Developments and Challenges in Mixture Models, Bump Hunting and Measurement Error Models

POSTER SESSION
June 2, 12:30-2:30 p.m.

Tumor Classification by Mixture Modeling of Gene Expression Data

Roxana Alexandridis, Shili Lin
Ohio State University

Accurate classification of tumor samples is an essential tool in cancer diagnosis and treatment. The DNA microarray technology enables successful cancer classification based only on gene expression analysis, without prior biological insights. We present a new method for disease classification based on microarray gene expression levels. Most of the thousands of genes whose expression levels are measured do not contribute to the separation between types of tumors/classes. We address the issue of gene selection by using several measures of expression ordering and a mixture modeling approach with an unknown number of components. For common measures of gene expression ordering, the number of genes selected in a classifier is achieved by cross-validation. For the mixture modeling approach, genes in a classifier are selected based on their posterior probabilities of belonging to the mixture components that correspond to differential expressions among the classes. A classification scheme is then applied to the selected genes, also based on mixture modeling of their expression patterns. Class assignment for a test sample is based on the posterior probabilities of class memberships. Preliminary results show that the methods work well for the leukemia data of Golub et al. (Golub et al., 1999), giving accurate cross-validation and prediction results.

An Alternative Approach to the A-test for Obtaining Unbiased Estimates of Linkage Parameters in the Presence of Heterogeneity

Swati Biswas and Shili Lin
Department of Statistics, Ohio State University

Linkage heterogeneity is often seen in complex genetic diseases. In some families the gene of interest causes the disease. These families are referred to as linked families. However, in other families (unlinked families) some other gene and/or nonhereditary factors cause the disease. A widely used method for analyzing linkage data when heterogeneity is suspected is the admixture test (A-test). It is based on maximizing the mixture log likelihood, where the mixing (heterogeneity) parameter represents the proportion of linked families in the sample. The main parameter of interest is the recombination fraction in the linked families and the heterogeneity parameter can be treated as a nuisance parameter. But, in general, this heterogeneity parameter varies across different types of families. In particular, Janssen et al. (Hum Hered 1997;47:223-233) show that the estimates used in A-test are biased when the linked and unlinked families do not have the same size. They use Effective Number of Informative Meiosis (EFNIM) as a measure of family size, i.e., linkage information in a family. They further say that if the EFNIM distributions of the linked and unlinked families are the same, then the estimates have little or no bias. We show that this is not true both theoretically as well as using simulations. To get unbiased estimates, the distributions of different types of families in the linked group and the unlinked group should be the same. This condition does imply the EFNIM condition but the converse need not hold.

An alternative approach to get around this problem of bias is to form groups of families that have same distributions. For each group we use a different parameter to denote the level of heterogeneity in that group. We consider the mixture likelihood using this approach and apply three maximization algorithms to obtain the maximum likelihood estimates. These algorithms are Expectation Maximization (EM), Classification Expectation Maximization (CEM) and Stochastic Expectation Maximization (SEM). Our simulation study shows that both the EM and SEM algorithms give satisfactory results while CEM algorithm does not. An advantage of using EM algorithm is that we get the standard error of the estimates directly by applying methods available in the literature. These simulations also show that when different types of families have different levels of heterogeneity, this approach gives unbiased estimates in contrast to the A-test approach.

The Use of Microarray Techniques to Study Normal and Pathogenic Mechanisms of Neuronal Copper Metabolism: Implications for Understanding Alzheimers Disease

Hsien W. Chan, Tianbing Liu, Glenda M. Bishop, Mark A. Smith, George Perry and Craig S. Atwood
Institute of Pathology, Case Western Reserve University

Alzheimers disease (AD) is the most common form of dementia in patients older than 40 and is characterized by neuronal cell death, synapse loss and the deposition of amyloid plaques and neurofibrillary tangles. Accompanying these
pathological features is evidence of damage to biological macromolecules as a result of oxidative stress. The redox-reactive properties of copper make it an essential element in iron transport and also in the functioning of enzymes such as cytochrome c oxidase, superoxide dismutase (SOD) and lysyl oxidase. However, in excess, copper induces the formation of free radicals via Fenton and Haber-Weiss chemistry, leading to lipid peroxidation, amino acid oxidation as well as DNA damage. The presence of oxidative stress markers in AD, the role of copper in the formation of oxygen free radicals, the fact that the concentration of copper is markedly elevated in the AD compared with age-matched control brain and the presence of copper binding sites on the amyloid precursor protein suggest that copper plays an important role in the pathogenesis of AD.

