PROGRAM

Sixth Annual Winter Workshop: Data Mining, Statistical Learning, and Bioinformatics

Department of Statistics
University of Florida
January 8-10, 2004

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

Alan Agresti
Jim Booth
George Casella
Mike Daniels
Malay Ghosh
Jim Hobert
André Khuri
Bhramar Mukherjee
Clyde Schoolfield

Invited Speakers

Keith Baggerly, MD, Anderson Cancer Center
William Cleveland, Bell Labs
Richard De Veaux, Williams College
Mario Figueiredo, Technical University of Lisbon, Portugal
David Hand, Imperial College
Alan Izenman, Temple University
Xihong Lin, University of Michigan
Giovanni Parmigiani, Johns Hopkins University
J. Sunil Rao, Case Western Reserve University
David Scott, Rice University
Scott Schmidler, Duke University
Rob Tibshirani, Stanford University

Acknowledgements

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

Conference Schedule

Thursday, January 8, 2004

7:00-10:00pm **Reception** (Keene Faculty Center, Dauer Hall)

Friday, January 9, 2004

8:00-8:30am **Breakfast** (Reitz Union Grand Ballroom)

All Sessions Meet in Reitz Union Grand Ballroom

8:30–10:30am **Session 1: Data Mining and Data Quality: An Overview**

Chair: George Casella

Speakers:
Dick De Veaux, "Data Mining: Where do we start?"

David Hand, "Data Quality"

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

11:00am–1:00pm **Session 2: Advances in Modeling Gene Expression Data**

Chair: Mark Yang

Speakers:
Giovanni Parmigiani, "Multilevel Models in Gene Expression Data Analysis"
Xihong Lin, "Semiparametric Regression for Microarray Gene Expressions: Support Vector Machines and Mixed Models"

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

2:30–4:30pm **Session 3: New Techniques for Dimension Reduction and Model Selection**

Chair: Trevor Park

Speakers:
J. Sunil Rao, "Spike and Slab Shrinkage for Analyzing Microarray Data"
Alan Izenman, "Reduction of Dimensionality Using Nonlinear Methods"

4:30-6:30pm **Poster Session**

Saturday, January 10, 2004

8:00-8:30am **Breakfast** (Reitz Union Grand Ballroom)

All Sessions Meet in Reitz Union Grand Ballroom

8:30–10:30am **Session 4: Model Selection in Statistical Learning**

Chair: Malay Ghosh

Speakers:
Rob Tibshirani, "Least Angle Regression, Forward Stagewise and the Lasso"
Mario Figueiredo, "On Feature Selection for Supervised and Unsupervised Learning"

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

ll:00am-l:00pm **Session 5: Proteomics**

Chair: Mike Daniels

Speakers:
Keith Baggerly, "The Analysis of Proteomic Spectra from Serum Samples"
Scott Schmidler, "Statistical Shape Methods for Data Mining in Protein Structure Databases"

l:00-2:30pm **Lunch** — free time

2:30–4:30pm **Session 6: Statistics for Large Datasets**

Chair: Alex Trindade

Speakers:
David Scott, "Partial Mixture Estimation for Outlier Detection, Mapping, and Clustering of Large Datasets"

Bill Cleveland, "Intensive Learning for Very-Large Databases"

5:00-8:30pm **Barbecue** - Linda Young's home, 13414 NW 19^ PI (map appended)

Invited Talks

Data Mining: Where do we start?
Richard De Veaux, Williams College

One definition of data mining states: ”Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. ” (D.Hand) Much of exploratory data analysis (EDA) and inferential statistics concern the same problems. What’s different about data mining? Part of the challenge of data mining is the sheer size of the data sets and/or the number of possible predictor variables. With 500 potential predictors, just defining, summarizing them and graphing them to start the process is nearly impossible. Problems of model selection, creation of new variables, dimension reduction and missing values, to name a few, overwhelm the data miner from the start. To get a handle on the problem,sometimes we start by creating a preliminary model just to narrow down the set of potential predictors. This exploratory data modeling (EDM) seems to be at odds with standard statistical practice, but, in fact, it’s simply using models as a new exploratory tool. We’ll take a brief tour of the current state of data mining algorithms and map out what I think are some of the biggest challenges for the statistician/data miner and how statisticians can focus their attention to solving them.

Data Quality
David J. Hand, Imperial College, London

High quality data is a prerequisite for high quality conclusions. As such it is central to data analysis and to all that hinges on the conclusions of such analysis. Without good data, our understanding and predictions are suspect and often wrong. This paper gives an overview of data quality, covering definitions, measurement, monitoring, and improvement. Some important special topics are discussed in detail, including missing values, anomaly detection, deliberate data distortion, and what to do in the absence of baseline data. The talk is illustrated with real examples from a wide variety of areas.

Multilevel Models in Gene Expression Data Analysis
Giovanni Parmigiani, Johns Hopkins University

Molecular biology is being revolutionized by technologies measuring simultaneously the level of expression of a large number of genes. In statistics, joint estimation of many related quantities is typically approached by multilevel modeling, and the associated Empirical Bayesian and Hierarchical Bayesian techniques. In genomics, multilevel models represent variation in at least two stages, one describing variation of expression at the gene level, and the other describing variation of gene-specific parameters across the genome.

In this lecture I will present three vignettes applying multilevel models in the genomic context. The first is a simple comparison of alternative ways of modeling the joint distribution of signal and noise in multilevel models for gene selection. The second is a more elaborate multilevel mixture model used to determine sample sizes for gene selection: we use multilevel models to estimate false discovery while accounting for heterogeneity of alternative hypotheses. The third is a multilevel mixture model used in denoising molecular classification data and developing discrete latent classes that can be used to hypothesize novel molecular subtypes.

