Case Studies in Bayesian Statistics
Workshop 9
Poster Session

Following is a tentative list of posters being presented during the workshop:

1. Edoardo Airoldi, Curtis Huttenhower, Olga Troyanskaya and David Botstein
2. Dipankar Bandyopadhyay, Elizabeth Slate, Debajyoti Sinha, Dikpak Dey and Jyotika Fernandes
3. Brenda Betancourt and Maria-Eglee Perez
4. Sham Bhat, Murali Haran, Julio Molineros and Erick Dewolf
5. Jen-hwa Chu, Merlise A. Clyde and Feng Liang
6. Jason Connor, Scott Berry and Don Berry
7. J. Mark Donovan, Michael R. Elliott and Daniel F. Heitjan
8. Elena A. Erosheva, Donatello Telesca, Ross L. Matsueda and Derek Kreager
9. Xiaodan Fan and Jun S. Liu
10. Jairo A. Fuquene and Luis Raul Pericchi
11. Marti Font, Josep Ginebra, Xavier Puig
12. Isobel Claire Gormley and Thomas Brendan Murphy
13. Cari G. Kaufman and Stephan R. Sain
14. Alex Lenkoski
15. Herbie Lee
16. Fei Liu
17. Jingchen Liu, Xiao-Li Meng, Chih-nan Chen, Margarita Alegria
18. Christian Macaro
19. Il-Chul Moon, Eunice J. Kim and Kathleen M. Carley
20. Christopher Paciorek
21. Susan M. Paddock and Patricia Ebener
22. Nicholas M. Pajewski, L. Thomas Johnson, Thomas Radmer, and Purushottam W. Laud
23. Mark W. Perlin, Joseph B. Kadane, Robin W. Cotton and Alexander Sinelnikov
24. Alicia Quiros, Raquel Montes Diez and Dani Gamerman

25. Eiki Satake and Philip Amato
26. James Scott
27. Russell Steele, Robert Platt and Michelle Ross
28. Alejandro Villagran, Gabriel Huerta, Charles S. Jackson and Mrinal K. Sen
29. Dawn Woodard
30. David C. Wheeler, Lance A. Waller and John O. Elliott
31. Chun-yin Yip, Sujit Sahu
32. Tingting Zhang, Jun Liu
33. Wei Zhang, Jun Zhu, Jun Liu

A Bayesian perspective on cellular growth

by

Edoardo Airoldi, Curtis Huttenhower, Olga Troyanskaya & David Botstein

Lewis-Sigler Institute for Integrative Genomics, Department of Computer Science & Department of Molecular Biology, Princeton University

eairoldi@Princeton.EDU

Abstract

Growth is a fundamental process in cellular proliferation, and its disruption plays a role in a variety of disorders from viral infection to cancer. Cellular growth is so essential that our ability to probe and reveal the inner mechanisms of the cell crucially depends on our ability to control it. In fact, most experiments are performed on cellular cultures growing in artificial environments. An important example of this is the investigation of the environmental stress response (ESR) in yeast. In this setting, each gene's transcriptional response may be considered to arise from a mixture of two extreme models: either a gene is expressed directly in response to stress, or it is expressed purely in response to the change in growth rate caused indirectly by stress. A clear understanding of the ESR is confounded by the admixture of growth-related and stress-related effects in the magnitude of transcriptional responses. Our goal is thus to isolate the effects of these two extremes, or in the general case, to deconvolute transcriptional responses to primary biological stimuli from responses from indirect growth effects.

Bayesian analysis of a carefully designed experimental probe enables us to estimate (in both continuous and batch cultures) the "effective growth rate" of new collections of expression data, and to establish the portion of response that cannot be attributed to growth. The effective growth rate of a cellular culture is a novel biological concept, and it is useful in interpreting the system-level connections among growth rate, metabolism, stress and the cell division cycle.

