. Although in theory, a likelihood function could be constructed by exploiting the Markov property or independence of increments, in practice explicit or tractable expressions for the relevant densities are rarely available. One way out is to resort to simulation based techniques; the problem with these methods is that simulation of OU, and more generally, Lévy processes, is difficult owing to their jump character.

The present work proposes a characteristic function (ch.f.)-based estimation approach. A unifying criterion is introduced by exploiting the peculiarity of self-decomposable random variables and by providing rules to construct estimators starting from a univariate ch.f., which is available in explicit form for several models of interest. The techniques proposed are quite general as they are not restricted to processes with positive jumps.

The relevant asymptotic theory is presented and the applicability of estimators is shown by way of simulations and an application. An extension to OU-based stochastic volatility models is provided.

SatAM-TimeSlot-l: Session #21

A Rough Set based Approach to Detect Plagiarism

[1*] K. Thammi Reddy, [2] M. Shashi, and [3] L. Pratap Reddy
[1]VIIT-Visakhapatnam, India; [2]Andhra University-Visakhapatnam, India; [3]JNTU-Hyderabad, India
(*Paper Presenter; E-mail: thammireddy@yahoo.com)

Abstract

Information Retrieval systems used in digital libraries need to find the extent of similarity between a pair of text documents for providing access to topically relevant documents on one hand and for identifying document replication on the other hand. In this paper the details of a Rough Set based Document Ranking system (RSDRS) developed to detect plagiarism by the authors are presented. Plagiarism is the practice of claiming, or implying, original authorship or incorporating material from someone else's written or creative work, in whole or in part, into one's own without adequate acknowledgement. Unlike cases of forgery, in which the authenticity of the writing, document, or some other kind of object, itself is in question, plagiarism is concerned with the issue of false attribution (www.wikipedia.org).

The terms associated with related concepts are grouped together to form equivalence classes by clustering the terms of the vocabulary. The query passage and the documents are represented as rough sets using these equivalence classes of terms and further partitioned into families of rough sets in higher level approximation spaces which impose partial ordering on the families of documents with reference to the query passage. Documents falling in the same family are ordered in accordance with their similarity to the query to form the relevance ranking of the documents. The performance of the RSDRS is analyzed using Average Precision metric by computing the precision at different positions in the ranking list.

FriPM-TimeSlot-2: Session #40

Asymptotic Properties of Functions of Order Statistics

[1]G. Jogesh Babu, [2]Zhidong Bai, [2]Kwok P. Choi, and [3*]Vasudevan Mangalam
[1]Pennsylvania State University-University Park, USA;
[2]National University of Singapore- Singapore;
[3]Universiti Brunei Darussalam-Brunei
(*Paper Presenter; E-mail: mangalam@fos.ubd.edu.bn)

Abstract

We have a sequence of d-dimensional random vectors such that the component sequences are i.i.d. for each component. We also have a real-valued measurable function on the d-dimensional space. Under some mild regularity conditions on the function, we find the almost sure limit for the average of the function evaluated at the order statistics, defined componentwise. The asymptotic normality of the average is also established.

It is shown that the almost sure convergence does not require the random vectors to be i.i.d. It is also shown that the limit depends only on the marginal distributions and not on the joint distribution of the components of the random vector. Asymptotic normality is proved under strengthened conditions, which include the condition that the random vectors are i.i.d.

SatPM-TimeSlot-1: Session #44

Assessment of Spatial Point Process Models by Thinning and Augmenting

Alejandro Veen
IBM T. J. Watson Research Center-Yorktown Heights, USA
(E-mail: aveen@us.ibm.com)

Abstract

This work discusses how thinning and augmenting point patterns can be used to assess the goodness-of-fit of spatial point process models. After presenting available test statistics and their distributional properties, the methodology will be applied to point patterns representing the geographic locations of settlements.

SatAM-TimeSlot-l: Session #39

The R Package Implementing Maximum Entropy Bootstrap for Dependent Time Series

[1*]Hrishikesh D. Vinod and [2]Javier López-de-Lacalle
[1]Fordham University-New York, USA;
[2]University of the Basque Country-Bilbao, Spain
(*Paper Presenter; E-mail: Vinod@fordham.edu)

Efron's bootstrap shuffle destroys time series dependence properties of the original data, whereas block bootstrap is available only for stationary m-dependent data. Only Vinod's (2004, *J. Empirical Finance*, 11, 353–377) the maximum entropy bootstrap algorithm creates an ensemble for statistical inference when the data are state-dependent evolving time series. Stationarity is not required and the ensemble satisfies the ergodic theorem and the central limit theorem. The meboot R-package implements such algorithm. This document introduces the procedure explaining the rationale behind it, and illustrates its scope by means of a broad range of applications including several guided applications.