Semiparametric Regression for Microarray Gene Expressions: Support Vector Machines and Mixed Models

Xihong Lin*, University of Michigan

We consider a semiparametric regression model to relate a normal clinical outcome to clinical covariates and gene expressions, where the clinical covariate effects are modeled parametrically and gene expression effects are modelled nonparametrically using the support vector machine. The nonparametric function of gene expressions allows for the fact that the number of genes is likely to be large and the genes are likely to interact with each other. We show that the dual problem of the primal support vector machine problem can be formulated using a linear mixed effects model. Estimation hence can proceed within the linear mixed model framework using standard mixed model software. Both the regression coefficients of the clinical covariate effects and the support vector estimator of the nonparametric gene expression function can be obtained using the Best Linear Unbiased Predictor in linear mixed models. The smoothing parameter can be estimated as a variance component in linear mixed models. A score test is developed to test for the significant gene expression effects. The methods are illustrated using a prostate cancer data set and evaluated using simulations.

* Collaborated with: Dawei Liu and Dabashis Ghosh, University of Michigan.

Spike and Slab Shrinkage for Analyzing Microarray Data

J. Sunil Rao, Case Western Reserve University

We introduce a class of variable selection procedures that come from what we refer to as rescaled spike and slab models. We study the importance of the prior hierarchical specifications and draw connections to frequentist generalized ridge regression estimation. Specifically, we study the usefulness of continuous priors to model hypervariance parameters, and the effect scaling has on the posterior mean through its relationship to penalization. We demonstrate the importance of selective shrinkage and model averaging for effective variable selection in terms of risk performance. We then show how these strategies might be used to model microarray data - in particular for the problem of detecting differentially expressing genes. A large microarray database consisting of samples from various stages of colon cancer through to liver metstasis is used to illustrate the methodology. This is joint work with Hemant Ishwaran of the Cleveland Clinic Foundation.

Reduction of Dimensionality Using Nonlinear Methods

Alan J. Izenman, Temple University

Classical multivariate analysis has focused on linear methods for dimensionality reduction. Examples include principal components analysis, canonical variate analysis, and discriminant analysis. New techniques are now being introduced into the statistical literature and the machine-learning literature that attempt to generalize these linear methods for dimensionality reduction to nonlinear methods. One of the main aspects of such generalizations is that there are different strategies for viewing the presence of nonlinearity in high-dimensional space. Thus, we have recently seen a number of versions of ”nonlinear” PCA, including polynomial PCA, principal curves and surfaces, autoassociative multilayer neural networks, smallest additive principal components, and kernel PCA. The kernel method, which has also been used to generalize canonical variate analysis and independent component analysis, was originally proposed for use in support vector machines. In this talk, we discuss several of these methods and their connections to each other.

Least Angle Regression, Forward Stagewise and the Lasso

Rob Tibshirani, Stanford University

We discuss ”Least Angle Regression” (”LARS”), a new model selection algorithm. This is a useful and less greedy version of traditional forward selection methods. Three main properties of LARS are derived. (1) A simple modification of the LARS algorithm implements the Lasso, an attractive version of Ordinary Least Squares that constrains the sum of the absolute regression coefficients; the LARS modification calculates all possible Lasso estimates for a given problem in an order of magnitude less computer time than previous methods. (2) A different LARS modification efficiently implements Forward Stagewise linear regression, another promising new model selection method; this connection explains the similar numerical results previously observed for the Lasso and Stagewise, and helps understand the properties of both methods, which are seen as constrained versions of the simpler LARS algorithm. (3) A simple approximation for the degrees of freedom of a LARS estimate is available, from which we derive a Cp estimate of prediction error; this allows a principled choice among the range of possible LARS estimates. LARS and its variants are computationally efficient: the paper describes a publicly available algorithm that requires only the same order of magnitude of computational effort as Ordinary Least Squares applied to the full set of covariates. Connections to ”boosting” are also described. This is joint work with Bradley Efron, Trevor Hastie and Iain Johnstone.

On Feature Selection for Supervised and Unsupervised Learning

Mario A. T. Figueiredo, Technical University of Lisbon, Portugal

It has recently been shown that feature selection in supervised learning can be embedded in the learning algorithm by using sparsity-promoting priors/penalties that encourage the coefficient estimates to be either significantly large or exactly zero.

In the first half of this talk, I will review this type of approach (which includes the well-known LASSO criterion for regression) and present some recent developments: (i) simple and efficient algorithms (with both parallel and sequential updates) for both binary and multi-class problems; (ii) generalization bounds; (iii) feature selection ”inside” the kernel for kernel-based formulations. Experimental results (on standard benchmark data-sets and also on gene expression data) reveal that this class of methods achieves state-of-the-art performance.
While feature selection is a well studied problem in supervised learning, the important issue of determining what attributes of the data better reveal its cluster structure is rarely touched upon. Feature selection for clustering is a difficult because, in the absence of class labels, there is no obvious optimality criterion. In the second half of my talk I will describe two recently proposed approaches to feature selection for mixture-based clustering. One of the approaches uses a new concept of ”feature saliency” which can be estimated using an EM algorithm. The second approach extends the mutual-information-based feature relevance criterion to the unsupervised learning case. The result is an algorithm which ”wraps” mixture estimation in an outer layer that performs feature selection. Experiments show that both methods have promising performance.

