Program

Eighth Annual Winter Workshop
Frontiers of Theoretical Statistics

Department of Statistics
University of Florida
January 13-14, 2006

Contents

Sponsors

This year's symposium is funded by the National Science Foundation and Info Tech, Inc., along with the Graduate School, the College of Liberal Arts and Sciences and the Department of Statistics of the University of Florida.

Organizing Committee

George Casella
Malay Ghosh
André Khuri
Ramon Littell
Clyde Schoolfield
Alex Trindade

Invited Speakers

James Berger, ISDS, Duke University, and SAMSI
Peter Bickel, University of California at Berkeley
Ron Butler, Colorado State University
R. Dennis Cook, University of Minnesota
Jim Fill, Johns Hopkins University
Peter McCullagh, University of Chicago
Susan Murphy, University of Michigan
Nancy Reid, University of Toronto
Christian Robert, Ceremade University, Paris, France
Michael Steele, University of Pennsylvania
Stephen Stigler, University of Chicago
Bin Yu, University of California at Berkeley

Acknowledgements

The organizers thank the Department of Statistics staff, especially Carol Rozear, Marilyn Saddler, Robyn Crawford and Tina Greenly for their tremendous efforts in helping to set up this meeting and make it run smoothly.

Conference Schedule

Thursday, January 12, 2006

7:00-10:00 pm Welcome reception at the Keene Faculty Center, Dauer Hall

Friday, January 13, 2006

7:30-8:15am Breakfast (J. Wayne Reitz Union, Room 282)

All Sessions Meet in the J. Wayne Reitz Union, Room 282

8:15-8:30am Welcome by André Khuri, Conference Chair
Neil Sullivan, Dean, College of Liberal Arts Sciences
George Casella, Chair, Department of Statistics

8:30-10:30am **Session 1 Chair:** André Khuri

Speakers:
Stephen Stigler: *Expect a High Casualty Rate at the Frontier*

R. Dennis Cook: *Subspace Models and Methods*

10:30-11:00am **Break and Conference photo at JWRU South side**

11:00-1:00pm **Session 2 Chair:** Malay Ghosh

Speakers:
Peter McCullagh: *Partition Models And Cluster Processes With Statistical Applications*

Nancy Reid: *Likelihood-Based Inference In Complex Models*

1:00-2:30pm **Lunch** Gator Corner Dining Center

2:30-4:30pm **Session 3 Chair:** Clyde Schoolfield

Speakers:
Jim Fill: *Perfect Sampling From The Dickman Distribution And Other Perpetuities Using Coupling Into And From The Past*

Christian P. Robert: *Minimum Variance Importance Sampling Via Population Monte Carlo*

4:30-6:30pm **Poster Session:** J. Wayne Reitz Union, Rooms 272-273, Rooms 276-277 and Rooms 288-291

Saturday, January 14, 2006

8:00-8:30 Breakfast (J. Wayne Reitz Union, Room 282)

All Sessions Meet in Reitz Union

8:30-10:30am **Session 4 Chair:** Brett Presnell

Speakers:

Bin Yu: *Lasso: Blasso Algorithm And A Model Selection Consistency Result*

Susan Murphy: *Experiments and Dynamic Treatment Regimes*

10:30-11:00am **Break**

11:00-1:00pm **Session 5 Chair:** Alex Trindade

Speakers:

J. Michael Steele: *Yield Curve Models and Their Applications*

Ron Butler: *Cybernetics and Stochastic Systems: From Electrical Engineering to Biostatistics*

1:00-2:30pm **Lunch** – free time

2:30-4:30pm **Session 6 Chair:** George Casella

Speakers:

James Berger: *Some Recent Developments in Bayesian Model Selection*

Peter Bickel: *Estimating Large Covariance Matrices*

5:30-8:30pm. **Dinner**: United Church of Gainesville, Fellowship Hall
1624 NW 5th Ave., Gainesville, FL 32603

Invited Talks

Expect a High Casualty Rate at the Frontier

Stephen Stigler, University of Chicago

Any frontier is a dangerous place, and the risks of fatality are high. No advance comes without cost; we can only hope that some gain rewards the sacrifice. The frontier of theoretical statistics is no exception. Some brave explorers left no trace - who remembers Robert Adcock, Erhard Tournier, or Isidore Didion? Others left the landscape permanently changed, despite a long series of unsuccessful forays. These eternal truths are illustrated by the 250 year history of maximum likelihood, a history that has shown remarkable progress emerging from a long sequence of brave and ingenious errors.

