Michigan Student Symposium for Interdisciplinary Statistical Sciences 2009

Paper and Poster Abstracts

April 10, 2009
Abstracts for Paper Presentation

1. Composite Graphical Model
Jian Guo, Statistics

Gaussian graphical models prove useful in capturing network structures in a group of variables. Conventional Gaussian graphical models assume that the data are homogenously sampled from a population such as Gaussian distribution. This assumption, however, is problematic for heterogenous data, i.e., the data with several categories. Consider, for example, patients with different diseases or even different types of the same disease. Estimating a single graphical model for all the patients may mask the heterogeneity in the population. On the other hand, trying to estimate a specific graphical model for each group of patients may result in completely different network structures that do not share any common links. To address this problem, we propose a composite graphical model that takes into consideration the structure imposed by the various groups of patients. To achieve this, first note that a specific network can be learned from the data in each category by estimating a sparse precision matrix. Subsequently, we propose to impose a group penalty that would encourage common links to be preserved across all categories. Consequently, the remaining individual links capture the heterogenous patterns between these categories and hence provide important information to biologist. In this article, we address the computational issues of composite graphical model and prove its theoretical properties including the consistency and the sparsistency. Evaluation of its performance is achieved by examining a number of synthetic and real data sets.

2. Proxy Pattern-Mixture Analysis for Survey Nonresponse
Rebecca Andridge, Biostatistics

We consider assessment of nonresponse bias for the mean of a survey variable Y subject to nonresponse. We assume that there are a set of covariates observed for nonrespondents and respondents. To reduce dimensionality and for simplicity we reduce the covariates to a proxy variable X that has the highest correlation with Y, estimated from a regression analysis of respondent data. We consider adjusted estimators of the mean of Y that are maximum likelihood for a pattern-mixture model with different mean and covariance matrix of Y and X for respondents and nonrespondents, assuming missingness is an arbitrary function of a known linear combination of X and Y. We propose a taxonomy for the evidence concerning bias based on the strength of the proxy and the deviation of the mean of X for respondents from its overall mean, propose a sensitivity analysis, and describe Bayesian versions of this approach. We propose using the fraction of missing information from multiple imputation under the pattern-mixture model as a measure of nonresponse bias. Methods are demonstrated through simulation and data from the third National Health and Nutrition Examination Survey (NHANES III).

3. Extrapolation of Classifier Performance
Eric Laber, Statistics

A common task in medical diagnostics is to construct a rule that discriminates among variants of a disease using easily-measured patient characteristics. Such a rule might be constructed, for example, by running a clinical trial then inputting the collected data into a learning algorithm to produce a classifier. If the learned rule is not sufficiently discriminative, the researcher may consider collecting additional data in order to improve the rule’s performance. However, the cost of extra data collection must be weighed against the expected performance gain of the learned rule. In this paper, we propose a novel method for extrapolating the discriminative power (e.g., one minus the misclassification rate) of a learning method beyond the original sample size. The proposed procedure uses the parametric bootstrap combined with an additional bias correction. The procedure applies to the class of rules commonly known as large-margin classifiers, which includes soft-margin support vector machines, boosting and logistic regression. For the special case of logistic regression we show our method is the maximum likelihood estimate with an additive bias correction. Numerical properties of our method are also studied for this case. Preliminary results are
promising when moderate signal is present in the data. We compare our method with the “gold standard”: an estimate of discriminative power when extra data is actually available. Our method is competitive with this estimator and is able to reliably extrapolate discriminative power for sample sizes up to 1.5 times the original sample size. In the case where little or no signal is present our procedure tends to overestimate the gain in discriminative performance. A regularized estimator is proposed and shown to work well with carefully-chosen tuning parameters that govern the amount of regularization. A data-driven approach for choosing these parameters is currently an open problem.

4. Stochastic Functional Data Analysis: A Diffusion Model-based Approach
Bin Zhu, Biostatistics

We consider the problem of estimating an unknown smooth function given functional data. The unknown function is treated as the realization of a stochastic process, which is incorporated into a proposed diffusion model, called a stochastic velocity model. The resulting model offers great flexibility to capture the dynamic features of functional data, and allows straightforward and meaningful interpretation. The method of smoothing splines is a special case of this approach. The likelihood of the model is derived with Euler approximation and data augmentation. Bayesian inference is carried out via a Markov Chain Monte Carlo algorithm with simulation smoother. The proposed model and method are illustrated using a blood oxygenation-level dependent signal data, and prostate specific antigen data. Keywords: Data augmentation, Euler approximation, MCMC, Smoothing spline, Simulation smoother, Stochastic velocity model.

