

1. An Inequality for Principal Components and Regression

Andreas Artemiou, The Pennsylvania State University

Artemiou and Li (to appear: *Statistica Sinica*) gave a probabilistic explanation of a natural phenomenon that is frequently observed but whose reason is not well understood. That is, in a regression setting, the response (Y) is often highly correlated with the leading principal components of the predictor (X) even though there seems no logical reason for this connection. In this work we make three extensions of the previous result. First, we extend the result in a nonlinear regression setting, where the link function is unknown. Second, in both the linear and nonlinear setting we extend the result to multivariate response Y and third in both linear and nonlinear settings we show the result for fixed covariance matrices.

2. Inferring Likelihoods and Climate System Characteristics from Climate Models and Multiple Tracers

K Sham Bhat, The Pennsylvania State University

Abstract: To understand the current state of the climate system and to predict future behavior, good estimates of key climate system parameters are critical. Due to the difficulty in measuring these parameters directly, we must infer their values based on two sources: (i) spatiotemporal observations of 'tracers' that indirectly provide information about these parameters, and (ii) output from computationally expensive climate models run at several climate parameter settings. Here we describe an inferential approach using Gaussian processes to emulate the climate models, thereby establishing a connection between the climate parameters and the multiple tracers. We carry out statistical inference for the climate parameters, accounting for dependence and sources of uncertainty. We also discuss identifiability issues and computational challenges posed by the size of the data.

3. A Hierarchical Bayesian Dynamic Model: The Case of Customer Satisfaction and Shareholder Value Association

Zhe Chen, The Pennsylvania State University

Hierarchical Bayesian Model has been widely used in the Marketing literature to address various marketing phenomena. In this study, the authors propose a Hierarchical Bayesian Dynamic model to give insights upon the association between customer satisfaction and shareholder value. Traditionally, marketing research has indicated a positive association; however, the authors believe that the association is more complex and the complexity stems from changing environmental conditions as well as individual firms differences. The proposed model allows estimation of time varying associations between customer satisfaction and shareholder value at individual firm level. Their findings suggest that individual firm level associations change over time and customer satisfaction tends to create shareholder value in the long run but not necessarily in the short term. The results also indicate that industry and firm characteristics can affect the associations and the impacts vary over time.

4. A Multivariate Likelihood-Tuned Density Estimator and Modal Inference

Yejin Chung and Bruce G. Lindsay, The Pennsylvania State University

We consider an improved multivariate nonparametric density estimator which arises from treating the kernel density estimator as an element of the model that consists of all mixtures of the kernel, continuous or discrete. One can obtain the kernel density estimator with "likelihood-tuning" by using the uniform density as the starting value in an EM algorithm. The second tuning leads to a fitted density with higher likelihood than the kernel density estimator. The two-step likelihood-tuned density estimator reduces asymptotic bias. The new density estimator captures interesting signals better than the others but does not have clear improvement in root MSE. We compare the performance of the new density estimator with other modified density estimators in higher dimensions.

5. Using a Stepping Algorithm to Fit an Exponential Random Graph Model for a Biological Network

Ruth M Hummel, Mark S Handcock (Univ. Washington), David R Hunter

Because of the intractability of directly calculating the maximum likelihood estimate in many exponential family random graph models, a Markov chain Monte Carlo method can be used for obtaining approximate maximum likelihood estimates. This approximation to the likelihood ratio relies heavily on the starting value being close to the MLE, while providing no intuition into how we might choose this starting value. In this poster we describe the difficulties surrounding maximum likelihood estimation in ERG Models, and we introduce as a promising improvement in MCMC ML estimation an iterative method of moving toward the MLE by jumping between the original parameter space and a reparametrized space in which we know the value of the MLE. We illustrate this method on a biological network model for which the MCMC MLE was previously not calculable.

6. Local Rank Inference for Varying Coefficient Models

Bo Kai, The Pennsylvania State University

By allowing the regression coefficients to change with certain covariates, the class of varying coefficient models offers a flexible semiparametric approach to modeling nonlinearity and interactions between covariates. We propose a novel estimation procedure for the varying coefficient models based on local ranks. The new procedure provides a highly efficient and robust alternative to the local linear least squares method, and can be conveniently implemented using existing R software package. Theoretical analysis and numerical simulations both reveal that the gain of the local rank estimator over the local linear least squares estimator, measured by the asymptotic mean squared error or the asymptotic mean integrated squared error, can be substantial. The new estimator may achieve the nonparametric convergence rate even when the local linear least squares method fails due to infinite random error variance. We establish the large sample theory of the proposed procedure by utilizing results from generalized U-statistics, whose kernel function may depend on the sample size. We also extend a resampling approach, which perturbs the objective function repeatedly, to the generalized U-statistics setting; and demonstrate that it can accurately estimate the asymptotic covariance matrix. This is a joint work with Lan Wang and Runze Li.