Bayesian analysis of zero-inflated count data with applications to dental caries

by

Dipankar Bandyopadhyay

Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina

Elizabeth Slate

Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina

Debajyoti Sinha

Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina

Dipak Dey

Dept. of Statistics, University of Connecticut

Jyotika Fernandes

Dept. of Endocrinology, Diabetes and Medical Genetics, Medical University of South Carolina

463 Huntsman Hall

3730 Walnut Street, Philadelphia, PA 19102

bandyopd@musc.edu

Abstract

The Gullah-speaking inhabitants of the Sea Islands of South Carolina are a unique population because of their minimal Caucasian genetic admixture and high propensity for diabetes. A clinical study was conducted to determine their dental health status of Type-2 diabetic Gullah African Americans. Dental caries was assessed using the total number of decayed, missing and filled surfaces, an index known as DMFS in the dental literature. Data resulted from examining 4 (for canines and incisors) or 5 (for premolars and molars) surfaces per tooth, for all (up to 32) teeth, for over 260 individuals. Also recorded were covariates including age, gender, smoking and brushing/flossing habits, etc., which may influence caries development. We model the tooth-level contributions to DMFS, which range from 0 to 5, and evaluate associations with covariates.

Histograms suggest a zero-inflated binomial model for the tooth-level counts. As in a Hurdle Model, the process determining a healthy tooth (with a count of 0) is treated as a structural zero and hence separated from the remaining counts (1 to 5), which are modeled using a zero-truncated binomial distribution. We develop a multivariate model where covariates enter through a random effects logistic regression on the logit of the probability of a carious surface. To preserve marginal logit structure for interpretability, we use a bridge density (Wang and Louis, 2003) for the subject-specific random effects. The tooth-specific zero-inflation probability is modeled as arising from a beta distribution whose shape/scale parameters are linked to the odds of a healthy tooth (Song et al., 2006). We compare our model with alternatives to assess improvements in fit, prediction and interpretability.

References:

1. Epidemiological study of Periodontal Disease and Diabetes: Cytokine Genes & Inflammation Factors," Dr. J. Fernandes, PI. A project of the South Carolina COBRE for Oral Health, NIH/NCRR P20RR017696, Dr. S. Lanier, PI.

2. Wang, Z. and Louis, T.A. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function, Biometrika, 90, 765-775.

3. Song, S., Dey, D.K. and Holsinger, K.E. (2006). Differentiation among populations with migration, mutation and drift: implications for genetic inference. Evolution, 60, 1-12.

Bayesian Objective Testing of Hardy-Weinberg Equilibrium

by
Brenda Betancourt C.
University of Puerto Rico. Department of Mathematics.
b.betancourt@uprrp.edu

María-Eglée Pérez
University of Puerto Rico. Department of Mathematics.
meglee@uprrp.edu

Abstract

Assessment of Hardy-Weinberg equilibrium is one of the basic problems in population genetics and it is far for being closed from the statistical point of view, as recent efforts in this direction prove. The selection of prior distributions for testing Hardy-Weinberg equilibrium is a difficult issue, as we are dealing with a low dimensional null hypothesis. Recent advances in Objective Bayesian Analysis allow the construction of priors specially suited for Bayesian testing (intrinsic priors). In this work a intrinsic prior is calculated in closed form, and it is applied to some examples for the problem of testing Hardy-Weinberg equilibrium.

Estimating the risk of a crop epidemic from coincident spatiotemporal processes

by

Sham Bhat, Murali Haran, Julio Molineros and Erick Dewolf

Department of Statistics, Pennsylvania State University and Plant Pathology, Kansas State University

326 Thomas, University Park, PA 16802.

mharan@stat.psu.edu

Abstract

Fusarium Head Blight (FHB) or "scab" is a very destructive disease that affects wheat crops resulting in massive financial losses. Learning about areas that are particularly prone to such epidemics is therefore of great interest. Recent research in this field has involved developing accurate models that estimate the probability of an FHB epidemic based on weather conditions available by satellite across the U.S. However, these predictions ignore two crucial aspects of FHB epidemics: (1) An epidemic can only occur at times when the plants are flowering, and (2) FHB spreads by its spores, resulting in spatial and temporal dependence in risk. We use survey data on flowering dates to produce risk maps that combine weather-based probabilities with estimated flowering dates based on survey data, while simultaneously accounting for spatial and temporal dependence. To allow for scalability, we model spatiotemporal dependence via Higdon's (1998) process convolutions approach. Our approach produces a more realistic assessment of the probability of an FHB epidemic, along with associated estimates of error. We will discuss the application of our methodology in the context of a case study from North Dakota in 2005.