5. Deconfounding Small Quasi-Experiments Using Propensity Scores and Other Dimension Reduction Techniques
Yevgeniya Kleyman, Statistics

When the sample is small and the number of potential confounders large, propensity adjustment may seem to have little to offer. However, in combination with dimension reduction, flexible matching, formal diagnostics and simple post-matching adjustments, it works surprisingly well. In our motivating example, a richly observed quasi-experiment comparing a faith-based and a conventional substance abuse treatment, close propensity score matching was not feasible. Still, by matching relatively coarsely on the propensity score but also matching on other scores summarizing the covariate, it was possible to balance k=27 covariates with only n=67 subjects; closer propensity matching was no better in these terms, and worse in others. In a thorough simulation study, we find that our combination of techniques reduces bias and Type I error rates in comparison to those resulting from rival approaches.

6. Meta-Analysis of fMRI Data via a Bayesian Cox Cluster Process (Highlight Student Talk)
Jian Kang, Biostatistics

Most functional magnetic resonance imaging studies (fMRI) are small in size due to cost and difficulty in recruiting special patient populations. Since the same psychological paradigms are used in many studies, there is growing interest in meta-analyses of these data. Typical data available for fMRI meta-analyses consists of all activation foci from several experiments. To date the most widely used method is the Activation Likelihood Estimation (ALE) method, either on its own or as part of a larger multi-stage method. The ALE method has known shortcomings, in particular only producing null-hypothesis inferences and providing no interpretable fitted model. In contrast, our model, a Bayesian spatial Cox cluster process model, provides an explicit fitted model and interpretable parameters. In particular our model provides information, via posterior intensity functions, about the most likely locations of activation centers at a population level and the inter-experiment spread of activated foci about these population centers. In our model, the observed activated foci are the offspring of a latent realization (population centers) of a parent process. A priori, the offspring arise from a Cox cluster process while the parent process is a homogeneous Poisson process. We demonstrate our method on an emotion activation meta-analysis of 169 studies.
Abstracts for Poster Presentation

Session 1 (10:30-11:25am)

1. Application of the Kalman filter algorithm to estimate a functional mixed model
Meihua Wu, Biostatistics

Applications of functional data analysis (FDA) methods in epidemiology remain limited in large population-level studies. One such FDA method is functional mixed models (FMMs). In functional mixed models, the dependent variable for each individual in the study is a function, which is sampled at a finite set of time points. The average across individuals is a function, which may depend on covariates. Covariate effects are also represented by functions. In small studies, FMMs can be easily estimated using standard mixed model software. However, when using this software, the computation time grows proportional to the cube of the number of observations. Thus, in large studies (in terms of the number of individuals) mixed model software fails. The Kalman filter algorithm has been proposed as a computationally efficient approach to estimate FMMs, and has been used effectively in moderate size datasets where the set of sample times is the same for all individuals. We employ the Kalman filter to estimate a FMM for a large dataset where the sample times differ across individuals. We discuss some challenges of this application, and propose some solutions to make the estimation feasible.

2. P-value weights as an integrative tool for omics data
Laila M. Poisson, Biostatistics

Cancer research is embracing the multiple "-omics" technologies available for global scale measurement of molecular events. Transcriptomics, as the global measure of gene expression, has been well developed through microarray technology and new sequencing methods. Metabolomic profiling involves chromatography coupled mass spectrometry to measure the global activity of metabolites or small molecules. Here we integrate the analysis of these two data sources to enhance the ability to find molecular changes between two disease states. Measures of differential gene expression are used to construct p-value weights for tests of differential metabolite levels in an effort to enhance the power of this platform. In this presentation we describe a shrinkage estimate for the weights, i.e. $w = w_1 + (1- \theta)w_2$, in which a gene expression based gene set enrichment score is used to construct the weight component from the prior data ($w_1$) and the metabolomic data associated with the set is used to construct the $w_2$ component. In this way tests of metabolites within a pathway that shows differential gene expression will be given a boost in power. We use simulation to explore the properties of various functions for the $w_1$, $w_2$ and components. This work is motivated by a prostate cancer progression study in which tumor samples of varying stage and benign tissue were assessed for both gene expression and metabolomic levels.

3. A Risk-Adjusted O-E CUSUM with a V-mask for Monitoring Medical Outcomes
Rena Jie Sun, Biostatistics

Motivated by a risk-adjusted one-sided CUSUM procedure within a continuous time setting, we introduce an O-E (Observed-Expected) CUSUM along with a V-mask decision criterion to simultaneously monitor for failure time outcomes that are 'worse than expected' or 'better than expected'. Appropriate V-masks are obtained for facilities of different sizes by controlling the false alarm rate over a period of given length, and simulation studies are conducted to test the performance of the proposed method, and to compared it to the one-sided CUSUM approach. A case study is carried out for 67 liver transplant programs, using the proposed O-E CUSUM, the one-sided CUSUM, and methods currently in use for flagging centers for review. The use of CUSUM methods for quality improvement is stressed.
4. Relationship Between Dialysis Schedule and Day-of-Week Association with Mortality in the Dialysis Outcomes and Practice Patterns Study
Hui Zhang, Biostatistics