Subspace Models and Methods

R. Dennis Cook, University of Minnesota

Dimension reduction in statistics often hinges on the idea of projecting the data onto a low dimensional reductive subspace, defined so that the projected data contains the same relevant information as the full data. Both model-free and model-based methods for dimension reduction can hinge on estimating a reductive subspace. Models may be parameterized by a minimal reductive subspace while model-free methods rely on reductive subspaces that apply across large classes of models.

We will discuss subspace models and methods as a class, focusing on unifying themes and outstanding issues. This unification encompasses classical sufficient statistics and leads to new approaches for dimension reduction, including ideas for reducing discrete multivariate data.

Partition Models And Cluster Processes With Statistical Applications

Peter McCullagh, University of Chicago

A partition of the set $[n]=\{1 \uparrow n\}$ is a set of disjoint non-empty subsets whose union is $[n]$, and a random partition of $[n]$ is a probability distribution on set E_n of partitions of $[n]$. Two statistical applications involving partition models are described, one connected with Bayesian multiple comparisons, and one connected with the notion of an exchangeable cluster process. Both applications require the concept of a partition process in which P_n is the marginal distribution of P_{n+1} under unit deletion.

A cluster process is a sequence of random variables $Y_1, Y_2, \ldots$ together with a random partition B of the index set. The observation on n individuals consists of a finite sequence $Y_1,\ldots,Y_n$ together with a random partition B of $[n]$. For an exchangeable cluster process, the joint distribution P_n of $(B, Y_1,\ldots. Y_n)$ is invariant under permutation of units, and P_n is the marginal distribution of P_{n+1}. A simple process having these properties is described together with applications to classification problems.

Likelihood-Based Inference In Complex Models

Nancy Reid, University of Toronto, Canada

Models for large or highly structured data often lead to likelihood functions that are difficult to use for inference. A variety of likelihood-motivated approaches have been suggested, such as quasi-likelihood, pseudo-likelihood, composite likelihood, pairwise likelihood, etc. I will survey some of the recent work on likelihood-like functions in complex models, with a view to understanding how these approaches are useful for inference, and what their limitations are.

Perfect Sampling From The Dickman Distribution And Other Perpetuities Using Coupling Into And From The Past

Jim Fill, Johns Hopkins University

I plan to review the algorithm known as "Coupling Into And From The Past" (CIAFTP) and show how it can be used for efficient perfect sampling from the Dickman distribution, a distribution of importance in number theory, the analysis of algorithms, and (historically) the study of unimodal distributions. Our perfect sampling approach extends to other perpetuities, including the Vervaat family. This is joint work with Mark Huber of Duke University.

Minimum Variance Importance Sampling Via Population Monte Carlo

Christian P. Robert, Ceremade University, Paris France

In the design of efficient simulation algorithms, one is often beset with a poor choice of proposal distributions. Although the performances of a given kernel can clarify how adequate it is for the problem at hand, a permanent on-line modification of kernels raises concerns about the validity of the resulting algorithm. While the issue is quite complex and most often intractable for MCMC algorithms, the equivalent version for importance sampling algorithms can be validated quite precisely. We derive sufficient convergence conditions for a wide class of population Monte Carlo algorithms and show that Rao--Blackwellized versions asymptotically achieve an optimum in terms of a Kullback divergence criterion, while more rudimentary versions simply do not benefit from repeated updating. In particular, since variance reduction has always been a central issue in Monte Carlo experiments, we show that population Monte Carlo can be used to this effect, in that a mixture of importance functions, called a D-kernel, can be iteratively optimised to achieve the minimum asymptotic variance for a function of interest among all possible mixtures.The implementation of this iterative scheme is illustrated for the computation of the price of a Europeanoption in the Cox-Ingersoll-Ross model.

Lasso: Blasso Algorithm And A Model Selection Consistency Result

Bin Yu, University of California at Berkeley

Information technology advances are making data collection possible in most if not all fields of science and engineering and beyond. Statistics as a scientific discipline is challenged and enriched by the new opportunities resulted from these high-dimensional data sets. Often data reduction or feature selection is the first step towards solving these massive data problems. However, data reduction through model selection or L_0 constrained optimization leads to combinatorial searches which are computationally expensive or infeasible for massive data problems. A computationally more efficient alternative to model selection is L_1 constrained optimization or Lasso optimization.