Through the use of microarray analytical techniques to analyze the effect of copper on M17 human neuroblastoma cells, we plan to examine the mechanisms by which neurons respond to copper by measuring changes in gene expression.

Genes that are likely to be affected by the presence of copper include those involved in intracellular copper transport, copper storage and antioxidant defense systems. Intracellular copper transport proteins identified in yeast have been found to be highly conserved in humans, although not all human homologs have been identified.

It is currently believed that cellular uptake of copper is facilitated by the copper specific transporter, hCTR1 which is located on chromosome 9q31/32. Copper is then distributed by the protein chaperones, Atox-1, Cox17p and CCS. The destinations of copper bound to Cox17p and CCS are cytochrome c oxidase and superoxide dimutase-1 respectively. The Atox-1 chaperone transports copper to ATP7A and B, and has been shown to promote neuronal survival during serum starvation and oxidative stress. Other known mechanisms of coping with copper load include the induction of Cu/Zn-SOD to help protect cells from oxidative damage as well as the increased expression of the MTF-1 zinc finger transcription factor which leads to the upregulation of metallothionein I and II, although it is not known how copper is stored in brain. In addition to these known human genes, we expect to identify novel genes involved in the modulation of copper metabolism.

Through the use of microarray techniques to analyze the genetic expression of the aforementioned copper proteins, and any new proteins, we hope to determine the basic mechanisms of neuronal copper metabolism. Examination of the compensatory mechanisms utilized by neurons in handling excess copper will provide insights into the role of copper in the pathogenesis of AD and potential new therapies for the treatment of this disease. Our in vitro model system also can be used to examine altered copper metabolism in other diseases, such as Menkes disease and Wilson=92s disease, that result from mutations in the genes encoding the copper transport proteins ATP7A and B ATPases, respectively. Altered copper metabolism also is thought to play a role in other neurodegenerative diseases such as Creutzfeldt-Jakob disease, Parkinson=92s disease and amyotrophic lateral sclerosis. Therefore our studies have implications not only for understanding normal copper homeostatic mechanisms, but also in understanding the basic pathogenic mechanisms involved in a number of copper-related diseases.

Acknowledgement: This research is funded by the National Institute of Health (NIA #R016426337)

Testing Homogeneity in a Mixture Distribution via \( L^2 \)-type Statistics

Richard Charnigo and Jiayang Sun
Case Western Reserve University

Testing for the number of components in a mixture has long been an interesting and challenging problem for statisticians. Chen, Chen and Kalbfleisch (2001) recently proposed a modified likelihood ratio test (MLRT), which is distribution-free and locally most powerful, asymptotically. In this paper we present a new method for testing whether a finite mixture distribution is homogeneous. The test is based on a simple \( L^2 \)-type distance between a fitted homogeneous model and a fitted heterogeneous model. For mixture components from standard distributions, our test statistic \( d(k) \) has closed-form expressions, whereas likelihood ratio-type test statistics do not. Thus, our test has potential for data mining applications. The \( d(k) \) test is competitive with the MLRT when the mixture components are normal. The MLRT performs better when the mixture components are exponential, but in such a case the \( L^2 \) and visual separations between the homogeneous and heterogeneous models are extremely small. Thus, it is proposed that a reasonable transformation of the data be considered before applying the \( d(k) \) test. Such a transformation - which is equivalent to altering the measure underlying the \( L^2 \) norm according to a weight function - renders the \( d(k) \) test more competitive in the aforementioned case. The convergence rate of the \( d(k) \) test under a null hypothesis of homogeneity is established. We conclude by applying our method to a data set in which the observations are measurements of firms’ financial performances.

A histogram smoothing estimate of the multivariate density level set

Nidhan Choudhuri
Case Western Reserve University

Often, large scale multivariate data are encountered where it is unreasonable to assume any parametric structure. Analyzing this data requires some nonparametric tool. A commonly used approach is to study the multivariate density function generating the data through a nonparametric density estimate. A relatively newer approach is to study the density level sets through their nonparametric estimates. This work presents a new method of estimating multivariate density level sets, which is based on histogram smoothing of the data. This method is easy to execute, has computational advantage over the existing methods and thus may be applied to fairly large data sets. This work also finds the
rate of convergence of these estimates, and the conditions required on both the level sets and the underlying density function to achieve this rate. Attention is given in finding a data driven value of the smoothing parameter to achieve the optimal rate of convergence. The illustration of this method is then done on some two dimensional data sets.