Bayesian Hierarchical Models for Joint Analysis of Gene Expression and Copy Number Data

by
Jen-hwa Chu
Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School
181 Longwood Avenue
Boston, MA 02115
jen-hwa.chu@channing.harvard.edu
Merlise A. Clyde and Feng Liang
Department of Statistical Science, Duke University
PO Box 90251
Durham, NC 27708

Abstract

Array comparative genomic hybridization (CGH) is a technology used to detect DNA copy number alterations which help identify the relevant genes for cancer development. This recent technology development calls for new statistical methods for analyzing array CGH data. Here we propose a Bayesian method to analyze multiple samples at the same time. It is based on hierarchical model with samples grouped according to the disease progression and survival status. The posterior probabilities of copy gain/loss are estimated for each gene at the group level using a reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithm. From the results we can also classify new patients and identify the genes relevant to the group difference. We demonstrate the performance of our method using simulated and real data sets.

This model can be extended to estimate the copy number changes and the association between expression level and DNA copy numbers jointly. The level of association between gene expression level and DNA copy number is defined by a linear regression coefficient θ. From the output we can readily identify genes of which the expression level are associated with copy number changes and those are potential candidates for future research.

We performed analysis on the breast cancer cell line data published by Hyman *et al.* (2002). We provide estimates for both the gene expression levels and DNA copy numbers, along with the degree to which the two types of data are associated. We identify a subset of genes of which the expression levels are most likely attributable to gene copy number alterations across the samples, including some of the oncogenes that were previously associated with breast cancer and some new targets.

Using Predictive Power to Determine Sample Size in an Adaptive Bayesian Clinical Trial

by
Jason T. Connor, Scott Berry and Don Berry
Berry Consultants
117 Westchester Blvd, Noblesville, IN 46062
jason@berryconsultants.com

Abstract

The U.S. FDA requires medical device manufacturers to demonstrate their products are safe and effective. With fixed sample size trials, if *a priori* estimates of effect sizes or variability are poor, these studies may be too small to ensure high statistical power or unnecessarily large and costly.

I illustrate an adaptive design using subjective prior distributions during planned "sample size analyses." After the 50th patient is enrolled, then after every 5 patients up to 100 patients, we look at the interim data (some patients will have reached 6- month primary outcomes some will not) combined with our subjective priors and calculate the predictive probability of each endpoint's success (safety & efficacy) if we stop enrolling now or keep enrolling to our prespecified maximum sample size. We effectively perform power calculations during the course of the trial. If stopping accrual and waiting until all enrolled subjects reach 6-month outcomes provides high power, we stop for early success. If there is low power even if we accrual to 100 patients, we stop for futility.[1]

For example, the manufacturer expects their efficacy and adverse event rates to be $p_T = 0.84$ and $q_T = 0.06$, respectively. The FDA requires $Pr(p_T > p_C - \Delta_p) > 1 - \alpha$ and $Pr(q_T < q_C + \Delta_q) > 1 - \alpha$ for approval ($p_C = 0.7$, $\Delta_p = 0.1$, $q_C = 0.14$, and $\Delta_q = 0.05$). Simulations show this adaptive design has a 95% chance of demonstrating safety and efficacy with an average sample size of 56.4 patients. A classical design with 57 patients has 87% power (same α). If the benefits of the new device are smaller than expected, $q_T = 0.10$, classical power drops to 50% whereas in this adaptive design the average trial size would increase to 66 subjects and power would decrease only to 68%. This loss of power in the adaptive design could be diminished by *a priori* choosing a larger maximum sample size. Using subjective priors during planned sample size analyses ensures we do not make poor early decisions based on limited data. The subjective priors are not used in the final analysis once all outcomes are observed.

Adaptive designs may offer smaller sample sizes in general, and are especially beneficial when trial designers may incorrectly predict efficacy or safety rates, effect sizes, or variability. Adaptive trials will adjust accrual to compensate for higher variability or lower effect sizes, thereby assuring adequate power, or worst case stop the trial early for futility saving the manufacturer time and money. Likewise when variability is smaller than expected or effect sizes are more beneficial than expected, the trial will require fewer patients, helping get the device to market faster.

[1]This trial design has been accepted by FDA and we (Jason Connor, Don Berry & Scott Berry) are completing a manuscript for submission to a statistical journal.