Hemodialysis (HD) is an intermittent method of renal replacement therapy by which harmful waste products are removed from the body. With a thrice weekly HD schedule, the highest risk of death is thought to be on Monday/Tuesday since these days are preceded by the longest interval without dialysis. However, whether this phenomenon occurs uniformly across countries is not known. Using data from the Dialysis Outcomes and Practice Patterns Study (DOPPS), we analyzed 18,483 HD patients from the U.S. and European countries to evaluate the association between dialysis schedule and day-of-week-specific mortality. Cox proportional hazards models were constructed, with day-of-week (an external time-dependent covariate) serving as the factor of interest. Models were adjusted for gender, race, time on dialysis, body mass index (BMI) and 14 comorbid conditions; stratified by dialysis schedule, age group, country and phase; and accounted for facility clustering. Patients on a MWF schedule experienced 51% and 42% higher mortality in the U.S. and Europe, respectively; the corresponding values being 60% and 30% for TTS patients. In general, the day-of-week effect was stronger for cardiovascular disease (CVD) death compared with non-CVD death. In the U.S., CVD mortality risk was higher by 45% for MWF schedule and 83% for TTS schedule. In European countries, CVD mortality risk was higher by 66% for MWF schedule and 110% for TTS schedule. Although a large number of interactions were evaluated, only neurological disease, serum sodium and gender were observed to modulate the day-of-week effects. Conclusions. The thrice weekly HD schedule may contribute to the elevated risk of death on Monday/Tuesday. The European experience is less pronounced than U.S., although some elevation in death risk for Monday and Tuesday is observed, especially for CVD death. Future studies should examine practice patterns which may explain the regional differences.

5. Regression analysis on a covariate with heteroscedastic measurement error
Ying Guo, Biostatistics

We consider the estimation of the regression of an outcome D on a covariate X, where X is unobserved, but a variable Y which measures X with error is observed. A calibration sample that measures pairs of values of X and Y is also available; we consider calibration samples where D is measured (internal calibration) and not measured (external calibration). One common approach for measurement error correction is Regression Calibration (RC), which substitutes the unknown values of X by predictions from the calibration curve of X on Y. An alternative approach is to multiply impute the missing values of X given Y and D based on an imputation model, and then use multiple imputation (MI) combining rules for inferences. A recent paper by Freedman et al (2008) compares these two approaches, suggesting that RC is more efficient under plausible assumptions. However, their work assumes the measurement error of Y has a constant variance, whereas in many situations, the variance varies as a function of X. We consider modifications of the RC method and the MI method that allow for heteroscedastic measurement error, and compare them by simulation. The MI method is shown to provide better inferences in this setting.

6. Propensity Score Matching in Randomized Clinical Trial
ZhenZhen Xu, Biostatistics

Cluster randomization trials with relatively few clusters have been widely used in recent years for evaluation of health care strategies. On average, randomized treatment assignment achieves balance in both known and unknown confounding factors between treatment groups, however, in practice investigators can only introduce a small amount of stratification and cannot balance on all the important variables simultaneously. The limitation arises especially when there are many confounding variables and in small studies. Such is the case in the INSTINCT trial designed to investigate the effectiveness of an education program in enhancing the tPA use in stroke patients. In this paper, we introduce a new randomization design, the balance match weighted (BMW) design, which applies the optimal matching with constraints technique to a prospective randomized design and aims to minimize the mean squared error of the treatment effect.
estimator. When true confounding effects are known, we construct the BMW design to produce treatment effect estimators with minimal MSE; these results suggest that, even when the confounding effects are unknown, the BMW design with appropriately chosen parameters can generate treatment effect estimators with substantially improved MSE properties. The simulation study shows that, under various confounding scenarios, the BMW design can reduce the MSE for the treatment estimator by 10% to as much as 80% compared to a completely randomized or matched-pair design. We illustrate these methods in proposing a design for the INSTINCT trial.

7. A new approach to Cholesky-based covariance regularization in high dimensions
Adam J. Rothman, Statistics

We propose a new regression interpretation of the Cholesky factor of the covariance matrix, as opposed to the well known regression interpretation of the Cholesky factor of the inverse covariance, which leads to a new class of regularized covariance estimators suitable for high-dimensional problems. Regularizing the Cholesky factor of the covariance via this regression interpretation always results in a positive definite estimator. In particular, one can obtain a positive definite banded estimator of the covariance matrix at the same computational cost as the popular banded estimator proposed by Bickel and Levina (2008), which is not guaranteed to be positive definite. We also establish theoretical connections between banding Cholesky factors of the covariance matrix and its inverse and constrained maximum likelihood estimation under the banding constraint, and compare the numerical performance of several methods in simulations and on a sonar data example.