7. Recent History Functional Linear Model for Longitudinal Data

Kion Kim, The Pennsylvania State University

We propose a variant of historical functional linear models for cases where the current response is affected by the predictor process in a window into the past. Different from the rectangular support of functional linear models, the triangular support of the historical functional linear models and the point-wise support of the varying coefficient models, the current model has a sliding window support into the past. This idea leads to models that bridge the gap between varying coefficient models and functional linear (historic) models. We propose an algorithm for this model that can be applied to longitudinal data where the measurements are taken on irregular time points and missing values are allowed. The proposed estimation algorithm is shown to be fast, involving one dimensional basis expansions and one dimensional smoothing procedures.

8. Order Thresholding

Min Hee Kim, The Pennsylvania State University

A new thresholding method, based on L-statistics and called order thresholding, is proposed as a technique for improving the power when testing against high-dimensional alternatives. The new method allows great flexibility in the choice of the threshold parameter. This results in improved power over the soft and hard thresholding methods. Moreover, order thresholding is not restricted to the normal distribution. The performance of the basic order threshold statistic is evaluated with extensive simulations.

9. On Dimension-Folding of Matrix- or Array-Valued Statistical Objects

Bing Li, Min Kyung Kim, Naomi Altman, The Pennsylvania State University

This talk is concerned with a dimension reduction problem where the predictors are in the form of matrices or arrays. A data set that has motivated this research is the study of EEG correlates of genetic predisposition to alcoholism. This study involves two groups of subjects: an alcoholic group of 77 subjects and a control group of 45 subjects. The dimension of X , however, is almost 16,000, far exceeding the sample size n , which is 122. This type of data sets presents two challenges - that they are very large and that they have are matrices or arrays. We introduce a "dimension folding" method to handle this type of data. Applying dimension folding to the EEG data set we are able to correctly identify 86 out 122 alcoholic or nonalcoholic subjects based on their EEG patterns.

10. Estimation in Covariate-Adjusted Nonlinear Regression Models

Esra Kurum, The Pennsylvania State University

We propose a new estimation procedure for covariate adjusted nonlinear regression models for situations where both the predictors and response in a nonlinear regression model are not directly observed, however distorted versions of the predictors and response are observed. The distorted versions are assumed to be contaminated with a multiplicative factor that is determined by the value of an unknown function of an observable covariate. We demonstrate how the regression coefficients can be estimated by establishing a connection to nonlinear varying coefficient models. Simulation studies are used to illustrate the efficacy of the proposed estimation algorithm.

11. Variable Selection for Clustering

Hyang Min Lee, The Pennsylvania State University

A new variable selection algorithm is developed to achieve good separation between clusters. In contrast to the conventional measure of separation by the ratio of between- and within-cluster dispersion, we exploit the prominent geometric features of the density function so that the exact shape and orientation of the density matter. The computational foundation for this separability measure consists of the recently developed Modal EM (MEM) algorithm which solves modes of a density in the form of a Gaussian mixture, and the Ridgeline EM (REM) algorithm which finds the anti-modes between two uni-mode clusters. We propose a way to combine the pairwise separability between clusters into the so-called aggregated distinctiveness (AD), which measures separation for the entire clustering result. Forward variable selection is applied in the attempt to maximize AD. The multivariate density estimation is obtained using the Mclust package. Components acquired from the mixture modeling in Mclust are examined for potential merging of multiple components into a single uni-mode cluster. This variable selection procedure enables us to find lower dimensional subspaces retaining the major clustering structure, useful for both visualization and discovery of important variables. We will show experimental results based on both simulated and real data sets.

12. Sampling Tables Given a Set of Conditionals

Juyoun Lee, The Pennsylvania State University

Federal agencies and other organizations publish a data summarized in arrays of non-negative integers, called a contingency table. When releasing the data, it is necessary to prevent the sensitive information of the individuals from being disclosed. In statistical disclosure limitation, we must maintain a balance between disclosure risk and data utility for the purposes of statistical inference. One method of achieving this balance is to release partial information about the original data; in practice, many federal agencies and medical institutions release data summarized in the form of marginal sums, conditional probabilities, or odds-ratios. Sampling methods for multi-way contingency tables given a set of marginal sums have been studied in diverse ways while there is almost no literature about sampling of tables given a set of conditional probabilities. Here, we focus on a set of conditional probabilities instead of a set of marginal sums. We describe the Markov chain Monte Carlo (MCMC) algorithm based on the algebraic tools for sampling contingency tables with given conditional probabilities. This algorithm can be used for Bayesian computation of posterior distribution and assessment of data utility and disclosure in statistical disclosure limitation. We demonstrate the MCMC algorithm with examples and discuss its advantages and disadvantages. We then discuss their feasibility for sampling tables given a set of conditional probabilities.