In this talk, we will describe the Boosted Lasso (BLasso) algorithm that is able to produce an approximation to the complete regularization path for general Lasso problems. BLasso consists of both a forward step and a backward step. The forward step is similar to Boosting and Forward Stagewise Fitting, but the backward step is new and crucial for BLasso to approximate the Lasso path in all situations. For cases with finite number of base learners, when the step size goes to zero, the BLasso path is shown to converge to the Lasso path. Experimental results are also provided to demonstrate the difference between BLasso and Boosting or Forward Stagewise Fitting. We can extend BLasso to the case of a general convex loss penalized by a general convex function and illustrate this extended BLasso with examples.

Since Lasso is used as a computationally more efficient alternative to model selection, it is important to study the model selection property of Lasso. If time allows, I will present some (almost) necessary and sufficient conditions for Lasso to be model selection consistent in the classical p << n setting. This is joint work with Peng Zhao at UC Berkeley.

Experiments and Dynamic Treatment Regimes

Susan Murphy, University of Michigan

We discuss experimental designs that can be used in constructing and refining dynamic treatment regimes. These experimental designs accommodate settings in which both the timing of when to alter treatment and the type of treatment alteration is important. Treatments may have multiple factors including both therapeutic and "encouragement to adhere" factors. We discuss how special care in the design and inference must be taken in order to preserve causal inferences and discuss issues in defining main effects and interaction effects between factors.

Yield Curve Models and Their Applications

J. Michael Steele, University of Pennsylvania

This talk, which presupposes no background in finance, will first describe the motivation for yield curve models and then briefly summarize the evolution of these models from their inception to the present. We then consider the uses that are made of yield curve models in financial markets. In particular, we consider how yield curve models provide models for interest rate futures and for interest rate swaps. Financial theory and the empirical financial efficiency of these real-world markets are then found to imply that our models should have certain martingale properties. Subsequently, these martingale properties are shown to follow from certain simpler analytic conditions. Finally, we call on market data to discern the circumstances where the simpler martingale criteria are met. Somewhat surprisingly, we find that in some important situations, the martingale criteria are violated empirically.

Cybernetics and Stochastic Systems: From Electrical Engineering to Biostatistics

Ron Butler, Colorado State University

The cybernetics movement from the ’50s to ’70s was instrumental in developing the theory of stochastic systems for use as models to explain diverse phenomenon in the engineering and physical sciences. Much of this theory was based in the “frequency domain” rather than in “time domain” and the completion of this theory was hampered by the difficulties encountered with inverting the transforms involved. In addition, the theory was probability based and lacked statistical methods that would allow statistical inference to enter into the framework. This talk argues that this original work finds its natural completion when two tools are added into the existing methods: saddlepoint approximations, for fast and accurate transform inversion, and the bootstrap, for nonparametric statistical inference. Saddlepoint approximations are also needed to make the bootstrap methods practically feasible. Multi-stage survival models provide the primary examples that will be used to show how such methods derived in EE can be used in a biostatistical setting to explain the dynamics of degenerative diseases. The patient is essentially modelled as a “charge” passing through a electrical DC circui whose sink(s) are various sorts of death states. Complete statistical inference about a patient with the degenerative disorder requires the use of the double bootstrap. Practical implementation of the double bootstrap in all but the simplest models requires the use of saddlepoint methods in lieu of the resampling effort at the inner stage of resampling.

Some Recent Developments in Bayesian Model Selection

James Berger, ISDS, Duke University, and SAMSI

We review two fairly recent developments in Bayesian model selection:

1. When the space of models is large, Search Strategies need to be carefully developed for exploration of the space. One successful strategy in variable selection is to perform a stochastic search that (roughly) adds or removes variables based on their current estimated posterior inclusion probabilities. This approach, and related diagnostics, will be illustrated. An interesting phenomenon that seems to be frequently encountered is that no model receives significant posterior probability, so that the meaning of model selection and the questions we pose concerning models may need to be reconsidered.

2. A generalization of BIC has recently been developed, for contexts common in the social sciences, that appropriately assesses the dimension of a model and the effective sample size for each parameter in a model. The generalization allows for the model dimension to grow with the sample size.