The Analysis of Proteomic Spectra from Serum, Samples

Keith Baggerly, Dept, of Biostatistics, MD Anderson Cancer Center

Just as microarrays allow us to measure the relative RNA expression levels of thousands of genes at once, mass spectrometry profiles can provide quick summaries of the expression levels of hundreds of proteins. Using spectra derived from easily available biological samples such as serum or urine, we hope to identify proteins linked with a difference of interest such as the presence or absence of cancer. In this talk, we will briefly introduce two of the more common mass spectrometry techniques, matrix-assisted laser desorption and ionization/time of flight (MALDI-TOF) and surface-enhanced laser desorption and ionization/time of flight (SELDI-TOF). We then describe two case studies, one using each of the above techniques. While we do uncover some structure of interest, aspects of the data clearly illustrate the need for careful experimental design, data cleaning, and data preprocessing to ensure that the structure found is due to biology.

Statistical Shape Methods for Data Mining in Protein Structure Databases

Scott Schmidler, Duke University

Understanding protein structure and function is one of the great post-genome challenges of biology and molecular medicine. The 3D structure of a protein provides fundamental insights into its biological function, mechanism, and interactions, and enables rational design of drugs to modify or inhibit these properties. Large-scale DOE- and NIH-funded efforts are now under way to collect high-resolution structural data for wide coverage of the protein universe.

We present a statistical framework for analysis, prediction, and discovery in protein structural data using methods adapted from the statistical theory of shape. Our approach provides natural solutions to a variety of problems in the field, including the study of conservation and variability, examination of uncertainty in database searches, algorithms for flexible matching, detection of motions and disorder, and clustering and classification techniques. Several of these will be discussed.

Partial Mixture Estimation for Outlier Detection, Mapping, and Clustering of Large Datasets

David Scott, Rice University

The covariance matrix is a key component of many multivariate robust procedures, whether or not the data are assumed to be Gaussian. We examine the idea of robustly fitting a mixture of multivariate Gaussian densities, but when the number of components is intentionally too few. Using a minimum distance criterion, we show how useful results may be obtained in practice. Application areas are numerous, and examples will be provided. We will reexamine the classical Boston housing data, with spatial views of the residuals, as well as similar California data.

Intensive Learning for Very-Large Databases

William S. Cleveland, Bell Labs

Very large databases today are the rule rather the exception. One approach to their analysis is relaxed learning: carry out queries that retrieve functions of the detailed data that summarize behavior, and then analyze the resulting reduced information. While the study of summaries is important, and can sometimes suffice, it often does not succeed in exploiting fully the information in a very large database. Intensive learning is needed; a significant fraction of the data are examined in immense detail. Research in intensive learning for very large databases must cut across many areas of statistics and computer science. In statistics, even the very foundations must be examined because for some large databases, while stochastic variation is salient, sample sizes are so large that statistical variability in estimators becomes negligible, and residual variability is largely model misspecification. In computer science, even the basic issues of leading-edge distributed computing environments are highly relevant because a distributed architecture appears ideal for attacking very large databases. In the talk, several areas of statistics and computer science will be discussed in the context of an ongoing study of a very large database of Internet traffic. Issues in this project are wide ranging, from the pros and cons of using an off-the-shelf database management system vs. a roll-our-own system, to strategies of analysis that balance statistical information needs with computing realities, to when the theorems on the convergence of superposed point processes to Poisson can be applied to real life.

Poster Abstracts

Discovering Crime Patterns from a State Database

Sikha Bagui, The University of West Florida

This article presents a knowledge discovery effort to retrieve meaningful information about crime from a State database obtained from http://www.unl.edu/SPPQ/datasets.html, a US State Politics & Policy Quarterly Data Resource website. This raw data was preprocessed, and data cubes were created using Structured Query Language (SQL). The data cubes were then used in deriving quantitative generalizations for the data in the form of $_{t_{we}A_g}hts$ and $_{d_{we}A_g}hts$. An entropy based attribute relevance study was done to determine the relevant attributes. Finally, a machine learning software called WEKA was used for mining association rules, developing a decision tree and clustering.

On a Modification of the Naive Bayes Approach to Statistical Text Categorization, with Applications to Detecting Spam,

Sudip Bose, The George Washington University

We discuss a modification of the Naive Bayes approach to statistical text categorization, using the "bag-of-words" representation (Madigan 2003). This approach relaxes the assumption of independence for words and also uses groups of words arising from clusters involving predefined "phrases of interest".

Bayesian Nonlinear Regression for Large p Small n Problems

Sounak Chakraborty*, University of Florida

Statistical modelling and inference problems with sample sizes substantially smaller than the number of available covariates are challenging. This is known as large *p* small *n* problems. We develop nonlinear regression models in this setup for accurate prediction. In this paper, we introduce a full Bayesian support vector regression model with Vapnik's c-insensitive loss function, based on reproducing kernel Hilbert spaces (RKHS). This provides a full probabilistic description of support vector machine (SVM) rahter than an algorithm for fitting purposes. Also, we have considered Relevance Vector Machine (RVM) introduced by Bishop, Tipping and others. Instead of the original treatment of the RVM relying on the use of type II maximum likelihood estimates as the hyper-parameters, we put a prior on the hyper-parameters and uwse Markov chain Monte Carlo technique for computation. We apply our model for prediction of blood glucose concentration in diabetics using florescence based optics. We have extended the full Bayesian support vector regression (SVR) and relevance vector regression (RVR) models when the response is multivariate. The multivariate version of the SVM and RVM is illustrated with a prediction problem in near-infrared (NIR) spectroscopy.

Key Words: Gibbs sampling; Markov chain Monte Carlo; Metropolis-Hastings algorithm; Near infrared spectroscopy; relevance vector machine; reproducing kernel Hilbert space; support vector machine; Vapnick's c-insensitive loss.