Predicting Event Times in a Masked Clinical Trial Using the Weibull Distribution

by

J. Mark Donovan (1), Michael R. Elliott (2), Daniel F. Heitjan (3)

(1) Bristol-Myers Squibb, (2) University of Michigan, (3) University of Pennsylvania

(1) 311 Pennington-Rocky Hill Road

Pennington, NJ 08534

mark.donovan@bms.com

Abstract

In event-based clinical trials one typically performs interim and final analyses at predetermined event counts. In planning and conducting such a trial, it can be helpful to predict the times of these analyses well in advance of their occurrence. Previous papers described a prediction method that combines prior information with accruing trial data via Bayess theorem. The analysis involves modeling both the time from enrollment to the occurrence of an event and the time from trial commencement until enrollment, which is necessary for predicting event times for subjects who have not yet enrolled. We have extended this work from an exponential-gamma model to incorporate richer families of distributions for both event and enrollment times. We also allow for masking of treatment group membership and blocked randomization, both common design elements in pharmaceutical clinical trials. Our analysis uses latent-class models whose estimation requires use of Markov Chain Monte Carlo (MCMC) sampling with an embedded sampling importance resampling (SIR) step. We applied our analysis to predict time of the 92nd mortality event that would trigger final analysis in REMATCH, a clinical trial of a left ventricular assist device in 129 patients with end stage heart failure. The trial design incorporated both masking and site-based randomization with randomized block sizes. Predictions occurred every 6 months after 1/4 of the planned number of events had occurred. Results were consistent with simulations showing that with sufficient information, our method for prediction yields similar or improved prediction interval coverage probabilities with only modest loss of precision.

Hierarchical Bayesian modeling of marijuana use trajectories in young adults and adolescents

by
Elena A. Erosheva
Department of Statistics
University of Washington
Padelford, Box 354322
Seattle, WA 98195-4322
elena@stat.washington.edu
Donatello Telesca
Department of Statistics
University of Washington
Padelford, Box 354322
Seattle
Washington 98195-4322
telesca@stat.washington.edu
Ross L. Matsueda
Department of Sociology
University of Washington
223J Condon Hall, Box 353340
1100 NE Campus Parkway
Seattle, WA 98195-3340
matsueda@u.washington.edu
and Derek Kreager
Department of Sociology and Crime, Law, and Justice
Pennsylvania State University
dkreager@psu.edu

Abstract

Understanding within- and between-individual variability in longitudinal data on crime and substance use remains an important topic in criminology. A popular method for analyzing life-course crime data accounts for between-individual variability by assuming existence of several distinct groups of offenders, and group-specific polynomial relationships between age and behavior (Roeder, Lynch and Nagin 1999). An extension of this method adds individual-specific random effects such as age, age-squared and age-cubed (Muthen and Shedden 1999). We begin this study with an observation that the usual mixture models with polynomial age-dependence explain little variation in the individual trajectories. We develop an alternative model by assuming the existence of a single natural age-crime curve, and then modeling individual departures from that curve. Following Sampson and Laub (2003), who observed that between-individual heterogeneity in the age-crime relationship "seems to be the age at desistance and level of offending," we introduce individual-specific parameters of phase (to capture desistance) and amplitude (to capture level). We model the natural age-crime curve nonparametrically using B-splines, and factor out individual-specific temporal misalignment and amplitude via Bayesian curve registration methods. We apply this method for analyzing longitudinal data of marijuana use from the Denver Youth Survey. As expected, we find the estimated natural age-crime curve of marijuana use to have one major hump. The individual-specific predictions of marijuana use trajectories fit the observed data reasonably well.

A Case Study of Parallel Model Selection with Hierarchical Structure

by
Xiaodan Fan and Jun S. Liu
Department of Statistics, Harvard University
One Oxford Street
Cambridge, MA 02138
xfan@fas.harvard.edu

Abstract

One type of classification problem involves model selection for multiple objects where the objects of the same model class have shared parameters or parameters governed by a same distribution. In this scenario, the relationship of the model selection part and the parameter estimation part is like the chicken and egg problem, which adds a level of complexity to the already hard model selection problem. Traditional approaches employed a two-step procedure: first gain an estimation of all shared parameters or hyper-parameters from training data or other pre-analysis, then conduct model selection for each object independently. In this paper we developed a Bayesian method to perform all model selection and parameter estimation in an integrative fashion. The task of identifying periodically expressed genes is one case of this problem. Starting from a genome-wide microarray time series data produced by synchronization experiment for yeast cells, we are asked to classify each gene as either periodically expressed or aperiodically expressed. All periodically expressed genes share the same period equal to the inter-division time, the same de-synchronization effect, and the same phase shift between different experiments. The noise variance for all periodically expressed genes share a same distribution with a unknown hyper-parameter. The noise variance for all aperiodically expressed genes also share a same distribution, but with another unknown hyper-parameter. We adopted a Bayesian approach to solve this hierarchical model. MCMC is used to sample the parameter space and the model space iteratively. Reversible Jump MCMC is used to dynamically select models for each gene. Transformation move, group move, block swapping are used to improve the mixing of the chain. The results showed our integrative approach performed much better than the traditional two-step approach.