Estimating Large Covariance Matrices

Peter Bickel, University of California at Berkeley

We discuss different notions of sparsity for covariance matrices and the different assumptions and goals underlying these. We proceed to introduce an OO dimensional "nonparametric" model for covariances. We generalize in this context a result of Bickel and Levina (2004)Bernoulli on naive Bayes rules and show that if log (Dimension)/sample size ->0 covariance matrices and their inverses can be estimated consistently in a suitably uniform way by matrices requiring on the order of np rather than np^2 operations.

Poster Abstracts

Restricted Likelihood Inference for GeneralizedLinear Mixed Models

Alessandra R. Brazzale and Ruggero Bellio
National Research Council Padova and University of Udine, Italy

Restricted Maximum Likelihood (REML) is probably the most widely used criterion for making inference on the variance components of a linear mixed model. REML estimators have several desirable properties such as a more reliable finite-sample behaviour and a lower sensitivity to influential observations. Furthermore, several of the commonly used statistical software packages implement a version of restricted maximum likelihood. The REML criterion can be defined following several approaches, all of which give rise to the same analytical expression for the residual likelihood function. Indeed, the restricted likelihood can be seen as a conditional or marginal likelihood, as a Bayesian posterior density or, borrowing from the theory of likelihood asymptotics, as a modified profile likelihood. The extension of the REML criterion to generalized linear mixed models GLMMs) is not straightforward. The marginal and conditional arguments, which are advocated to derive the restricted likelihood in the linear case, are no longer meaningful in this context. Bayesian inference is in most cases too demanding as it involves integrals of likelihood quantities over several dimensions and, furthermore, depends on the choice of the prior distribution for the fixed effects. We promote the use of the modified profile likelihood function as criterion for estimating the variance parameters of a GLMM. As it is the case for the Bayesian approach, the major difficulty is the efficient calculation of the multidimensional integrals involved, although there are typically less random effects than fixed ones. Our fitting routine exploits a suitable implementation of quasi Monte Carlo integration, which turns out to be by several orders of magnitude faster than competing methods such as Monte Carlo integration and Gauss-Hermite quadrature. The properties of the resulting estimators are explored in a number of simulation studies. Applications with real data will be given. Attention is focused on regression models with binary response.

Error Rate Estimation for Bayesian Support Vector Regression Model

Sounak Chakraborty, University of Missouri

Support vector machines (SVM) originally developed by Vapnik (1995) are a system for efficiently training the linear learning machines in the kernel induced feature space. It has gained popularity due to its attractive, analytic and computational features, and promising empirical performance. The formulation embodies the Structural Risk Minimization (SRM) principle which has been shown to be superior to the Empirical Risk Minimization (ERM), used by other conventional methods.

Our goal is to accurately estimate the error in any prediction of a support vector regression (SVR) model. We propose a probabilistic regression framework for this purpose which can also be used for any other basis function regression models such as nonlinear ridge regression. The purpose is to estimate the probability that the predicted output ˆy is in the _-neighborhood of the original value, P(d(ˆy, y) _ _). In this paper, we describe the well-known relationship between Gaussian Processes (GP) and Support Vector Machines (SVM) and obtain a simple formula for the error rate estimation. We will also establish a result which indicates that as the number of samples in the training set increases estimates of P(d(ˆy, y) _ _) will approach the true value. Some initial simulation results indicate that our method is successful in simultaneously controlling mean square error rate.

On Posterior Consistency in Nonparametric Regression Problems

Taeryon Choi, University of Maryland
Mark J. Schervish, Carnegie Mellon University

We give sufficient conditions in order to establish posterior consistency in nonparametric regression problems with Gaussian error when suitable prior distributions areused for the unknown regression function and the noise variance. When the prior under considertation satisfies certain properties, the crucial condition for posterior consistency is to construct tests that makes the true parameter be separated from the outside of the suitable neighborhoods of the parameter. Under appropriate conditions on the regression function, we show there exist tests, of which the type I error and the type II error probabilites are exponentially small for distinguishing the true parameter from the complements of the suitable neighborhoods of the parameter. These sufficient conditions enable us to establish almost sure consistency based on the appropriate metrics with multi-dimensional covariate values fixed in advance or sampled from a probability distribution. We consider two examples of nonparametric regression problems.