Statistical Analysis of Life-Times with Two Types of Failure with Mixture Models and Likelihood Based Methods

Elomsa Dmaz-Francis
Centro de Investigacion en Matematicas

Log lifetimes, assumed to come from two exclusive types of failure, are modelled with a mixture of two extreme value (Gumbel) distributions. Inferences about the parameters based on the use of the relevant full profile likelihood functions are given in terms of likelihood-confidence intervals. Based on these, various aspects of the model are assessed and suitable parameterizations can be suggested. The logical distinction between the roles played by the mixing proportion parameter \( p \) vs. the location-scale parameters is exhibited. Aircraft component failure times, which were censored at a fixed test termination time, are used to illustrate these ideas.

Acknowledgement: This is joint work with Dr. David A. Sprott

Optimal Adaptivity of Higher Criticism

David Donoho, Jiashun Jin
Statistics Department, Stanford University

Higher Criticism, or second-level significance testing, is a multiple comparisons concept mentioned in passing by Tukey (1976). It concerns a situation where there are many independent and unrelated tests of significance. Tukey suggested to compare the fraction of observed significances at a given \( \alpha \)-level to the expected fraction under the joint null, in fact he suggested to standardize the difference of the two quantities and form a z-score; the resulting z-score tests the significance of the body of significance tests.

We consider a generalization, where we maximize this z-score over a range of significance levels \( 0 < \alpha < \alpha_0 \). We are able to show that the resulting Higher Criticism statistic is effective at resolving a very subtle testing problem where we are testing whether \( N \) normal means are all zero versus the alternative that a very small fraction is nonzero.

The subtlety of this ‘sparse normal means’ testing problem can be seen from work of Ingster (1999) and Jin (2002), who studied such problems in great detail. In their studies, they identified an interesting range of cases where the small fraction of nonzero means is so small that the alternative hypothesis exhibits little noticeable effect on the distribution of the body of significance tests given by

\[
H_0: \theta_1, \theta_2, \ldots, \theta_m \\
H_1: \theta_1 = \theta_2 = \cdots = \theta_r = 0, \theta_{r+1}, \ldots, \theta_m \text{ are all nonzero}
\]

We show that the higher criticism is successful throughout the same region where the precisely specified likelihood ratio test would succeed, and thus, higher criticism is in a sense optimally adaptive to unknown sparsity and size of the non-null effects. While our theoretical work is largely asymptotic, we provide simulations in finite samples, and suggest some possible applications.

Acknowledgement: The authors would like to thank Yoav Benjamini and Iain Johnstone for extensive discussions, references, and encouragement. This work has been partially supported by National Science Foundation grants DMS 00-77261, and DMS 95-05151, and by AFOSR MURI 95-P40620-96-1-0028. This is joint work with David L. Donoho.

Trends in the Times Between Maximum Brightnesses for Variable Stars

Mickey Dunlap, Jeffrey D. Hart, and Cai-xia Zhao
Texas A&M University

Of interest to astronomers is determining if times between maximum brightnesses of a variable star display a trend over time. We investigate this question using data from 378 variable stars. The problem is considered from two points of view: (1) We assess the evidence for trends on a star by star basis, and (2) we attempt to cluster stars according to the types of trends they exhibit. Both stochastic and deterministic trend models are entertained, and a variety of tools are utilized, including smoothing, time series analysis, deconvolution and mixture models.

Acknowledgement: This work was supported in part by NSF Grant DMS 99-71755.

Estimating Finite Mixture Cumulative Density Functions Using Multinomial Mixtures

Ryan T. Elmore
The Pennsylvania State University

Consider a two-component finite mixture model

\[
f(x) = 3D\lambda f_1(x) + (1 - \lambda) f_2(x)
\]

with continuous component densities \( f_j(x), j = 1, 2 \) and weights \( \lambda, 1 - \lambda \). Consider a repeated measures setting, such as a psychological assessment, where it is known that different children employ different cognitive strategies to solve the same cognitive problem (e.g., Thomas & Horton, 1997). In these settings different mixture components may be usefully regarded as representing different cognitive strategies. Furthermore, it is often plausible to assume that each individual can provide \( j = 3D1, 2, \ldots, m \) (conditionally) iid responses from component \( r \), say, where the \( m \) responses are to essentially equivalent cognitive problems of constant difficulty. Wanting to avoid specifying the shape of the components \( f_r(x) \), Hettmansperger and Thomas (2000) develop a semiparametric approach to inference for this model. They map \( X_i \), the vector with subject \( i \)'s \( m \) responses, into a binary vector according to

\[
Y_{ij} = 3D1_{[X_i \leq c]} \text{ for some } c \text{ in the support of } X.
\]