Bayesian Modeling of Vocabulary Distributions

by
Martí Font, Josep Ginebra, Xavier Puig
Polytechnical University of Catalonia
D. of Statistics, Avgda. Diagonal 647
08028Barcelona, Spain
josep.ginebra@upc.edu, marti.font@upc.edu

Abstract

The zero truncated inverse gaussian-Poisson model is very useful when modeling highly skewed positive integer data like word frequency count data. This model provides a simple mechanistic explanation for this type of data that allows one to interpret the mixing distribution as the vocabulary distribution from which the text was sampled from.

We propose a Bayesian model for this kind of data based on this statistical model, and we fit it to the word frequency count of texts by Macaulay, Lewis Carroll, Wells, Doyle and Dekker. To check the model we look into the posterior distribution of their Pearson errors and carry out posterior predictive consistency checks. We also look into the role of the posterior of the mixing density function in the characterization of literary style.

As an alternative we propose a second model based on the inverse gaussian-zero truncated Poisson model, and we fit it to the same data, check the model and explore the posterior of its mixing distribution.

A GRADE OF MEMBERSHIP MODEL FOR RANK DATA

by
Isobel Claire Gormley & Thomas Brendan Murphy
School of Mathematical Sciences, University College Dublin
Dublin 4
Ireland
claire.gormley@ucd.ie & brendan.murphy@ucd.ie

Abstract

Irish elections use an electoral system known as Proportional Representation by means of a Single Transferable Vote (PR-STV). Under this system voters are required to rank some or all of the candidates in order of their preference. During counting votes are transferred between candidates, according to the voter preferences, as candidates are elected or eliminated from the race. The system is designed to be candidate rather than party orientated and minimizes wasted votes.

The Irish electorate is examined to highlight homogeneous clusters of voters and to determine whether voter preferences are motivated by party or candidate. Moreover, a 'soft' clustering of the electorate is performed such that each voter has an associated probability vector describing the likelihood of their membership of each group within the electorate. This Grade of Membership Model is incorporated with the Plackett-Luce model for ranking data which exploits the information contained in the ranked preferences.

Model fitting is performed within a Bayesian framework subsequent to the imputation of latent variables. An MCMC (Metropolis within Gibbs) sampler is employed to estimate model parameters. Ideas from the MM algorithm are used to construct valid and tractable proposal distributions for the Metropolis step of the algorithm.

Bayesian Robustness in Binary Data for Clinical Trials

by
Jairo A. Fúquene P.
University of Puerto Rico. Department of Mathematics.
jairo.a.fuquene@uprrp.edu

Luis Raúl Pericchi.
University of Puerto Rico. Department of Mathematics.
lrpericchi@uprrp.edu

Abstract

The use of binary data for clinical trials is very common. The usual Bayesian approach for binary data is using Beta/Binomial conjugate analysis. However, in this conjugate analysis the influence of the Beta Prior distribution could be very high leading to non robust results.

Our proposal is using an analysis based on the Cauchy prior for the natural parameter which is more robust than the conjugate analysis diminishing the influence of the prior as a function of the conflict between prior and data. We use strongly the fact that the Binomial likelihood belongs to the exponential family. Finally, this result is illustrated with an example.

Keywords: Bayesian Robustness, Exponential family, Posterior distribution, Clinical Trials.

Functional ANOVA Modeling of Regional Climate Model Experiments

by
Cari G. Kaufman and Stephan R. Sain
Geophysical Statistics Project, National Center for Atmospheric Research
P.O. Box 3000
Boulder, CO 80307
cgk@ucar.edu, ssain@ucar.edu