* Collaborated with: Malay Ghosh, University of Florida, and Bani K. Mallick, Texas A&M University.

Error Density and Distribution Function Estimation in Nonparametric Regression

Fuxia Cheng, Illinois State University

This talk will discuss some asymptotics of some error density and distribution function estimators in nonparametric regression models. In particular, the histogram type density estimators based on nonparametric regression residuals obtained from the full sample are shown to be uniformly almost surely consistent. Uniform weak convergence with a rate is obtained for the empirical d.f. of these residuals. Furthermore, if one uses a part of the sample to estimate the regression function and the other part to estimate the error density, then the asymptotic distribution of the maximum of a suitably normalized deviation of the density estimator from the true error density funtion is the same as in the case of the one sample setup. Simiarly, a suitably standardized nonparametric residual empirical process based on the second part of the sample is shown to weakly converge to a time transformed Brownian bridge.

Mining Serial Data: Time-Frequency Methods Using Wavelet Packets

J. Wade Davis, Baylor University

Techniques for discrimination and clustering of serial data have been employed in many different scientific areas. Applications include speech pattern recognition, the Categorization of seismic records as originating from either nuclear explosions or earthquakes (Shumway and Stoffer, 2000; Shumway, 2003), the classification of cities to various climate regimes based on temperature patterns (Davis and Cavanaugh, 2001), and the determination of different states of consciousness based on EEG readings (Gersch, 1981).

Since many time series exhibit nonstationary dynamics, a great need exists for discrimination and clustering procedures that exploit rather than ignore temporal frequency shifts. This can be achieved using a time-frequency representation (TFR), which characterizes the frequency dynamics. The discriminant method we primarily consider is based on TFRs computed using the discrete wavelet packet transform, in conjunction with Kullback information measures. However, any general TFR can be used in conjunction with our proposed discriminant scoring function. The resulting scores can then be used for classification or clustering. We demonstrate the effectiveness of our methods using simulations, and we apply the methods to a well-known data set of underground explosions and earthquakes. Key Words: Discriminant analysis; nonstationarity; time series analysis; wavelets.

Identifying Differentially Expressed Genes in an Unreplicated Microarray

Rhonda R. DeCook, Iowa State University

Microarray technology has become widespread as a means to investigate gene function and metabolic pathways in an organism. A common experiment involves probing, at each of several time points, the gene expression of experimental units subjected to different treatments. Due to the high cost of microarrays, such experiments are often unreplicated, which means that the gene expression of only one experimental unit is measured for each combination of treatment and time point. Though an experiment with replication would provide more powerful conclusions, it is still possible to identify differentially expressed genes in this type of experiment and to control the false discovery rate at a pre-specified level. We present such a method utilizing polynomial regression models to approximate underlying expression patterns over time for each treatment and gene. Initially, all possible regression models involving treatment effects, terms polynomial in time, and interactions between treatments and polynomial terms are considered. Using a two-stage permutation approach, a "best" model is first chosen at each gene with a method incorporating the traditional F-statistic. In the second stage of the approach, we identify genes whose "best" model differs significantly from the overall mean model suggesting the gene may be biologically interesting. The expected proportion of false positive results among all positive results is estimated using a method presented by Storey (2001).

Multifractal and Recurrence Time Based Methods for Deciphering the Structures of Genomic DNA Sequences

Jianbo Gao, University of Florida

The completion of the human genome and genomes of many other organisms calls for the development of faster computational tools which are capable of easily identifying the structures and extracting features from DNA sequences. Here we report two novel algorithms. One is for finding genes based on the observation that multifractal signatures of a DNA segment often have abrupt changes near the borders of gene-gene and coding-noncoding regions, due to compositional and structural differences between different DNA regions. The other is based on recurrence time statistics, which has its root in nonlinear dynamical systems theory and can conveniently study all kinds of periodicity and exhaustively find all repeat-related features from a genomic DNA. A convenient coding region indicator based on the recurrence time statistics will also be discussed.

A Fast Clustering Algorithm with Application to Cosmology

Woncheol Jang, Carnegie Mellon University

We present a fast clustering algorithm for density contour clusters (Hartigan, 1975) that is a modified version of the Cuevas, Fabrero and Fraiman (1998) algorithm. By Hartigan's definition, clusters are the connected components of a level set $S_c = \{f >$

$c\}$ where f is the probability density function. We use kernel density estimators and orthogonal series estimators to estimate f and modify Cuevas et al to extract the connected components from level set estimators $S_c = \{f > c\}$. Unlike the original algorithm, our method does not require an extra smoothing parameter and can use the Fast Fourier Transform (FFT) to speed up the calculations.

We show the cosmological definition of clusters of galaxies is equivalent to density contour clusters and present an application in cosmology.

On the Control Gene Based Normalization in cDNA Microarrays

Masha Kocherginsky*, University of Chicago

Several methods of normalization of the cDNA arrays have been proposed, some of which are based on the use of control genes. These control genes are believed to be biologically stable across the experimental conditions. Some authors also propose methods that select internally stable control genes (Schadt, 2001; Vandesompele, 2002, etc.). After selecting an appropriate set of control genes, they can be used for normalization by scaling the intensities of other genes by the average expression level of the control genes. We argue that any statistical inference based on the normalized intensities need to account for the variability in the normalization constants. We propose an adjustment to account for this additional variability in the ¿-statistic for testing whether a gene is differentially expressed. This method is illustrated using cDNA microarrays containing 7653 cattle genes which are hybridized with Cy3 and Cy5 labeled RNA from bovine leukemia virus (BLV) infected and uninfected cell lines, but its potential applicability is much broader.