Higher-Order Asymptotic Normality Of Approximations To The Modified Signed Likelihood Ratio Statistic For Regular Models

Heping He and Thomas A. Severini
University of Chicago and Northwestern University

Approximations to the modified signed likelihood ratio statistic are asymptotically standard normal with error of order n^{-1} where n is sample size. Proofs of this fact generally require that the sufficient statistic of the model be written as $(\hat{\theta}, a)$ where $\hat{\theta}$ is the maximum likelihood estimator of parameter θ of the model and a is an ancillary statistic. This condition is very difficult or impossible to verify for many models. However, calculation of the statistics themselves do not require this assumption. This paper is devoted to exploring higher-order asymptotic normality of these statistics under general conditions. It focuses on the case that θ may be parameterized as $\theta = (\psi, \lambda)$, where ψ is the scalar parameter of interest and λ is a nuisance parameter vector. Under general assumptions, the asymptotic properties of the statistics are proved. These proofs do not put any requirements of the sufficient statistic, they just assume general conditions which are easy to verify for and satisfied by commonly used models. Therefore this research removes the theoretical obstacle for applying these statistics to commonly used models.

Uniform Confidence Sets for Densities with Componentwise Shrinkage Estimators

Woncheol Jang, Institute of Statistics and Decision Sciences, Duke University

In most statistical problems, often we need to address uncertainty of an estimator to make effective inferences. Beran and D¨umbgen (1998) show how to construct uniform confidence sets for regression using modulation estimators, componentwise shrinkage estimators. Their confidence set provides (asymptotically) uniform coverage for the whole function. Inferences can then be generated by searching this set, possibly with added constraints from available side information. We extend results by modulation estimators to density estimation. The confidence set is obtained by showing that a pivot process, constructed from the quadratic loss function, converges uniformly to a Gaussian process. Inverting this pivot yields a confidence sets for functionals of the density.

Keywords and Phrases: Confidence Sets, nonparametric density estimation, shrinkage methods, empirical processes.

References:
1. Beran, R (2000). REACT Scatterplot Smoothers: Supereffciency through Basis Economy. Journal of American Statistical Association, 63, 155-171.
2. Beran, R., and D¨umbgen, L. (1998). Modulation of Estimators and Confidence Sets. Annals of Statistics.,26, 1826-1856.
3. Jang, W. (2005). A Functional Central Limit Theorem for Sums of Weakly Dependent Processes.
4. Stein , C. (1981). Estimation of the mean of a multivariate normal distribution. Annals of Statistics, 9, 1135-1151.

Testing For The Equality Of Two NonparametricRegression Curves With Long Memory Errors

Fang Li, Indiana University-Purdue University, Indianapolis

This paper discusses the problem of testing the equality of two nonparametric regression functions against two-sided alternatives for uniform design on [0,1] with long memory moving average errors. The standard deviations and the long memory parameters are possibly different for the two errors. The paper adapts the partial sum process idea used in the independent observations settings to construct the tests and derives their asymptotic null distributions. The paper also shows that these tests are consistent for general alternatives and obtains their limiting distributions under a sequence of local alternatives. Since the limiting null distributions of these tests are unknown, we first conducted a Monte Carlo simulation study to obtain a few selected critical values of the proposed tests. Then based on these critical values, another Monte Carlo simulation is conducted to study the finite sample level and power behavior of these tests at some alternatives. This power is found to be high for the chosen values of the long memory parameters below 0.75. The paper also contains a simulation study that assesses the effect of estimating the nonparametric regression function on an estimate of the long memory parameter of the errors. It is observed that the estimate based on direct observations is generally preferable over the one based on the estimated nonparametric residuals.

Research partly supported by the NSF grant DMS 0071619.
1990 IMS Subject Classification: Primary 62M10, Secondary 62F03.
Key words and Phrases: Partial sum process, fractional Brownian motion, Monte Carlo simulation.

Robust And Efficient Estimation Under Data Grouping

Nan Lin, Washington University, St. Louis

The minimum Hellinger distance estimator is known to have desirable properties in robustness and efficiency. We propose an approximate minimum Hellinger distance estimator by adapting the approach to grouped data from a continuous distribution. It is easier to compute the approximate version for either the continuous data or the grouped data. Given certain conditions on the model distribution and reasonable grouping rules, the approximate minimum Hellinger distance estimator is shown to be consistent and asymptotically normal. Furthermore, it is robust and can be asymptotically as efficient as the maximum likelihood estimator. The merit of the estimator is demonstrated through simulation studies and real data example.