FriPM-TimeSlot-2: Session #36

REU (Research Experience for Undergraduates) in Statistics (SUMSRI) at Miami University

Vasant Waikar
Miami University-Ohio, USA
(E-mail: waikarvb@muohio.edu)

Abstract

In this paper I will describe the working of this REU named the Summer Undergraduate Mathematical Sciences Research Institute or SUMSRI that I have directed for the last nine summers at Miami University. SUMSRI is funded by the National Security Agency (NSA) and the National Science Foundation (NSF). I will also discuss the nature and content of the research papers written by the undergraduates at this REU under my supervision. Some of these papers have won awards in the student paper competition sponsored by the American Statistical Association (ASA).

SunAM-TimeSlot-2: Session #45

An Efficient Nonparametric Estimator for Judgment Post-Stratified Data

[1*]Xinlei (Sherry) Wang, [2]Johan Lim, and [1]Lynne Stokes
[1]Southern Methodist University, USA;
[2]Seoul National University, Korea
(*Paper Presenter; E-mail: swang@smu.edu)

Abstract

MacEachern et al. (2004) introduced a data collection method, called judgment post-stratification (JP-S), based on ideas similar to those in ranked set sampling, and proposed methods for mean estimation from JP-S samples. In this paper, we propose an improvement on their methods, which exploits the fact that the distributions of the judgment post-strata are often stochastically ordered, so as to form a mean estimator using isotonic sample means of the post-strata. This new estimator is strongly consistent with asymptotic properties similar to those in MacEachern et al. (2004). It is shown to be more efficient for small sample sizes, which appears to be attractive in applications requiring cost efficiency. Further, we extend our method to JP-S samples with imprecise ranking or multiple rankers. The performance of the proposed estimators is examined on three data examples through simulation.

SunAM-TimeSlot-l: Session #32

High Dimensional Volatility Modeling and Analysis for High-Frequency Financial Data

Yazhen Wang
National Science Foundation and University of Connecticut-Storrs, USA
(E-mail: yzwang@stat.uconn.edu)

Abstract

It is very popular in financial economics to estimate volatilities of asset returns by realized volatility based on high-frequency financial data. For a large number of assets, due to matrix size the volatility matrices are very hard to estimate by realized co-volatility matrices at the reasonable level of accuracy. Dimension reduction is needed to reduce the effective size of volatility matrices and produce better volatility estimator. As high-frequency financial data for multi-assets are non-synchronized, the usual factor models are not directly applicable to high-frequency return data. In this paper we propose a factor model directly for volatility matrices to reduce the effective size of volatility matrices and a vector autoregressive (VAR) model for the loading factors to estimate model parameters and predict volatility. The factor and VAR models are at low-frequency level. We estimate realized co-volatility matrices from high-frequency data and then fit the estimated co-volatility matrices to the low-frequency models for parameter estimation.

Sat 11:15AM – 12Noon: Special Named Lecture-IV

V. S. Huzurbazar Lecture

Model Selection based on Prediction Precision

L. J. Wei
Harvard University-Cambridge, USA
(E-mail: wei@hsph.harvard.edu)

Abstract

Consider a standard model selection or evaluation problem with two competing working models. Suppose that the second model contains predictors which may be expensive or invasive to obtain over the first model. The question is how to quantify the difference of these two models so the practitioners can make a cost-benefit decision. The usual likelihood ratio type of statistics may not give us a heuristically interpretable way to assess the benefit from the complex model over the simple one. Moreover, any estimated difference should be evaluated via the corresponding sampling variation, that is, the observed difference may be well within the sampling error, indicating the gain from the complex model may not be real. In this talk we discuss a simple approach to quantify such model differences. The method is illustrated with continuous, binary and event time response for the regression (working) models.

FriPM-TimeSlot-1: Session #10

A Systematic Assessment of Population Structure and Its Effects on Genome-Wide Association Studies

*Hongyan Xu and Varghese George
Medical College of Georgia-Augusta, USA
(*Paper Presenter; E-mail: hxu@mcg.edu)

Abstract

Large-scale genome-wide association studies are promising for unraveling the genetic basis of complex diseases. Population structure is a potential problem, the effects of which on genetic association studies are controversial. Systematic assessment of the effects of population structure on large-scale genetic association studies is needed for valid analysis of the data and correct interpretation of the results. In this study, we performed extensive coalescent-based simulations of samples with varying levels of population structure to investigate the effects of population structure on large-scale genetic association studies. The effects of population structure are measured by the multiplicative changes of the probability of type I error rate, which is then correlated with the levels of population structure. It is found that at each nominal level of association tests, there is a positive relationship between the level of population structure and its effects, which could be summarized well with a regression function. It is also found that at a specific level of population structure, its effect on association study increases drastically as the significance level of the test decreases. Therefore in genome-wide association studies, the effects of population structure cannot be safely ignored and must be accounted for with proper methods.