Abstract

Regional climate models (RCMs) are used by climate scientists to model the evolution of Earth's climate system, using discretized versions of physical processes. These models address smaller spatial regions than do global climate models (GCMs), but their higher resolution better captures the impact of local features such as lakes and mountains. GCM output is often used to provide boundary conditions for RCMs, and it is of interest how much variability in the RCM output is attributable to the choice of RCM, and how much is due simply to large-scale forcing from the GCM. We are analyzing data from the Prudence Project, in which RCMs were crossed with GCM forcings in a designed experiment. To compare model similarities across spatial regions, we propose a functional ANOVA approach, in which we decompose the long-run mean temperature response into a common mean, effects due to regional model, effects due to global model, and interactions. Our hierarchical Bayesian model assigns Gaussian process priors to each of these functional effects, with separate spatial covariance parameters for each effect. Posterior inference can then be carried out either in a summary fashion, by examining the joint posterior of these covariance parameters, or locally, by studying functional and fully Bayesian versions of the usual ANOVA decompositions. For example, although the "prior R^2" in our model is constant across space, time, and model, the "posterior R^2" varies, giving insight into where there is disagreement among climate models and where further model development should focus.

Mode Oriented Stochastic Search for Covariance Estimation in Regional Bank Stock Prices

by
Alex Lenkoski
University of Washington
lenkoski@stat.washington.edu

Abstract

The Banking Industry has undergone extensive consolidation in the past twenty years and the recent rate of mergers has increased considerably. As this implies a convergence in asset valuation across firms, hedge funds are interested in finding small banks whose assets appear undervalued in the market and will thus be targets for acquisition. While depressed price to book ratios suggest arbitrage opportunities, stock price fluctuations along both regional and industry lines are highly prevalent. In order to efficiently mitigate this risk, an accurate estimate of the covariance in stock prices amongst savings, regional and major banks is required.

We consider covariance estimation in a Bayesian framework through Gaussian graphical models with conjugate priors. With well over ten-thousand publicly traded banks, we focus on developing methodology for model selection and model averaging in high dimensions. We develop a novel technique, the Mode Oriented Stochastic Search (MOSS) algorithm, which quickly finds regions of the model space with high posterior probability while evaluating as few irrelevant models as possible.

The use of conjugate priors gives a precise way of scoring models, though with some computational difficulties, as we include nondecomposable graphs in our model space. We review techniques for obtaining normalizing constants of non-decomposable graphs via Monte Carlo integration originally developed in Atay-Kayis and Massam (2005) and also investigate the use of the Laplace approximation to aid the search process.

After model selection has been performed we use the results of Piccioni (2000) to develop a Block Gibbs Sampler, which helps form a model averaged estimator. Comparing results of a simulation study to those reported by Yuan and Lin (2007) we show that the MOSS algorithm performs better than likelihood-based techniques in estimation over a number of models.

After verifying the validity of our approach, we apply the MOSS algorithm to a dataset consisting of savings, regional and major banks across the country, with an eye towards assessing the covariation structure relative to a specific firm of interest. Potential extensions of this approach to Semi-Parametric Gaussian Copulas (Hoff 2007), thus accounting for heavy-tailed marginal distributions, will be discussed.

Joint work with Adrian Dobra, Department of Statistics and Center for Statistics and the Social Sciences, University of Washington.

Selecting a Representative Sample for Calibration of Computer Code for a Circuit Experiment

by
Herbie Lee
Department of Applied Math & Statistics
University of California, Santa Cruz
herbie@ams.ucsc.edu

Abstract

Collaborators at Sandia National Laboratories have run a physical experiment involving the behavior of circuit devices, and plan to use the data to calibrate and validate their computer models, so that the computer simulator can be used for future predictions. The statistical problem here is to select a small representative sample from each of the datasets to be used for the training of the computer model, with the remaining points held out for validation. A mixture model approach is proposed for choosing this sample.

Computer Model Validation with Functional Output

by
Fei Liu
University of Missouri, Columbia
146 Middlebush Hall
University of Missouri, Columbia
Columbia, MO, 65211 liufei@missouri.edu

Abstract

Functional data analysis (FDA) – inference on curves or functions – has wide application in statistics. An example of considerable recent interest arises when considering computer models of processes; the output of such models is a function over the space of inputs of the computer model. The output is functional data in many contexts, such as when the output is a function of time, a surface, etc. A nonparametric Bayesian statistics approach, utilizing separable Gaussian Stochastic Process as the prior distribution for functions, is a natural choice for smooth functions in a manageable (time) dimension. However, direct use of separable Gaussian stochastic processes is inadequate for irregular functions, and can be computationally infeasible in high dimensional cases. In this talk, we will develop and extend several Bayesian FDA approaches for high dimensional irregular functions in the context of computer model validation, tailored to interdisciplinary problems in engineering.