8. Predicting Traffic on Computer Networks
Joel Vaughan, Statistics

In order to maintain consistent quality of service, computer network engineers face the task of monitoring the traffic fluctuations on the individual links making up the network. However, due to constraints in resources and access, it is not possible to directly measure all the links consistently. Therefore, it is of interest to be able to predict the traffic fluctuations on an unobserved set of links using the measurements from the observed links. To this end, we introduce a mechanistic model describing the network-wide traffic that is consistent with features known to be present in traffic over a single, fixed link. We then use the spatial and temporal correlation structure specified by the model to develop a method of predicting the fluctuations on the unobserved links using fluctuations on the observed links in the network. Although similar in spirit to kriging methods from spatial statistics, our method makes use of the additional structure imposed on computer networks by routing tables. We illustrate the effectiveness of our approach with simulations from NS/2, a discrete-event network simulator, and with traffic data from the Internet2 backbone network.

9. A Bootstrapped Linear Two-Stage Plan
Runlong Tang, Statistics

Motivated by dose-finding problems in toxicology and drug development, two-stage procedures for the estimation of the inverse of a monotone regression function at a specific point are considered. It is known that the convergence rate of an estimator from a one-stage isotonic regression plan (OSIRP) built on a budget of \( n \) sample points is \( n^{1/3} \). Taking OSIRP as the first stage with some sample points to obtain an initial estimator of the targeted quantity, a linear two-stage plan (LTSP) then evenly allocates the remaining design points at two shrinking points beside the initial estimator, locally fits a linear model to the second-stage sample points, and estimates the target quantity again. In LTSP, more design points are placed around the true target quantity and the convergence rate of the more efficient estimator increases to \( n^{1/2} \). Furthermore, a bootstrapped version of LTSP (BLTSP) is tailored to avoid estimating an unknown quantity, which is difficult to estimate but needed in the construction of confidence intervals. Consistency and asymptotic distributions of the estimators from both LTSP and BLTSP are derived. A finite-sample relationship be-
tween an adaptive LTSP (ALTSP) and an adaptive BLTSP (ABLTSP) is analyzed. Finally, finite-sample performance of an implementable ABLTSP is studied through simulations.

10. The solution path of Markovitz’s portfolio problem
Zach Zhanyang Zhang, Statistics

In this paper we investigated the solution of Markovitz’s portfolio problem (Markovitz 1952) in the perspective of constrained optimization and variable selection. Inspired by the piecewise linear solution of LASSO, we revealed that the solution path of the portfolio problem is also piecewise linear. And by proposing an on-off algorithm, we were able to give the solution path efficiently. Finally, we applied our method on the S&P500 index data set, and illustrated helpful ways of using the solution path obtained to make decisions.

11. Logistic Regression Under Uncertainty
Patrick Harrington, Statistics

We present a robust estimate of the parameters involved in linear logistic regression when confronted with bounded uncertainties in the data. The relationship to worst case logistic regression under bounded data uncertainties is presented and circumstances of equivalence are discussed through properties of the Lagrange dual. We also present the relationship between ‘2-regularized logistic regression and our minimax estimation formulation with bounded uncertainty radius playing a similar role as the regularization parameter in ridge logistic regression. A geometric interpretation of the closely related support vector machine classifier is presented, providing additional insight into how this robust estimation is operating. Numerical results are presented on a synthetic data set and future work on kernelizing this method are discussed.

12. Variable Selection With Grouped Penalties
Arnau Tibau Puig, Biostatistics

Our goal is to carefully identify a set of genes that discriminate between two classes of patients. Logistic regression is a good approach for variable selection when we have observations labelled with a binary response. However, as in any regression paradigm, this problem becomes computationally difficult to solve whenever the dimensionality of the data is smaller than the number of samples. Unfortunately, our dataset consists of tens of samples (patients) versus tens of thousands of variables (genes) which means that regularization is essential. Regularization in logistic regression is a well-studied problem. Typically, regularization was performed at the variable level using L1 or L2 type of penalties. More recently, the use of other types of penalties that incorporate further structure while performing regularization have become an important subject of study. The work initiated by Yuan and Lin [3] on group penalties for linear regression problems was extended by Meier et al. [1] to the logistic regression framework. Further extensions have been proposed to incorporate hierarchical dependencies in the data, see for instance [4] or [2]. In this work we address several practical and theoretical problems related to the application of these classes of methods to the variable selection problem when we group the genes according to a certain biological structure. This structure is obtained from the well-known KEGG database, which provides information about gene-pathway dependencies.