Sequential Estimation of Autoregressive Parameters in General Vector Autoregressive Model

Indranil Mukhopadhyay, University of Pittsburgh
A.K.Basu,University of Calcutta

This work centers on the problem of sequential estimation of the autoregressive parameters in a p-th order vector autoregressive model. The sequential estimator proposed here is based on the least square estimator and is shown to be asymptotically risk efficient as the reciprocal of cost of estimation error tends to infinity, under certain regularity conditions. The asymptotic normality and uniform integrability of the standardized stopping time are established. A simulation study is also being conducted to show the performance of the sequential estimator proposed here.

Use of Orthogonal Polynomials in Density Approximation

Deepak Sanjel, McMaster University

Often the exact moments of a continuous distribution can be explicitly determined, however, its density function can not be expressed in simple mathematical form. In such situations, the density approximants can be obtained by using the orthogonal polynomials such as the Legendre, Laguerre, Jacobi and Hermite polynomials. The application will be illustrated using the various test statistics in order to detect the outliers in gamma and exponential samples, Durbin-Watson statistic and some mixture distributions. The density approximants will be verified using Monte Carlo simulations. Orthogonal polynomials have been scarcely discussed in the statistical literature in connection with the approximation of distributions. The proposed methodology of density approximation applies to a very wide array of distributions; moreover, their accuracy can be improved by making use of additional moments.

References:

[1] Sanjel, D. Provost, S. B. and MacNeill, I. B. (2005). On Approximating the Distribution of an Alternative Statistic for Detecting Serial Correlation, Journal of Probability and Statistical Science, Vol. 3(2), pp. 229-239.
[2] Sanjel, D. and Balakrishnan, N. (submitted, 5/2005). A Laguerre Polynomial Approximation for a Test of Exponentiality Based on Progressively Type-II Right Censored Data, Journal of Statistical Computation and Simulation.
[3] Sanjel, D. and Balakrishnan, N. (submitted, 10/2005). A Laguerre Polynomial Approximations for Interval Estimation and Prediction for Exponential Distribution Based on Doubly Type-II Censored Data, Computational Statistics & Data Analysis.
[4] Provost, S. B. and Sanjel, D. (2005). Inference About the First-Order Autoregressive Coefficient, Communications in Statistics-Theory and Methods, 35 , 1183-1201.
[5] Sanjel, D. and Balakrishnan, N. (submitted, 11/2005). Jacobi and Laguerre Polynomial Approximations for the Distribution of Statistics Useful in Testing for Outliers in Exponential and Gamma Samples. P*robability and Statistics Letters*

Semilinear Stochastic Differential Equations In Hilbert Spaces Driven By Non-Gaussian Noise And Their Asymptotic Properties

Li Wang, Oregon State University

A class of stochastic evolution equations with additive noise (compensated Poisson random measures) in Hilbert spaces is considered. We first show existence and uniqueness of a mild solution to the stochastic equation with Lipschitz type coefficients. The properties (homogeneity, Markov, and Feller) of the solution are studied. We then study the stability and exponential ultimate boundedness properties of the solution by using Lyapunov function technique. We also study the conditions for the existence and uniqueness of an invariant measure associated to the solution. At last, an example is given to illustrate the theory.

Confidence Bands For Regression Curve Under Weak Dependence

Li Wang, Michigan State University

Many types of confidence bands have been proposed for nonparametric regression constructed from i.i.d. data. In this paper, asymptotically simultaneous conservative confidence bands are obtained for nonparametric regression function obtained from geometrically alpha-mixing samples, based on piecewise constant and piecewise linear polynomial spline estimation, respectively. Simulation experiments and examples have provided strong evidence that corroborates with the asymptotic theory. An EKC example has been given to show the possible application of this method.

Nonparametric Inference for High Dimensional Longitudinal Data

Ke Zhang and Haiyan Wang, Kansas State University

High dimension, low sample size data have received increasing attention by statisticians in recent years. It is more intriguing when the data are correlated. In this paper, we consider the data structure that the sample size is fixed, as the number of groups tends to infinity, and each subject is measured repeatedly over time. A hypothesis test for main effect is proposed based on both the original observations and (mid-) ranks. The asymptotic distribution of the test statistics under null hypothesis is established through a modified Hájek's projection. Simulation results show that the rank statistics outperform those based on the original observations if the normality assumption is not held. As an application of the proposed test, a lipidomics dataset are analyzed to study whether the profiles of lipid molecule species are significantly different. The proposed test will be further utilized in a clustering algorithm to identify lipid species with significantly different contents under various treatment conditions.