Summing over \( j \) yields observations from the binomial mixture given by

\[
g(y) = 3D\lambda b(y; \theta_1, m) + (1 - \lambda) b(y; \theta_2, m)
\]
\[ b(y; \theta_r, m) \] is a binomial mass function with success parameter \( \theta_r = 3D \int_{\infty}^{\infty} f_r(x) \, dx = 3DF_r(c) \). Note that \( g(y) \) has the same number of components and same component weights as \( f(x) \).

We extend these results by mapping the original vectors \( X_i \) into \((R + 1)\)-dimensional vectors \( Y_i = 3D \sum_{j=1}^{m} D_1 \left( I[X_i \leq c_1], I[x_1 \leq c_2], \ldots, I[x_R \leq c_R] \right)^t \). The observations \( Y_i \) are distributed as a mixture of two multinomial distributions with the same mixing proportion \( \lambda \), see Cruz-Medina (2000). We exploit this discretization and show how to estimate the unknown component distribution functions \( F_r(x) \) in the finite mixture setting.

**Acknowledgement:** This is joint work with Hoben Thomas and Tom Hettmansperger. Supported by NSF grant SES0115619.

Jackkniﬁng Bias of the Box-type Approximation under Dependence

George Zhaozhi Fan  
Department of Statistics, Case Western Reserve University

In statistical analysis of factorial designs, we often use the so called Box-type approximation of the quadratic forms of independent normal random variables to test corresponding statistical hypotheses. The underlying quadratic forms can be approximated by a scaled \( \chi^2 \)-distribution. This kind of approximation performs worse in dependent case. It leads to conservativity of the underlying testing procedures. This conservativity can be greatly improved by reducing the biases of the estimates of the scale coefficient and the degrees of freedom of the scaled \( \chi^2 \)-distribution. Delete-1 jackknifing technique is used in reducing the biases of the corresponding estimates. Simulation study has been done in different correlation situations.

**Acknowledgement:** This is joint work with Professor Edgar Brunner (University of Göttingen, Germany)

Confidence Intervals for Relative Risks in Disease Mapping

B. Ibáñez  
Statistics and Operational Research Department, Public University of Navarra

Several analysis of the geographic variation of mortality rates in space have been proposed in the literature. Poisson models allowing the incorporation of random effects to model extra-variability are widely used. The typical modelling approach uses normal random effects to accommodate local spatial autocorrelation. Instead, when spatial autocorrelation is absent but overdispersion persists, a discrete mixture model is an alternative approach. However, a technique for identifying regions which have significant high or low risk in any given area has not been developed yet when using the discrete mixture model. Taking into account the importance that this information provides to the epidemiologists to formulate hypothesis related to the potential risk factors affecting the population, different methods for obtaining confidence intervals for relative risks are derived in this paper. These methods are the standard information-based method and both the asymptotic-bootstrap and percentile-bootstrap methods. All of them are compared empirically by their application to mortality data due to cardiovascular diseases in women from Navarra, Spain, during the period 1988-1994. In the small area example considered here, we find that the asymptotic methods are sensible at estimating standard errors of the component means in the discrete mixture model but are not appropriate for providing standard errors of SMRs and hence, for constructing confidence intervals for the relative risk associated to each region. Therefore, the percentile-bootstrap method is recommended for this matter.

**Acknowledgement:** This is joint work with A. F. Militino and M.D. Ugarte. The authors would like to thank the Non-transmissible Diseases and Vital Statistics Section of the Institute of Public Health in Navarra for providing the data. The work has been supported by the Health Department of the Government of Navarra.