13. Error Corrected Gaussian Mixture Model and Precision Measurements of the Properties of the Color Clustering in Galaxy Clusters
Jiangang Hao, Physics

Galaxy clusters are the largest gravitationally bound systems in our Universe and their abundance as well as distribution encodes rich information about the composition and evolution of the Universe. The clustered galaxies are also tightly clustered in colors, which provides an efficient way for detecting clusters. The
galaxies’ color distribution around a cluster is bimodal, one peak corresponding to the clustered galaxies and another corresponding to non-clustered galaxies, which can be well approximated by a mixture of two Gaussian distributions. Gaussian Mixture Model is well suited for fitting the color distribution. However, the colors have non-negligible measurement errors, which should be appropriately modelled during the fitting to avoid the bias. In this paper, we propose a generalized Gaussian Mixture Model with inclusion of measurement errors and derive the corresponding EM recursive relation for maximizing the likelihood. We applied this technique to 1,3000 galaxy clusters and provide the first ever precision and unbiased measurements on the galaxy color evolution across a wide redshift range. This precise information give tight constraints on the evolution of galaxies and has significant impact on theoretical modelling of galaxy formation.

14. The Use of Mixed Rasch Model to Measure and Account for Question Sensitivity
Zeina Mneimneh, Survey Methodology

One of the most troubling features that researchers are concerned about is question sensitivity. This is mainly because questions that are judged to be sensitive could lead to measurement and/or nonresponse biases in the estimates. Two approaches could be adopted to deal with question sensitivity: 1) a reduction approach and 2) an adjustment approach. In the first approach methodologists could try to reduce the effect of question sensitivity by using certain design strategies. In the second approach, analysts could try to model question characteristics when estimating the constructs of interest. Before implementing any design strategy, methodologists adopting the first approach need to properly identify and measure question sensitivity. Unfortunately, current methods used to measure question sensitivity employ either arbitrary broad categorization or rely on subjective assessments. On the other hand and in relation to the second approach, most adjustment techniques applied by psychometricians account for certain item characteristics such as difficulty, discrimination, and guessing. Yet, common models in practice have not been used to isolate and account for the item sensitivity. In this poster we propose to use Mixed Rasch Models (MRM) to have a better understanding of question sensitivity. We propose that these models could identify latent groups of respondents who employ different response strategies when answering sensitive items. Moreover, by investigating variations in item difficulty parameters within and across these latent groups, a more empirically driven method could be identified to measure item sensitivity. Thus such MRM will be of value for both approaches described above. Finally we describe how such models could be used to improve real-time adaptive measurement by tailoring questions with a level of difficulty and/or sensitivity that is appropriate to respondents.

Session 2 (1:55-2:50pm)

15. Analysis of Gene Sets Based on the Underlying Regulatory Network
Ali Shojaie, Statistics

Networks are often used to represent the interactions among genes and proteins. These interactions are known to play an important role in vital cell functions and should be included in the analysis of genes that are differentially expressed. Methods of gene set analysis take advantage of external biological information and analyze a priori defined sets of genes. These methods can potentially preserve the correlation among genes, however, they do not directly incorporate the information about the gene network. In this paper, we propose a latent variable model that directly incorporates the network information. We then use the theory of mixed linear models to present a general inference framework for the problem of testing the significance of subnetworks. Several possible test procedures are discussed and a network based method for testing the changes in expression levels of genes as well as the structure of the network is presented. The performance of the proposed method is compared with methods of gene set analysis using both simulated data sets as well as real data on genes related to the Galactose Utilization pathway in yeast.
16. Dealing with Non-regularity in Optimal Dynamic Treatment Regimes
Bibhas Chakraborty, Statistics

Dynamic treatment regimes are individually tailored treatments. They offer a way to operationalize the adaptive multistage decision making in clinical practice, thus providing an opportunity to improve such decision making. However, when using longitudinal data on patients to construct these treatment regimes, hypotheses concerning the choice of the optimal treatment at each stage may involve non-regular parameters. The non-regularity stems from the fact that parameters are functions of maxima. As a result, the parameter estimates can be biased, and traditional methods of constructing confidence intervals can have poor frequentist properties. In this paper, we present and evaluate a method that adapts to this non-regularity by the use of an empirical Bayes approach. We also demonstrate the use of this method through the analysis of a data set from a randomized smoking cessation trial.

17. Confidence Intervals for the Test Error In Classification Using the Adaptive Bootstrap
Eric Laber, Statistics

A small training set is fed into a learning algorithm producing a classifier. In the absence of a validation set, can we construct a valid confidence interval for the test error of the learned classifier? A common approach to construct a confidence interval is to form a point estimate of the test error and then assume the estimator follows a known distribution. Alternatively, one might resample the estimator of the test error to form a confidence interval. Unfortunately, these approaches do not reliably deliver the desired coverage in small samples. In fact, even in large samples, coverage is not guaranteed in many of these approaches, including the bootstrap, normal approximation, and repeated data splitting. The reason for this is that the test error is a non-smooth functional of the learned classifier. We propose a new method called the adaptive bootstrap for constructing a confidence interval for the test error in the small-sample setting. This method is based on bootstrapping an adaptively-regularized estimator of the test error. The procedure is adaptive in the sense that the estimated amount of non-smoothness in the test error governs the amount of regularization. The regularization induces smoothness, and consequently, bootstrapping this estimator provides asymptotically consistent confidence intervals. We evaluate this procedure on a suite of simulated and test data sets. It is shown to provide the desired coverage, markedly outperforming the standard bootstrap.