Detection of Sparse Normal Mixture: Effect of Unequal Variances

Xinge Zheng, Purdue University

Higher Criticism is a statistic recently proposed in [1], where it has been shown to be effective in resolving the sparse normal mean problem: test whether n normal means are 0 or a small fraction are nonzero.

In this talk, we consider a testing problem whether n samples are truly from N(0,1) or a two component normal mixture where two components may have different means as well as different variances. In detail, suppose we have n iid samples $X_i \sim (1 - \varepsilon_n)\, N(0,1) + \varepsilon_n\, N(\mu_n, \sigma^2)$, where (ε_n, μ_n) depend on n but not on i, and σ^2 is a constant. We are interested in the detection problem in which we test $\varepsilon_n = 0$ vs. $\varepsilon_n > 0$, and are particularly interested in the influence of the unknown parameter σ^2 on the testing problem.

We will describe the precise demarcation between detectable and undetectable: for which $(\varepsilon_n, \mu_n, \sigma)$ it is possible to reliably tell $\varepsilon_n > 0$ and for which it is impossible to do so. Particularly, $\sigma^2 = 2$ is the separating line for the influence of σ^2. For the same ε_n, when $\sigma^2 < 2$, σ^2 alone is not able to decide whether it is detectable or not, and μ_n always has its share of influence; while when $\sigma^2 > 2$, σ^2 alone can decide whether it is detectable or not, and different μ_n won't make any difference.

We will also compare our result with that in [1]. The case $\sigma^2 = 1$ has been studied in detail in [1], where it has been showed that the Higher Criticism is optimally adaptive to the unknown level of sparsity and unknown mean (ε_n, μ_n). Surprisingly, we found that the Higher Criticism is also optimally adaptive to the unknown variance σ^2.

This is in collaboration with Jiashun Jin.

[1] Donoho, D. and Jin, J. (2004). Higher Criticism for Detecting Sparse Heterogeneous Mixtures. *Ann. Statist.,* Vol 32, **3**, 962-994.

Statistical Analysis of Diffusion Tensor Images

Hongtu Zhu, Columbia University Medical Center

This paper presents a framework for use of the parametric models in constructing diffusion tensor (DT) images, and it establishes the validity of statistical inferences in diffusion tensor imaging (DTI). It will be shown that a Gaussian distribution is a reasonable representation of the estimated diffusion tensor, even though the postulated model is misspecified. Estimation procedures are developed for the Rician model and two other normal models of noise. In particular, an Expectation and Maximization (EM) estimation algorithm is proposed to maximize the likelihood function of the Rician model. Theoretical results further reveal that a conventional scheme for the acquisition of DT images may undermine statistical accuracy when estimating diffusion tensors, even when the number of diffusion weighted acquisitions is increased to infinity. A Bayesian design criterion and a gradient sampling index (GSI) are proposed to quantify, respectively, the noise characteristics of data acquisition and gradient sampling schemes for DT images. Optimization of the Bayesian design criterion and the GSI leads to accurate estimation of the DT. This paper also delineates the asymptotic distributions of four invariant measures of the diffusion tensor as well as for two measures of anisotropy. A scaled chi-squared distribution in isotropic tissue is proposed to approximate the asymptotic distribution of fractional anisotropy (FA), which provides a simple means of testing statistically whether the DT at each voxel is significantly anisotropic. Simulations characterize the bias of the estimated DTs under models for noise that are misspecified. Quantitative assessments confirm that the noise characteristics of differing schemes for acquisition of DT images are significantly associated with the GSI, and they suggest empirically that the scaled $\chi 2$ distribution proposed for FA has high statistical power for detecting anisotropy, while maintaining Type I error in isotropic tissues.

Key words and phrases: diffusion tensor; fractional anisotropy; gradient sampling index; misspecification; sampling acquisition scheme; sorting bias.

Map

United Church of Gainesville
1624 N.W. 5th Avenue
Gainesville, FL 32603