Iid Monte Carlo Algorithms for Semiparametric Linear Mixed Models

Hemant Ishwaran and Glen Takahara  
Cleveland Clinic Foundation and Queen’s University

Hybrid versions of iid weighted Chinese restaurant (WCR) seating rule algorithms are developed for inference in semiparametric linear mixed models under minimal assumptions for the random effects distributions. The WCR method of working with the posterior partition structure leads to Rao-Blackwell estimates for higher order moments of random effects, such as skewness and kurtosis, and can be used to estimate densities for random effects. A key feature is our hybrid approach of incorporating external estimates into the WCR seating rule. The use of such information leads to simplified computational procedures, reduces the amount of user input required for specifying models and simultaneously improves numerical stability and accuracy.

Nonparametric Density Estimation and Clustering in Cosmology: Large-scale Structure of the Universe

Woncheol Jang  
Carnegie Mellon University

Clusters of galaxies provide powerful tools to trace the large-scale structure of the universe. We present a clustering method based on nonparametric density estimation. Kernel smoothing and orthogonal series estimator are used to estimate the galaxy density \( f \) given \( z \), redshift and then the connected component of the level set \( \{ f(z) > \delta \} \) are extracted as clusters with a modified Cuevas et al (1998) algorithm. Confidence interval for \( \Omega_m \), density parameter for
matter, is derived using a goodness-of-fit criterion since $\delta_c$, present overdensity, is a function of $\Omega_m$ and $z$, therefore it can be estimated by matching the number density to Press-Schechter model. The confidence set for clusters is proposed based on Molchanov (1998).

**Acknowledgment** This is joint work with Larry Wasserman, Chris Genovese and Bob Nichol.

**Measurement Error Comparison Using Single Factor Analysis with Minimum Message Length Estimation**

**Murray Jorgensen**  
*University of Waikato*

We introduce an EM algorithm for Minimum Message Length (MML) estimation in the Single Factor Analysis or Comparative Calibration model as an alternative to the algorithm used by Wallace and Freeman (1992), and compare the resultant parameter estimates with Maximum Likelihood, Maximum Posterior, and Wallace and Freeman’s MML estimates in a simulation study.

The relationship between the two forms of Minimum Message Length estimation is of theoretical interest as Wallace and Freeman treat the unknown true measurements as parameters to be estimated and the EM algorithm treats them as unobserved latent variables.

Both algorithms involve a form of penalized maximum likelihood estimation, but the penalty term is much easier to derive in the EM approach.


**On Computing Information in Semiparametric Mixture Models**

**Mary Lesperance**  
*University of Victoria*

Semiparametric mixture models pose many interesting inferential problems. When the mixing distribution is unspecified, the parameter space is infinite-dimensional, and hence standard asymptotic results do not apply in general. Fisher’s information arises naturally in standard models as the inverse of the asymptotic variance of maximum likelihood estimators. In this work, we define what is meant by information in a semiparametric mixture model, and consider a conjugate gradient method for calculating it.

**Acknowledgement:** This is joint work with Bruce Lindsay, Pennsylvania State University

**Comparison of Replicate Functional Measurements: The Physical Sciences Challenge**

**Walter Liggett**  
*National Institute of Standards and Technology*

On the cutting edge of physical-sciences metrology is the measurement of functions. As in functional data analysis, functional measurements are observations that portray a function of a continuous independent variable. An example is measurement of a particle-size intensity. The challenge in functional measurement is replication, a cornerstone of science. The basic purpose of replication in metrology is discovery of unanticipated sources of variation. Such discovery can be framed as a hypothesis-testing problem in which the null hypothesis is that the difference between replicates can be attributed to the known sources of variation. Consider testing hypotheses about pairs of unnormalized histograms that contain particle counts for particle-size intervals. The analysis involves generalized-linear-model fitting of cubic splines with irregularly-spaced knots. Of interest is testing the null hypothesis that two sets of particle counts correspond to intensity functions that differ only by a scale factor and a constant shift in horizontal registration. An unknown smooth function is common to the two intensities. The alternative hypothesis is that in addition, the difference between the two intensities is also an unknown smooth function. We consider three approaches to knot placement. First is specification of so many knots that adequate representations of the unknown functions cannot be doubted. Second is data-driven choice of knots. Third is choice of knots based on prior knowledge of what intensity differences are plausible. For the data at hand, we show that specification of too many knots leads to tests with too little power and that data-driven knot selection can lead to false rejection of the null hypothesis. The data at hand seem to call for use of prior knowledge to construct a semiparametric model that incorporates the distinction between the two hypotheses in the parametric part.