STATISTICS AND LIES: CORRECTING QUESTIONNAIRE ORDERING EFFECT VIA MULTIPLE IMPUTATION

by

Jingchen Liu, Xiao-Li Meng, Chih-nan Chen, and Margarita Alegria

Harvard University, Harvard University, Cambridge Health Alliance, Cambridge Health Alliance

1 Oxford Street, Cambridge, MA 02138

1 Oxford Street, Cambridge, MA 02138

120 Beacon Street, Somerville, MA 02143

120 Beacon Street, Somerville, MA 02143

jcliu@stat.harvard.edu, meng@stat.harvard.edu, cnchen@bu.edu, malegria@charesearch.org

Abstract

National Latino and Asian American Study (NLAAS) is a complex survey of psychiatric epidemiology, with multiple embedded experiments with alternative question forms. One objective of multiple imputation is to create analytic datasets corrected for response biases due to defects of the survey instrument, such as increasing rates of negative responses over the course of the interview induced by the respondents' learning to use skip patterns to reduce interview time. Multiple imputations are sampled independently from the posterior distribution to facilitate common analysis. The imputation modeling task is particularly challenging because of the existence of high-order interactions among self-reported psychiatric service use variables. As our preliminary works show, the classic multivariate probit model turns out to be not rich enough to fit the data pattern, since it only includes pairwise correlations among two variables. An extended family of distributions is used to capture the high-order interactions. This family uses the multi-probit models as the building blocks and also inherits its latent structure of multivariate Gaussian distribution to facilitate sampling from the posterior distribution. It is motivated by the continuation ratio model wildly used in censored survival data analysis and originally proposed by Cox (1972). Further challenges lie in the complexity of the questionnaire, the small sample sizes for subgroups of interests, forming a noninformative prior for nonlinear models, and the posterior simulation of complicated models. Future studies will be focused on exploring properties of the extended family, improving the efficiency of the corresponding sampling for more general applications.

The Impact of Identification Conditions on the Forecasting of Seasonal Adjusted Series

by
Christian Macaro
Roma Tre University and New York University
christianmacaro@gmail.com

Abstract

This work aims to present a full Bayesian framework to identify, extract and forecast unobserved components in time series. The major novelty of the approach is the definition of a probabilistic framework to analyze the identification conditions. More precisely, informative prior distributions are assigned to the spectral densities of the unobserved components. This entails a interesting feature: the possibility to analyze more than one decomposition at once by studying the posterior distributions of the unobserved spectra. Particular attention is given to an empirical application where the canonical decomposition of sunspot data is compared with some alternative decompositions. The posterior distributions of the unobserved components are recovered by exploiting some recent developments in the Wiener-Kolmogorov and circular process literature. An empirical application shows how to capture the seasonal component in the volatility of financial high frequency data. The posterior forecasting distributions are finally recovered by exploiting a relationship between spectral densities and linear processes. An empirical application shows how to forecast seasonal adjusted financial time series.

Parameter Sensitivity Analysis on Influence Network

by
Il-Chul Moon, Eunice J. Kim and Kathleen M. Carley
Carnegie Mellon, School of Computer Science, CASOS
5000 Forbes Avenue, 1325 Wean Hall, Pittsburgh, PA, 15213
imoon@andrew.cmu.edu, eunicek@andrew.cmu.edu, kathleen.carley@cs.cmu.edu

Abstract

Influence network is one of the semi-Bayesian networks extensively used for Effects-Based Operation. The network represents the causalities of event occurrences, and it models the likelihood of each event by propagating the promotion or inhibition of one event to another. Often, this network is built by subject matter experts by hand. Thus, we have introduced the influence network generator in Organization Risk Analyzer, a social network analysis software package. The generator produces an influence network of a multi-mode, multi-plex social network from an event-flow and organizational management perspective. To support the soundness of this automated generation, we analyze the sensitivity of baseline probability, the major parameter of the model, by using Gibbs sampling. This sensitivity analysis reveals how a little change in the baseline probability assignment shifts the conclusion we draw, hence we obtain the comfort range of automatic baseline probability generation.