* Collaborated with: Xuming He, University of Illinois at Urbana-Champaign; Moon-Ho Ringo Ho, McGill University; Robin Everts, University of Illinois at Urbanna-Champaign; Harris A. Lewin, University of Illinois at Urbana-Champaign.

Identifying and Quantifying SNPs Responsible for a Linkage Signal

Mingyao Li*, University of Michigan

Once genetic linkage has been identified for a complex disease, the next step is often association analysis, in which single-nucleotide polymorphisms (SNPs) within the linkage region are genotyped and tested for association with the disease. If a SNP shows evidence for association, a key question is to what degree the linkage result can be explained by the associated SNP. To answer this question, we developed a novel statistical method that quantifies the degree of linkage disequilibrium (LD) between the associated SNP and the putative disease locus. We describe a simple parametric likelihood of the marker data conditional on the trait data based on disease penetrances and disease-SNP haplotype frequencies. Special cases of the likelihood include complete LD and linkage equilibrium (LE). We estimate model parameters by maximum likelihood using a simplex method. We propose two likelihood ratio tests to distinguish the relationship of the associated SNP and the disease locus. The first

test assesses whether the associated SNP and the disease locus are in LE so that the SNP plays no causal role in the linkage signal. The second test assesses whether the associated SNP and the disease locus are in perfect LD so that the SNP or a marker in complete LD with it can fully account for the linkage signal. LD measures such as D' and r^2 can be estimated from the disease-SNP haplotype frequency estimates. These estimates are of particular interest given incomplete disease-SNP LD.

to investigate the performance of our method, we estimate 500 affected sibling pairs under a variety of genetic models with sibling гесшгепсе3 risk ratio between 1.3 and 1.5. Simulation results demonstrate that our method yields accurate parameter estimates under most circumstances. Estimates may be biased when hitting the boundaries of parameter estimates under most circumstances. Estimates may be biased when hitting the boundaries of parameter space. Accuracy of parameter estimates can be improved by including unaffected sibling pairs. For dominant model, our tests have > 80% power to reject LE when $r^2 > 0.25$, and > 80% power to reject complete LD when $r^2 < 0.33$ and disease allele frequency is moderate at significance level 0.05. Powers are similar for additive model, but slightly different for recessive model. Our method will be valuable for prioritizing SNPs in searching for disease-susceptibility alleles. Key Words: linkage disequilibrium, mapping complex traits, gene localization, identification of disease genes.

* Collaborated with: Concaio R. Abecasis, Michael Boehnke.

Gene Expression Mapping

Jessica M. Maia, North Carolina State University

Quantitative trait loci (QTL) mapping is used to find genomic regions associated with a quantitative trait such as wood density in trees or hypertension in humans. Recently, QTL mapping has been applied to gene expression levels. The idea behind this new trend is that instead of looking for associations between genes and a trait, one could look for associations between the expression of genes which underlie a trait and the genomic regions regulating those expressions.

The expression of genes within an organism is highly correlated. My research consists of developing methods that can summarize thousands of gene expressions levels into fewer components. There are many methods of data reduction in the literature, but few, if any, can can be assigned a biological interpretation. For example, principal component analysis has been applied to gene expression data. Once associations among genomic regions and a principal component are found, one still has to make sense of what these associations represent biologically. Interpreting principal component analysis results in terms of the original gene expression levels is part of my current research. In the near future, I would like to find or create a summary method that is useful to this type of data.

I have been working with a eucalyptus data set generated by a graduate student in the Forest Biotechnology Group at North Carolina State University. It consists of

gene expression levels of 2,600 genes measured in 88 trees from a backcross of two eucalyptus strains: Eucalyptus grandis and Eucalyptus globulus. In addition to the expression data, there are 18 phenotypic traits measured in these 88 trees. Principal component analysis has been done for the 2,600 genes in the data set. The first 28 principal components capture 90% of the total variance contained in the original data set. The first 28 principal components have been mapped to regions in the eucalyptus genome using composite interval mapping. Our analysis suggests that regions in chromosomes 2 and 12 regulate a large part of the gene expression levels in our data set.

Interactions and Variable Importance in Genomic Data

Ingo Ruczinski*, Johns Hopkins University

Exploring higher order interactions is a statistical challenge frequently occurring in the analysis of genomic data. For example, interactions may be present between single nucleotide polymorphisms or chromosomal deletions when considering associations with disease or disease stages, respectively. Previously, we introduced *Logic regression* as an adaptive regression methodology that can be used to address the problem of detecting such interactions, by constructing predictors as Boolean combinations of binary covariates. Initially, a sequence of model candidates were generated adaptively, and a single model was chosen using permutation tests or cross-validtion. Recently, we investigated some Monte Carlo Markov Chain based procedures that generate an ensemble of models and measures of variable importance.

Logic regression was introduced by Ruczinski, Kooperberg and LeBlanc to address problems arising when data of mostly binary covariates are analyzed, and the interactions between those predictors is of main interest. Given a set of binary predictors X (such as indicator variables in SNP data whether a variation at a particular site is present), we try to create new, better predictors for the response by considering combinations of those binary predictors. For example, if the response is binary as well (which is not required in general), we attempt to find decision rules such as "if Xi,X_2,%3 and X_4 are true", or "X_5 or X_6 but not Xj are true", then "the response is more likely to be in class 0". In other words, we try to find Boolean statements involving the binary predictors that enhance the prediction for the response. In more specific terms: Let $X\backslash, \ldots,$ Xk be binary predictors, and let У be a response variable. We try to fit regression models of the form $g(E\backslash Y]) = 6_0 + b\backslash L\backslash + \cdots + b_n L_n,$ where Lj is a Boolean expression of the predictors $X,$ such as $Lj = [\{X2$ Л X4^c) V $X7]$. The above framework includes many forms of regression, such as linear regression $(g(E[Y]) = E[Y]$) and logistic regression $(g(E[Y]) = log(E[Y]/(l — E[Y])))$. For every model type, we define a score function that reflects the "quality" of the model under consideration (for example, the residual sum of squares for linear regression and for the deviance logistic regression). We try to find the Boolean expressions in the regression model that minimize the scoring function associated with t his model type, estimating the parameters bj simultaneously with the Boolean terms Lj. In general,