**Acknowledgement:** This is joint work with Robert Fletcher of the Surface and Microanalysis Science Division at NIST.

**The Optimal Environmental Window in Marine Biology: Is Empirical Evidence Real?**

**Miguel Nakamura**  
*Centro de Investigación en Matemáticas*

One main hypothesis behind small pelagics’ (sardines and anchovies) wide fluctuations in abundance is the existence of a dome-shaped relationship between fish recruitment (number of new individuals in the population) and wind-related environmental variables (wind speed, upwelling, turbulence, and others). This so-called Optimal Environmental Window (OEW) concept has gained great acceptance among specialists. The basic idea is that larvae survival requires both adequate food concentrations and a stable ocean. Moderate wind levels (the OEW) induce a balance between the enrichment
effect of wind-driven upwelling (and therefore increased food availability) and the adverse effects of turbulence. Extreme values are thus expected to decrease recruitment, while an intermediate (optimal) range of values is assumed to exist at a wind speed of about 5–6 m/s.

While the theoretical aspects of this argument have been largely discussed, its statistical examination is relatively recent as it demanded non-parametric, exploratory methods to detect the two relevant aspects of the OEW structure: the dome-shape aspect of the relationship and the location of the dome. Empirical evidence for OEW’s, appearing in the biology literature, is mainly based on estimated transformations via the Alternating Conditional Expectations (ACE) Algorithm. The OEW is thought to be evidenced when an estimated transformation for the environmental variable results in a dome. However, typical data in these settings are weakly correlated, so that well-known shortcomings of ACE are in order. Alternative exploratory devices such as SiZer, do not always show significant evidence of dome-shaped behavior.

Some researchers have purported ACE results as conclusive, and furthermore have claimed that evidence suggests location of the OEW (at 5–6 m/s) is universal. This exemplifies the need of proper methodology for addressing two problems with regard to a scientific hypothesis in an experimental setting: assessing the existence of a dome, and inferring about its location.

Acknowledgement: This is joint work with Daniel Lluch-Cota and Jesús Bautista-Romero, at Centro de Investigaciones Biológicas del Noroeste, La Paz, Mexico.

Distance-based Estimation of the Number of Components in Multivariate Mixture Models: A Tool For Analyzing Gene Expression Data

Surajit Ray
Pennsylvania State University

Multivariate mixture models provide a convenient method of density estimation, model based clustering and provide an excellent insight into the actual data generation process. But the problem of choosing the number of components (k) in a statistically meaningful way is still a subject of considerable research. Available methods for estimating k include optimizing AIC and BIC, gradient checking in a nonparametric mixture model setup and Bayesian approaches with entropy distances. In this paper we present rules for selecting k based on a one-sided non-parametric confidence-set generated by a quadratic distance measure. In this methodology the goal is to find the minimal number of components that are needed to adequately describe the true distribution. We also present results for selecting k based on a risk analysis that includes a penalty for overfitting. The goal here is to find the fitted mixture that is closest to the true distribution. Finally we fine tune our methods to analyze gene-expression data from micro-arrays, and compare them with other competitive methods.

Acknowledgement: This is joint work with Bruce Lindsay

Modeling Repeated Bivariate Ordinal Outcomes Subject to Informative Dropout

David Todem
Department of Statistics, University of Wisconsin-Madison

Longitudinal studies of cardiovascular diseases are usually based on the analysis of a bivariate vector comprising information on surrogate endpoints such as the hypertension and hypercholesteremia status. Additionally, it is often the case that subjects drop out of these studies prematurely, yielding unbalanced data with unequal numbers of measures for each subject. A statistical analysis of these data raises a number of challenging issues. It is well known, for example, that the clustering due to repeated observations from the same subject and the multiplicity of outcomes necessitate the use of methods for correlated data. The dropout process is another important issue. If the censoring process is related to the unseen values, estimates of the models that do not incorporate this mechanism may be misleading. We propose a likelihood-based approach that simultaneously models bivariate ordinal outcomes and the attrition process within a unified model-based framework. We first use the concept of a latent variable to derive the joint distribution of bivariate ordinal outcomes, and then extend the model to allow for longitudinal data. Second, we formulate a model for the dropout process, but its dependence to the data is captured through a shared random effects term. Assuming conditional independence given random effects, the marginal likelihood is approximated using an adaptive Gaussian quadrature to numerical integration. This method is illustrated using data from a randomized clinical trial on a cardiovascular educational program and compared to an estimating equations methodology that conditions on the dropout times but makes fewer assumptions about the distribution of data.