18. Statistical Inference of Functional Connectivity in Neuronal Networks using Frequent Episodes
Kohinoor Dasgupta, Statistics

Identifying the spatio-temporal network structure of brain activity from multi-neuronal data streams is one of the biggest challenges in neuroscience. Repeating patterns of precisely timed activity across a group of neurons is potentially indicative of a microcircuit in the underlying neural tissue. Frequent episode discovery, a temporal data mining framework, has recently been shown to be a computationally efficient method of counting the occurrences of such patterns. In this paper, we propose a framework to determine when the counts are statistically significant by modelling the counting process. Our model allows direct estimation of the strengths of functional connections between neurons with improved resolution over previously published methods. It can also be used to rank the patterns discovered in a network of neurons according to their strengths and begin to reconstruct the graph structure of the network that produced the spike data. We validate our methods on simulated data and present analysis of patterns discovered in data from cultures of cortical neurons.

19. Sampling from Doubly-intractable Distributions
Jing Wang, Statistics

Parameter-dependent intractable normalizing constants occur in many statistical models, e.g. social net-
works, image analysis, protein design, etc. In a Bayesian framework, posterior distributions induced by such
likelihood functions are called doubly-intractable distributions. Standard Markov Chain Monte Carlo (MCMC)
fails to apply to this scenario, since the intractable normalizing constants of the likelihood are functions of
the unknown parameters. Existing MCMC algorithms for doubly-intractable distributions include auxiliary
variable method, exchange algorithm with bridging, and an adaptive MCMC sampler developed by Yves
Atchade et al. (08). However, the former two rely on the ability of performing perfect sampling, and thus
are computationally expensive and simply infeasible for many models of interest. The last method is more
general, but still suffers from heavy computation load. We propose a more efficient and general MCMC
method to sample from doubly-intractable distributions. In each iteration of the MCMC sampler, a path
is built between the posterior distributions evaluated at the current and proposed parameter values. An
ensemble of auxiliary variables are generated from transition kernels with invariant distributions the inter-
mediate distributions along the path. Plugging them into the path sampling identity gives an approximation
of the log-ratio of the normalizing constants. We prove that as the number of intermediate distributions
approaches infinity, the limiting distribution of this sampler is the target posterior distribution. We carry
out simulation studies for Ising model and conditional random field model. The results show that our sam-
pler converges to the target posterior distributions, and has good mixing rate. Moreover it outperforms the
existing methods in computing time greatly. We also apply the new sampler to the parameter estimation
of a collaboration network in a New England law firm. Some theoretical properties of the new method are
studied, and preliminary proofs will be given.

20. How to performance of a classification procedure depends on multiple distinct attributes
of the training set and its population structure
Juan Zhang, Statistics

The performance of a classification procedure depends on multiple distinct attributes of the training set
and its population structure. Several of these attributes are well understood (e.g. the effect of training
set sample size, the separation between the classes in the variable space, collinearity among the predictor
variables, and measurement error in the predictor variables). Using logistic regression as a familiar context,
at least two other attributes can also be demonstrated to be important: (i) the "effect concentration," which
we assess as the coefficient of variation of the population vector $\beta$ of regression coefficients, and (ii) the
"alternatingness" of effects, which broadly means whether positively correlated predictor variables tend to
have effects in the same direction. Through extensive simulations, we considered the relationship of these
attributes to predictive performance of a classifier as quantified by "Area Under the Curve" (AUC). A consis-
tent pattern emerges in which the AUC decreases with effect concentration for non-alternating populations,
while it increases with effect concentration for alternating populations. The amount by which AUC changes
with effect concentration is negligible for small levels of measurement error but is substantial for moderate
levels of measurement error. Our findings are generally relevant, but our main application of interest is to
medical diagnostic and prognostic procedures where biomarkers are expected to be correlated and observed
with measurement error. To aid in interpreting our findings, we used several molecular assay datasets to
identify the likely range of values for parameters influencing prediction performance, including measurement
error, predictor correlations, effect concentration, and alternatingness. The latter two attributes are partic-
ularly challenging as they relate to $\beta$ and not just to the structure of the predictors. We describe some
speculative approaches for getting an idea as to what these attributes might be in real data sets.
21. Uncovering population structures of marginal and conditional associations in high dimensional data
Ming-Chi Hsu, Statistics