SatPM-TimeSlot-1: Session #4

A Bayesian Modeling in Dual Response Surface Methodology

[1] Younan Chen and [2*] Keying Ye
[1] Discovery Biometrics Biopharm, GlaxoSmithKline, USA
[2] University of Texas at San Antonio, USA
(*Paper Presenter; E-mail: keying.ye@utsa.edu)

Abstract

In quality engineering, dual response surface methodology is a useful tool to model an industrial process by using both the mean and the standard deviation of the measurements as the responses. The least squares method in regression is often used to estimate the coefficients in the mean and standard deviation models, and various decision criteria are proposed by researchers to find the optimal conditions. Based on the inherent hierarchical structure of the dual response problems, we propose a Bayesian hierarchical approach to model dual response surfaces. Such an approach is compared with two frequentist least squares methods by using real data sets and simulated data.

SatPM-TimeSlot-2: Session #5

Median Loss Analysis and Median Cross Validation

*Chi W. Yu and Bertrand Clarke
University of British Columbia-Vancouver, Canada
(*Paper Presenter; E-mail: c.yu@stat.ubc.ca)

Abstract

In this talk, we justify using the median of the loss, in place of the expected loss in the conventional statistical decision theory. The resulting estimates can be derived in many standard examples. The estimates yield improved predictive performance, and have a high resistance to outliers and to the specific loss used to form them. For applications, we systematically replace the expectation in the usual methods with the median. In particular, we consider a median-based cross-validation under squared error loss for model selection, and a simulation study is conducted to verify that the median-based cross validation gives better results than the usual expectation-based approach outside the normal error case.

Sat 3:15PM – 4:15PM

Plenary Lecture-II

So I Missed My Bus – What Has This Got To Do With Anything?

Marvin Zelen
Harvard University-Cambridge, USA
(E-mail: zelen@hsph.harvard.edu)

Abstract

This lecture will discuss the idea of length biased sampling. Several examples will be discussed that arise in real life situations. The motivating example is that I always seem to be missing my bus. Do I have a perpetual "jinx" or is there a rational explanation for my apparent misfortunes. This example introduces the idea of length biased sampling. It arises in many applications of modeling in the biomedical sciences. Among the applications of length biased sampling which will be discussed are: (1) the early detection of disease and (2) the designation of the risk status for families for certain chronic diseases.

SatAM-TimeSlot-1: Session #7

Hierarchical Models for Signal Transduction Pathway Analysis from Single Cell Measurements

Ruiyan Luo and *Hongyu Zhao
Yale University-New Haven, USA
(*Paper Presenter; E-mail: hongyu.zhao@yale.edu)

Abstract

Recent technological advances have made it possible to simultaneously measure multiple protein activities at the single cell level. In contrast to measurements based on aggregated cells, e.g. gene expression analysis from microarrays, single cell-based measures provide much richer information on the cell states and signaling networks. In this presentation, we discuss a hierarchical model for signaling network reconstruction based on single cell measurements. This modeling framework can effectively pool information from different perturbation experiments and the network sparsity is also explicitly modeled. We will describe the Monte Carlo Markov Chain method for model inference. Simulation results demonstrate the superiority of the hierarchical approach. The usefulness of our model will also be illustrated through its application to the intracellular signaling networks of human primary naïve CD4$^+$ T cells, downstream of CD3, CD28, and LFA-1 activation.

SatPM-TimeSlot-l: Session #38

Omnibus Tests for Comparison of Competing Risks with Covariate Effects via Additive Risk Model

*Yichuan Zhao and Duytrac Nguyen
Georgia State University-Atlanta, USA
(*Paper Presenter; E-mail: matyiz@langate.gsu.edu)

Abstract

It is of interest that researchers study competing risks in which subjects may fail from any one of K causes. Comparing any two competing risks with covariate effects is very important in medical studies. In this paper, we develop omnibus tests for comparing cause-specific hazard rates and cumulative incidence functions at specified covariate levels. The omnibus tests are derived under the additive risk model by a weighted difference of estimates of cumulative cause-specific hazard rates. Simultaneous confidence bands for the difference of two conditional cumulative incidence functions are also constructed. A simulation procedure is used to sample from the null distribution of the test process in which the graphical and numerical techniques are used to detect the significant difference in the risks. In addition, we conduct a simulation study, and the simulation result shows that the proposed procedure has a good finite sample performance. A melanoma data set in clinical trial is used for the purpose of illustration.