any type of model can be considered, as long as a scoring function can be defined. For example, we also implemented the Cox proportional hazards model, using the partial likelihood as the score. In previous versions, a sequence of models was generated, allowing various numbers of logic expressions of various sizes. Then a single model was chosen using permutation tests and/or cross-validation. While this technique has been applied successfully to genomic data, the shortcoming is that variables that are highly correlated to the variables in the final model most likely go undetected due to the selection procedure. In general, there might be various models of a given size that score about as good as the chosen model, and we are interested in all variables that might be associated with the response. To address this issue, we propose some Monte Carlo Markov Chain based procedures that generate ensembles of models and measures of variable importance.

Key Words: adaptive model selection, boolean logic, genomic data, interactions, simulated annealing, SNPs, variable importance.

* Collaborated with: Charles Kooperberg, Fred Hutchinson Cancer Research Center.

Microarray Genome Scans for Complex Disease: Power Under Multilocus Disease Models

Paul Schliekelman, University of Georgia

Recent data from studies combining gene expression level data and genotype information for complex traits have shown that expression levels often have strong dependence on a single causative locus, even when correlated with a complex trait. This suggests that microarrays might be a powerful tool for location of complex disease loci. I will propose a two-tiered approach for locating complex trait loci using microarrays: 1) compare affected and unaffected individuals with microarrays to identify genes with expression patterns that are correlated with disease status, then 2) collect genotype information in these individuals and use expression levels for these identified genes as the response variable for gene mapping. I will show power and sample size calculations for this approach for different multilocus disease penetrance models, population parameters, and expression distributions. These results will show conditions under which the microarray genome scan approach is expected to be far more powerful than currently popular methods such as the affected sib pair design.

Systematic Use of GO Annotation to Analyze Microarray Data

Chad Shaw*, Baylor College of Medicine

Making sense of large gene lists is central to analysis of global expression profiles. One infromatic resource which can be helpful is the Gene Ontology project. Gene Ontology aims to locate each gene to a 3-fold heirarchy of terms-a controlled vocabulary for describing genes. We have developed a unique and extensive software toolkit for integrated analysis of microarray data with the Gene Ontology effort. Our toolkit

is assembled as an R package, and the package makes extensive use of the SJava interface. We have implemented a variety of analyses based on our toolkit.

First, we have implemented a count-based enrichment analysis of gene lists. The enrichment analysis proceeds by tabulating the abundances of gene-list elements mapped at or below each node in the GO tree. The null distribution for counts is estimated by mapping the entire clone set (collection of arrayed targets) to the GO structure.

Second, we have explored using the GO annotation as an explanatory variable to structure analysis of the normalized array data - integrating the numerical data with the GO analysis. In this context, the GO annotation is used to construct what we call "GO-stratified ANOVAs" of the microarray data. Other regression based use of the GO annotations are possible, such as the search for GO-patterns in timecourse microarray data.

Finally, we have developed a distance metric on microarray gene lists based on the GO profile of the list. This distance metric is appealing for a number of reasons: it allows for "content" based comparison of lists rather than "gene-identity" based comparison. In the microarray gene-clustering setting, such a content based comparison of lists is highly desirable, since the closters are by definition disjoint at the level of genes.
* Collaborated with: Nathan Whitehouse, Andrew Young, Baylor College of Medicine.

Dynamic Profiling of Online Auctions Using Curve Clustering

Galit Shmueli, University of Maryland

Electronic commerce, and in particular online auctions, have received an extreme surge of popularity in recent years. While auction theory has been studied for a long time from a game-theory perspective, the electronic implementation of the auction mechanism poses new and challenging research questions. Although the body of empirical research on online auctions is growing, there is a lack of treatment of these data from a modern statistical point of view.

In this work, we present a new source of rich auction data and introduce an innovative way of modeling and analyzing online bidding behavior. In particular, we use functional data analysis to investigate and scrutinize online auction dynamics. We describe the structure of such data and suggest suitable methods, including data smoothing and curve clustering, that allow one to profile online auctions and display different bidding behavior. We use data on a collection of closed auctions from eBay.com to illustrate the methods and their insights into the bidding process. Finally, we tie our results to the existing literature on online auctions.