13. The Statistical Analysis of Monotone Incomplete Multivariate Normal Data

Megan Romer, The Pennsylvania State University

We consider problems in finite-sample inference with monotone incomplete data drawn from a multivariate normal population with mean μ and covariance matrix Σ . In the case of two-step, monotone incomplete data, we show that the maximum likelihood estimators (MLEs) of μ and Σ are equivariant and we obtain a new derivation of a stochastic representation for the MLE of μ . Our new derivation allows us to identify explicitly in terms of the data the independent random variables that arise in that stochastic representation. Again in the case of two-step, monotone incomplete data, we derive a stochastic representation for the exact distribution of a generalization of Hotelling's T^2 , obtain exact ellipsoidal confidence regions for μ , and derive probability inequalities for the T^2 -statistic. Finally, we apply these results to construct confidence regions for linear combinations of μ .

14. Composite Likelihood: Issues in Efficiency

Jianping Sun, The Pennsylvania State University

Maximum likelihood is a popular statistical method largely because it provides estimators with optimal statistical efficiency. However, many realistic statistical models are so complex in structure that it becomes computationally infeasible to find the MLE, especially in large data sets. One approach to solving this problem is the method of composite likelihood, which can reduce the complexity of computation at the price of some loss of efficiency. Because of its promising features, composite likelihood has recently become more and more popular in many fields such as longitudinal data, survival analysis, time series, spatial data and genetic data. A composite likelihood is constructed by taking a product of likelihood terms, each one of which is a likelihood, conditional or marginal, for some subset of the data. The statistical efficiency of such a composite likelihood then depends on the how it was constructed. In this poster, we will introduce the composite likelihood approach and then compare several methods for constructing them from an optimal efficiency point of view. To illustrate the method I will consider its use in a recombination model for DNA sequence data.

15. Bayesian Quadrature: State-Price Density Estimation

Huei-Wen Teng & John Liechty, The Pennsylvania State University

Asset pricing theories give the option price as the discounted expected payoff function under the risk-neutral probability. Such a pricing density is called the state-price density. Extracting the state-price density using options prices and/or historic data of underlying assets have been widely developed over the past decade. A nonparametric approach is preferable in that it is free of the joint-hypothesis problem, which says that any test of the economic theory is a joint test of the theory and the assumed option pricing model. We propose a quadrature method to make inference for the state-price density from a Bayesian perspective. For numerical purposes, a Gibbs sampler with slice sampling is proposed. Simulation studies and studies based on real data using S&P 500 index options show that our approach produces good model fit.

16. Assessment of Measurement Agreement for Functional Longitudinal Data

Mina Yoo, The Pennsylvania State University

Assessment of agreement is essential in many areas of research. When new methodology, assay or instrument is developed, it is necessary to check whether it can reproduce measurements within a predetermined acceptable range. Many studies have proposed indices which quantify the magnitude of the conformity of readings compared to a true value or the extent of the consistency of multiple readings. Recently, general indexes were proposed when there are more than two raters with multiple readings even considered other explanatory variables via two way mixed-effects model. The important rationales underlying the repeated measures ANOVA analysis are a common set of occasions across subjects, a complete (balanced) dataset, and the restriction to discrete covariates. However, no missing observations maintaining a common set of time periods among subjects is difficult in practice. The linear mixed model is flexible in accommodating any degree of imbalance. The parametric assumption, however, likely introduces model bias and the omission of relevant predictor variable. Even if explanatory variables are well selected, estimation of covariate is impossible when a curve through data points wiggle widely compared with number of subject. Those are common limitations encountered so far, and they may disturb to detect concordance due to induced large error. Motivated by quantifying of measurements agreement, we propose a functional agreement index for densely measured longitudinal data.

17. Regime Switching Stochastic Volatility and Correlation

Lu Zhang, The Pennsylvania State University

We proposed a multivariate regime switching covariance model. In our model, covariances are decomposed into volatilities and correlations, both of which are driven by independent hidden regime processes. The hidden regime processes are not constrained to be Markovian as in the literature. Our model is of great interests in tracking the relationship of regime switches in volatilities and correlations, which is useful in portfolio allocation and forecasting devastating economic situations where financial assets are very volatile and are highly correlated. Simulation studies showed that our model can successfully recover correlations, volatilities and their hidden regime processes. We also present some empirical analysis in daily commodity data and international market indices which conclude with some interesting findings.