The associations between pairs of variables in a high dimensional population have a complex structure that can be inferred to some degree from data. Since the number of variable pairs is huge, we asked whether the "marginal pattern" of pairwise associations yields any interesting insights. Focusing on marginal and conditional correlation coefficients, we developed a statistical method that estimates this marginal pattern from sample correlation coefficients. The method primarily involves removing the overdispersion due to sampling variance in the estimates. An application of this framework is to assess the presence of sparsity in the population correlations based on data, where sparsity means that a large fraction of the associations are negligibly small. We did this by representing the pattern of population associations as a mixture of a negligible (but not degenerate) component, and a second component that captures the non-negligible effects. Simulation studies showed that our method accurately reconstructs various levels of sparsity when it is present. Using microarray gene expression data for a catalog of human tissues, our method failed to find any evidence of sparsity in the marginal correlations. Currently we are considering whether the level of sparsity increases when the correlations are conditioned on major variance components identified either empirically, or through external biological knowledge. It has been proposed that by exploiting sparsity, if present, may more efficiently identify molecular signatures for human diseases. But if the assumption of sparsity fails to hold, procedures attempting to exploit it may perform worse than traditional procedures. Our goal is to provide a means to assess whether sparsity exists in a given population, so researchers can make informed decisions on how to best analyze a high-dimensional data set.

22. Plug-In estimators of Information Theoretic measures with application to Image Registration
Kumar Sricharan K., EECS

The image registration problem is to align images gathered via multiple sensors located at different locations by applying a sequence of intensity preserving transformations so that they have an identical pose in a common coordinate system. The image registration performance depends on two critical factors - the selection of discriminating feature sets, and the choice of similarity measures to match these feature sets. Information theoretic similarity measures such as $\alpha$-Jensen difference, $\alpha$-Geometric Arithmetic difference and $\alpha$ Mutual Information [2] have been shown to be effective in this regard, particularly in the multi-modal image registration setting. Several estimators of information theoretic measures have been proposed [3, 2, 1] and applied to the image registration problem. These estimators have been shown to be asymptotically consistent. However the performance of these estimators in the finite sample setting is unknown. We present plug-in estimators for information theoretic measures and derive their performance properties in terms of the bias, variance and therefore mean square error, asymptotic distribution, and confidence intervals, as functions of sample sizes. The theoretical performance guarantees of our plug-in estimators allow us to optimize over the free parameters in our plug-in estimators for optimal performance. This also allows us to predict the performance of our estimators - which in turn translates to predicting registration error. Equivalently, these theoretical performance guarantees lets us specify the requirements in terms of sample size to obtain a desired standard of registration performance. We perform geo-registration - image registration on images obtained from airborne remote sensing platforms such as NASA’s AVIRIS - using the proposed plug-in estimates and show that the observed performance matches that predicted by our theory.

23. Shrinkage estimators for Gaussian covariance matrices
Yilun Chen, EECS

We address covariance estimation under mean-squared loss in the Gaussian setting. Specifically, we consider shrinkage methods which are suitable for high dimensional problems with small number of samples (large p small n). First, we improve on the Ledoit-Wolf (LW) method by conditioning on a sufficient statistic via
the Rao-Blackwell theorem, obtaining a new estimator RBLW whose mean-squared error dominates the LW under Gaussian model. Second, to further reduce the estimation error, we propose an iterative approach which approximates the clairvoyant shrinkage estimator. Convergence of this iterative method is proven and a closed form expression for the limit is determined, which is called the OAS estimator. Both of the proposed estimators have simple expressions and are easy to compute. Although the two methods are developed from different approaches, their structure is identical up to specific constants. The RBLW estimator provably dominates the LW method; and numerical simulations demonstrate that the OAS estimator performs even better, especially when n is much less than p.

24. A Bayesian Generalized Non-Linear Predictive Model of Treatment Efficacy Using qMRI
Jincao Wu, Biostatistics

The prognosis for patients with high-grade gliomas is poor with a median survival of one year after diagnosis. The assessment of treatment efficacy is typically unavailable until about 8 to 10 weeks post treatment. Investigators hypothesize that recently developed Quantitative MRI (qMRI) techniques can predict the treatment efficacy only 3 weeks from the initiation of the therapy thereby allowing second line therapies to begin earlier. The purpose of this work is to build a predictive model for the treatment efficacy based on qMRI data and baseline prognostic factors. We use 1 year survival status as the outcome and propose a Bayesian joint model. In the first stage, we smooth the qMRI data using a pairwise-difference prior and derive summary statistics. In the second stage, these statistics are used in a generalized non-linear model with a Multivariate Adaptive Regression Spline (MARS) basis in the systematic component and a probit link. Gibbs sampling and reversible jump Markov chain Monte Carlo are applied iteratively between the two stages to estimate the posterior. Bayesian model averaging is employed to derive the final predictive model.