Gaussian Mixture Regression

Hsi Guang Sung, Rice University

The key functions of a data model is to provide an effective way to combine information from each observation. Parametric models achieve the effectiveness by assuming the underline parametric form. Nonparametric models replace the parametric assumption by the general assumption that the closer data points contains more information than the farther away data. This neighborhood-based models work well in low-dimension, they are often ineffective in modelling high dimensional data due to the scarcity of data points in higher dimension. Aiming at an effective approach to model high dimensional data, we propose the Gaussian Mixture Regression (GMR) model and the algorithms to fit the model. GMR starts with a Gaussian mixture model for the joint density $fxy,$ where X G R^p and Y G R. Once the Gaussian mixture model is established, the regression function $m(x)$ can be easily derived from $f_{XY,}$ by definition: $m(x) = E(Y\backslash X = x) = f\ yfY\backslash x(y\backslash x)dy,$ where $fy\backslash x(y\backslash x) = fx,Y(x,y)/\ f\ fx,Yx,y)dy\cdot$ Under the joint Gaussian mixture density, $fx,y(x,y)$ = $Y^f_{=}i$ цЖт $y'i$ ^{u}r **^i), the** regression function enjoys a simple form, $m(x) = Ylf=i^{w}j(x)^{m}j(x)i$ where the weight function has a closed-form expression $^{w}j(x) = 7tj(f)(x]\ Uj,\ Yj)/\ \Sigma i\ ^{\pi}i\varphi\{x]\ Oi)$ and $m\ j\ (x)$ is a linear function of x. The weight function $Wj(x)$ can be interpreted as the global Gaussian clustering effect at the target point x. GMR model offers a template to blend the global and local structures, especially in high-dimensional data analysis. Once the model is established for the original feature space $R^p,$ a sequential algorithm can be implemented to search for the subspace of the lower dimension while still preserving the predictive power of the regression model. Following are two main themes of this research:

1. Fitting the joint mixture density: Implement IPRA algorithm to fit the mixture model of the joint density in R^{p+l}.

2. Finding the optimum subspace: Implement a search algorithm for the projection: $B : RF$ i—^ Ry such that $B = \text{argmin}_{S:ßPl_}{}^{\wedge}{}_{ß}$p-i $E(m(x) — m(Bx))^2$.

I will demonstrate the GMR applications on both simulated and real data sets.

On the Use of Artificial Neural Networks for Predicting Infant Mortality in Epidemiological Studies

Permual Venkatesan*, Department of Statistics,
Tuberculosis Research Center, ICMR, Chennai, India

Artificial neural Networks are a powerful tool for analyzing data sets where there are potentially complicated non-linear interactions between measured inputs and the quantity to be predicted. The most commonly used multilayer perception network model is comparable to nonlinear regression and discriminant models. The major methodological issue in drawing inference and making predictions from observational studies is that outcomes will be influenced by differences in the underlying substrates

in the population and interventions.

Our focus here is on prediction of infant mortality from infant- and mother-specific features observed at the time of birth. Prediction is pragmatically important for identifying children at high risk for infant mortality and for taking appropriate preventive interventions. Logistic regression is commonly used for making predictions from observational or epidemiological data. Neural networks have proven to be a useful tool for prediction problems in many applications, but their application to epidemiologic phenomena is not widespread. In this work, we compare the performance of logistic regression and a feed forward neural network for predicting infant mortality using data from a large epidemiological follow up study. We compare the empirical predictions from various models, and evaluate prediction capability in terms of predictive accuracy and variance. The relationship between model complexity and prediction capability is examined.

* Collaborated with: Anitha, S., National Institute of Epidemiology, ICMR, Chennai, India and Hogan, J.W., Center for Statistical Science, Brown University, Providence, RI USA.

Nonparametric Goodness-of-fit Test for Heteroscedastic Regression Models

Lan Wang*, University of Minnesota

For the heteroscedastic nonparametric regression model $Y_{ni} = m(x_{ni}) + a(x_{ni})e_{ni}$, $i = 1,..,n$, we propose a novel method for testing that the regression function m has certain parametric form. It differs from most existing smoothing-based procedures in that it does not require consistent estimation of $m(cdot)$ and $\sigma(\cdot)$, and is asymptotically unbiased under the null hypothesis. The test statistic is modelled after the usual lack-of-fit statistic for constant regression in the case of replicated observations, is computationally very simple. The asymptotic results use recent developments in the asymptotic theory for analysis of variance when the number of factor levels is large. Extensive simulation comparisons with many competing procedures demonstrate that the proposed method has desirable type I and II error probabilities with finite samples. A real data set is analyzed.

* Collaborated with: Michael Akritas, Penn State University; and Ingrid Van Keilegom, Université catholique de Louvain in Belgium.

Boosting Extensions to Find AnomмЬy Structure in Data

Virginia Wheway, University of New South Wales, Australia

Recent advances in data mining have led to the development of a method called "boosting". Instead of building a single model for each dataset, "boosting" sees several models being built using weighted versions of the original data. These models are then combined into a single prediction model via a voting method. Models, which

are more accurate on the original data, are given hight voting power. Studies have demonstrated that this method of model combination leads to significantly more accurate predictions on unseen data.

This research demonstrates how method of boosting may be extended beyond its original aims of improved prediction. A plot of error statistics may be used as a tool to detect noisy data and unearth structure within datasets that cannot be detected using standard methods. By retaining a series of basic statistics are each boosting iteration, and plotting their standard errors, anomaly structure in data may be unearthed. This technique is able to unearth clusters of data, which appear to be behaving differently to other sections of the data. Whether or not they occur in groups or as single observations, outliers may also be flagged using this method. The repeated success of this technique may be proven using multivariate statistics.

An example industrial dataset will be used to demonstrate the process, and show the power of being able to detect unknown clusters within datasets. If such clusters go undetected, model errors are higher and the resulting model is unnecessarily complex. In practice, an accurate model which is interpretable to all team members is favorable to a complex mathematical model, understood by few.