Acknowledgement: This is joint work with my thesis supervisor KyungMann Kim, Ph.D.

Density Estimation from Biased Data with Unknown Biasing Function and Memory Effect

Bin Wang and Jiayang Sun
Case Western Reserve University

Data from observational studies may come with selection biases. The problem of estimating the density of a population and its biasing function, based on one biased sample, both nonparametrically, is unidentifiable. However, if there are two biased samples, the nonparametric problem is identifiable. It is clear that a subject sampled before, if sampled again, is subject to a memory effect. This paper incorporates the memory effect into a general biased sampling model, with two biased samples from a finite population. Nonparametric estimators of the density, biasing and memory functions are then provided. An estimator of the population size follows directly from our density estimator. Asymptotic properties of these estimators are studied. Our procedures are compared with those ignoring the memory effect. Their applications are illustrated via simulated data and to a survey data set.
A Score Test for Mapping Quantitative Trait Loci with Sibships of Arbitrary Size When the Dominance Effect is not Negligible

Kai Wang and Jian Huang
University of Iowa

In the linkage analysis of quantitative traits, it is often assumed that the dominance effect of the trait does not exist or is negligible, and only the additive effect of the trait is considered in the analysis. Intuitively, when this assumption does not hold, a model that takes dominance effect into consideration would be more reasonable and should make more efficient use of the data. Since the size of the dominance effect is often unknown, it is not always appropriate to take this assumption as guaranteed. Here we introduce a score statistic for detecting quantitative trait loci when the dominance effect is not negligible. This statistic is derived from a normal likelihood function for data on sibships of arbitrary size. It is asymptotically equivalent to the corresponding likelihood ratio statistic, but it is much easier to compute. This statistic is applicable to sibships of arbitrary size. Genetic constraints on the model parameters are considered in the derivation of this statistic. The explicit asymptotic distribution of this statistic is derived, which is a mixture of $\chi^2_0$, $\chi^2_1$ and $\chi^2_2$, with the weights on distribution components a function of the informativeness of the marker data. The type I error rate and the power of the proposed statistic in finite sample are evaluated via simulations.

Acknowledgement: This work is supported, in part, by a University of Iowa College of Public Health - College of Medicine New Investigator Research Award (to K.W.), by National Institute of Mental Health grant K01-01541 (to J.H.), and by National Institute of Mental Health grant R01-52841.

Are Galaxies Clumped? A Statistical Analysis Using Finite Mixture Models

Liqun Wang
Department of Statistics, University of Manitoba

The Big Bang theory predicts that galaxies in the universe are clustered. We analyze a data set of recession velocities of 82 distant galaxies. Since the underlying distribution may have multi-modes, we use a model of finite mixture of distributions. The statistical analysis is complicated by the fact that the number of components in the mixture is unknown. Most previous analyses of this problem in the literature use normal mixtures for mathematical or computational convenience, though the data histograms show certain degree of skewness. It is well-known that the number of components tend to be over estimated, if symmetric distributions are used to model asymmetric components. In this paper, I use a mixture of inverse Gaussian distributions to model the data. The inference is done using a novel, non-iterative sampling procedure, which overcomes drawbacks and difficulties of Monte Carlo Markov chain (MCMC) algorithms.

Acknowledgement: This is joint work with James C. Fu. It is partially supported by the Natural Science and Engineering Research Council of Canada.

Mixture Density Estimation through Split and Merge Operations

Yanzhong Wang & Mike Titterington
Department of Statistics, University of Glasgow

Among techniques for learning mixtures, the most popular appears to be the EM algorithm. But as a local search algorithm, EM has a number of limitations such as slow to converge, sensitive to initialisation, and may get stuck in one of many local maxima of the likelihood function. To overcome these limitations, inspired by the idea of splitting and/or merging components sequentially or simultaneously based on certain criteria, a number of algorithms were introduced. We test 3 algorithms of this kind, namely, the IPRA algorithm (merge), the Greedy EM (split) and the SMEM (split and merge), on our simulated multivariate data set, and comment on their performances. Further, we extend the IPRA algorithm (Iterative Pairwise Replacement) into multidimensional problems by using minimum spanning tree to limit searches and updating the MST after each iteration. Simulation results show that the algorithm (without fine-tuning) is robust, less computational demanding, and provides better location for the estimated component centres.

Acknowledgement: This work was supported by a 2001 research scholarship from University of Glasgow.