25. A Markov Compliance Behavior and Outcome Model for Causal Analysis in Longitudinal Studies
Xin Gao, Biostatistics

We propose a Markov compliance behavior and outcome model for analyzing longitudinal randomized studies when non-compliance cannot be ignored. We consider longitudinal studies where subjects are randomized to the treatment or control group only at baseline, but subjects’ compliance behaviors may vary over time. The proposed model solves the problem in the potential outcome framework, and provides causal estimates on the treatment effect via principal stratification. Previous research (Lin, Ten Have, and Elliott, JASA 2007) considered the effect of subjects’ joint compliance behavior on the joint distribution of the longitudinal outcomes, but not the effect of outcomes at time t-1 on the compliance behaviors at time t, which is often of great interest to investigators. The proposed Markov compliance behavior and outcome model provides estimates both on the effect of the compliance behavior on the following outcome, and on the effect of the outcome on the following compliance behavior. The model requires assumptions to be made about the unobservable correlation among a subject’s potential outcomes. We conduct a sensitivity analysis by varying the correlation. We applied the proposed model on a longitudinal study and estimate the parameters and causal effect using Markov chain Monte Carlo (MCMC) methodology.

26. Assessing the Convergence of Multiple Imputation Algorithms Using a Sequence of Regression Models
Jian Zhu, Biostatistics

Multiple imputation algorithms using a sequence of regression models are commonly used to handle non-responses in complex survey studies. Although such algorithms have several advantages over joint modelling of all survey variables, they have a theoretical limitation that the specified conditional distributions could be incompatible and the underlying joint distribution of the survey variables may not exist. Although previous simulation studies show that imputation algorithms using incompatible conditional distributions work well
for some cases, the performance of such algorithms for complex data needs to be studied. We focus on
general multivariate data to assess the convergence properties of the imputation algorithms using various
types of conditionally specified models. We also evaluate the impact of incompatible models on imputation
results through simulation studies.

27. **Query large scale microarray compendium datasets using a model-based Bayesian approach with variable selection**
Ming Hu, Zhaohui S. Qin

In microarray gene expression data analysis, it is often of interest to identify genes that share similar expression profiles with a particular gene such as a key regulatory protein. Multiple studies have been conducted using various correlation measures to identify co-expressed genes. While working well for small datasets, the heterogeneity introduced from increased sample size inevitably reduces the sensitivity and specificity of these approaches. This is because most co-expression relationships do not extend to all experimental conditions. With the rapid increase in the size of microarray datasets, identifying functionally related genes from large and diverse microarray gene expression datasets is a key challenge. We develop a model-based gene expression query algorithm built under the Bayesian model selection framework. It is capable of detecting co-expression profiles under a subset of samples/experimental conditions. In addition, it allows linearly transformed expression patterns to be recognized and is robust against sporadic outliers in the data. Both features are critically important for increasing the power of identifying co-expressed genes in large scale gene expression datasets. Our simulation studies suggest that this method outperforms existing correlation coefficients or mutual information-based query tools. When we apply this new method to the Escherichia coli microarray compendium data, it identifies a majority of known regulons as well as novel potential target genes of numerous key transcription factors.

28. **Estimating the recurrence-free survival benefit of salvage androgen deprivation therapy following external beam radiation therapy for prostate cancer**
EH Kennedy, Biostatistics

Patients treated for prostate cancer are followed over time by monitoring prostate-specific antigen (PSA) levels. The rise of PSA is an indication of the possible recurrence of prostate cancer. Patients may then receive salvage androgen deprivation therapy (ADT), with the goal of delaying recurrence. An open question is how to estimate the effectiveness of salvage ADT in reducing the risk of clinical recurrence of cancer. As a time-dependent covariate, PSA is both a confounder and an intermediate variable in the relation between salvage ADT and clinical recurrence of prostate cancer. As a result, standard procedures for the adjustment of this confounder are biased. We compare two methods which appropriately adjust for time-dependent confounding, using data from 2,781 patients treated with external beam radiation therapy (EBRT) for prostate cancer. The first is a two-stage method which uses a linear mixed model to predict PSA profiles after EBRT, and then uses a time-dependent Cox proportional hazards model to estimate the recurrence-free survival benefit of salvage ADT, adjusting for the predicted PSA profiles. The second method, called sequential stratification, uses a stratified Cox proportional hazards model where strata are defined by salvage ADT patients and their respective controls (matched on predicted PSA profiles). Estimates of the recurrence-free survival benefit of salvage ADT provided by these two methods are similar. Using the first method, the estimated risk of recurrence among patients undergoing salvage ADT is 76.8% lower than that of patients who did not undergo salvage ADT (RR = 0.23, 95% CI: (0.17, 0.32)); using the second method, the risk is 76.6% lower (RR = 0.23, 95% CI: (0.16, 0.34)). Both methods can be extended to investigate the interaction effects of salvage ADT with time and other covariates.