Penalized, Spline Estimation for Generalized Partially Linear Single-Index Models

Yan Yu, University of Cincinnati

Single-index models are potentially important tools for multivariate nonparametric regression. They generalize linear regression by replacing the linear combination *afx* with a nonparametric component, *yo(aIx),* where *ηo(·)* is an unknown univariate link function. By reducing the dimensionality from that of a general covariate vector *x* to a univariate index *of x,* single-index models avoid the so-called "curse of dimensionality". I propose penalized spline (P-spline) estimation of *ηo(·)* in generalized partially linear single-index models, where the mean function has the form $\eta_0(\alpha\beta\chi) + \beta f z$ and responses are from the general exponential family such as binary or Poisson responses. The P-spline approach offers a number of advantages over other fitting methods for single-index models. All parameters in the P-spline single-index model can be estimated simultaneously by penalized likelihood or penalized quasi-likelihood. As a direct fitting method, our approach is rapid and computationally stable. Standard nonlinear optimization software can be used. Moreover, joint inference for *ηo(·),* o¡o, and *βo* is possible by standard estimating equations theory such as the sandwich formula for the joint covariance matrix. Using asymptotics where the number of knots is fixed though potentially large, I show -^/ū-consistency and asymptotic normality of the estimators of all parameters. These asymptotic results permit joint inference for the parameters. Several examples illustrate that the model and proposed estimation methodology can be effective in practice.

Statistical Methods for Estimating Mutation Rate and Effective Population Size from Samples of DNA Sequences

Feng Zhan*, University of Texas at Houston

Effective population size N and mutation rate m are two basic parameters in population genetics, and $\Theta = 4\ N\ \mu$, is very important and useful in genetic research. In the past twenty years there have been a number of statistical methods developed for estimating q from DNA sequence sample. However, there has not been a comprehensive comparison among those estimators thus their statistical properties, such as unbiasesness and variance as well as their relative performances, remain poorly understood, particularly for large sample size. In this study we will first give an extensive review for these estimators and compare them with simulated sequence samples in terms of unbiasesness and variance, which are the two statistical criteria commonly used to evaluate an estimator's performance. Second, using IJPGMA and BLUE procedure we will modify two of the recent methods, Fu's UPBLUE (Ihr, 1994) and Deng-Fu UPBLUE (Deng and Fu, 1996), to extend their usage for larger sample size sequence data. Third, these original and new estimators will be applied to an X-linked sequence data and a mitochondrial DNA sequence data. The advantage and disadvantage of these estimators will also be discussed.

* Collaborated with: Yun-Xin Fu.

Time-Dependent Diffusion Models for Term Structure Dynamics

Chunming Zhang, University of Wisconsin

In an effort to capture the time variation on the instantaneous return and volatility functions, a family of time-dependent diffusion processes is introduced to model the term structure dynamics. This allows one to examine how the instantaneous return and price volatility change over time and price level. Nonparametric techniques, based on kernel regression, are used to estimate the time-varying coefficient functions in the drift and diffusion. The newly proposed semiparametric model includes most of the well-known short-term interest rate models, such as those proposed by Cox, Ingersoll and Ross (1985) and Chan, Karolyi, Longstaff and Sanders (1992). It can be used to test the goodness-of-fit of these famous time-homogeneous short rate models. The newly proposed method complements the time-homogeneous nonparametric estimation techniques of Stanton (1997) and Fan and Yao (1998), and is shown through simulations to truly capture the heteroscedasticity and time-inhomogeneous structure in volatility. A family of new statistics is introduced to test whether the time-homogeneous models adequately fit interest rates for certain periods of the economy. We illustrate the new methods by using weekly three-month treasury bill data. This is a joint work with Jianqing Fan, Jiancheng Jiang, and Zhenwei Zhou.

Estimation in the Additive Risk Model and Its Application

Yichuan Zhao, Georgia State University

Cox's proportional hazards model has been the most popular model for the regression analysis of censored survival data. However, the additive risk model provides a useful alternative. We apply the empirical likelihood ratio method to the additive risk model with right-censoring and derive its limiting distribution. Extension to classification is discussed. Simulation results are presented to compare the proposed method with the existing method. We also demonstrate the use of the additive risk model as a tool to identify genes potentially influencing survival.

Evidence of Pathophysiological Heterogeneity with the Diagnosis of Schizophrenia: Subgroup Differences in Cognitive Deficits and Brain Structure Abnormalities

Hongtu Zhu, Columbia University

Background: One of the most important current questions in schizophrenia research is whether the diagnostic category includes more than one disease. We used a clustering algorithm to identify three subgroups of schizophrenia patients characterized by markedly different cognitive deficits suggestive of fundamental differences in underlying brain pathology. We now compared the subgroups with regard to the three brain structure abnormalities that have most often been identified in previous studies: volume of the temporal lobes, dorsal lateral prefrontal cortex, and lateral ventricles. Method: Eighty stable outpatients with schizophrenia or schizoaffective disorder were divided into groups with normal memory, a selective verbal memory deficit, or global memory deficits. Structural MRI scans from the 80 patients and 30 healthy subjects were analyzed by trained technicians blind to group assignments, using manual and semi-automated procedures to define and measure volumes of the lateral ventricles, temporal lobes and dorsolateral prefrontal cortex. Statistical tools such as clustering algorithm, finite mixture model and mixed model were used to analyze the data set. Results: The patient subgroups differed significantly from one another in ventricular enlargement and in temporal and frontal volume reductions. Normal memory patients had left-lateralized temporal and frontal volume reductions without lateral ventricle enlargement. Verbal deficit patients had bilateral temporal and frontal volume reductions with moderate ventricular enlargement. Global deficit patients had marked ventricular enlargement but had the least remarkable temporal and frontal volume reductions.

* Collaborated with: Bruce E. Wexler, Robert Fulbright, Patricia Goldman-Rakic, John C. Gore, Bradley S. Peterson, Departments of Psychiatry, Diagnostic Imaging and Neurobiology , and the Child Study Center, Yale University School of Medicine, and Departments of Psychiatry, Columbia University. This work was supported by National Institute of Mental Health grants MH56642 and MH02196 to Dr. Wexler and MH01232 to Dr. Peterson.