Integrating satellite and monitoring data to estimate air pollution concentrations

by
Christopher Paciorek
Department of Biostatistics
Harvard School of Public Health
Boston, MA
paciorek@hsph.harvard.edu

Abstract

Satellite retrievals of aerosol optical depth (AOD) hold promise for helping to estimate ground-level particulate matter (PM), receiving much attention in the environmental science community. However, AOD is a biased and noisy proxy for PM2.5. Bayesian statistical techniques provide a natural framework to integrate the two sources of information, while accounting for the shortcomings of each, with estimates then used in health analyses. We present raw comparisons of the monitoring and satellite data to help understand the bias of AOD as a proxy for PM2.5. Based on this, we describe a hierarchial spatial model with a latent process representing pollution and separate likelihoods for the monitoring and satellite data. The model is specified such that computation via MCMC is effective. The key issue in the model specification is whether the AOD bias is spatially correlated. Results assuming spatially correlated bias differ dramatically from those assuming independent errors, as has been done in analogous settings. Finally, we compare predictions of PM2.5 using the integrated model to those using monitoring data only to understand the amount of information added by the AOD observations.

Subjective Prior Distributions for Modeling Change in Substance Abuse Treatment Process Scores with Non-Ignorable Dropout

by
Susan M. Paddock and Patricia Ebener
RAND Corporation
1776 Main Street
Santa Monica, CA 90401
paddock@rand.org

Abstract

Substance abuse treatment research is complicated by the pervasive problem of non-ignorable missing data i.e., the occurrence of the missing data is related to the unobserved outcomes. Missing data frequently arise due to early client departure from treatment. Pattern-mixture models (PMMs) are often employed in such situations to jointly model the outcome and the missing data mechanism. Because PMMs require non-testable assumptions to identify model parameters, several approaches to parameter identification have been explored for longitudinal, continuous outcomes, and informative priors have been developed in other contexts. In this presentation, we describe an expert interview conducted with five substance abuse treatment clinical experts who have familiarity with the Therapeutic Community modality of substance abuse treatment and with treatment process scores collected using the Dimensions of Change Instrument. The goal of the interviews was to obtain expert opinion about the rate of change in client-level treatment process scores for clients who leave before completing two assessments and whose rate of change (slope) in treatment process scores is unidentified by the data. We find that the experts opinions differed dramatically from widely-utilized assumptions used to identify parameters in the PMM. Further, subjective prior assessment allows one to properly address the uncertainty inherent in the subjective decisions required to identify parameters in the PMM and to measure their effect on conclusions drawn from the analysis.

Characteristics of the Anterior Dental Arch in Adult Males: Applications to Bitemark Analysis in Forensic Odontology

Nicholas M. Pajewski, L. Thomas Johnson, Thomas Radmer*, and Purushottam W. Laud*
Medical College of Wisconsin
Division of Biostatistics
8701 Watertown Plank Rd
Milwaukee, WI 53226
npajewski@mcw.edu

Abstract

Bitemark analysis represents one of the comparative forensic sciences, e.g. fingerprint comparison, footwear patterns and ballistics. It commonly entails matching bitemark evidence from crime scenes with the same pattern of a potential suspect. Although dentists have assumed that individual dentition patterns are unique, because bitemarks usually involve an individual's six front, or anterior teeth, this had led the courts to question whether two individuals could create highly similar bitemark patterns. This issue has led to research on quantifying the population distribution of the anterior teeth. The primary goal of this research is to provide odontologists with an epidemiologic tool in evaluating the "rarity" of bite mark patterns,thereby allowing them to comment on the likelihood that a bite mark originated from another individual.
The current study targeted volunteer adult males between the ages of 18 to 44. Males within this age group tend to be the most frequent suspects in cases involving bitemark evidence. Anonymous dental imprints of the biting edges of the anterior teeth were obtained from 419 subjects and then measured for three characteristics: arch width,and the width and rotation of the incisors. From a statistical perspective, there are two primary difficulties with the analysis of these measurements. Because the measurements are taken manually, they are subject to measurement error both within and between examiners. Second, the measurements are not independent because of the inherent symmetry in the layout of human teeth.

We consider parametric and non-parametric Bayesian hierarchical models that treat the observed values of the measurements as arising from a latent factor. We illustrate how the models permit explicit probability statements about the distribution of the measurements while accounting for measurement error and the correlation between measurements. Future methodological work for this application could entail the modeling of specific covariance structures within non-parametric frameworks, as Graphical Association Models have revealed that the anterior teeth involve complex correlation patterns.

* From the School of Dentistry, Marquette University

Interpreting Uncertain DNA Evidence

by

Mark W. Perlin (1), Joseph B. Kadane (2), Robin W. Cotton (3), Alexander Sinelnikov (1)

(1) Cybergenetics, (2) Statistics Department, Carnegie Mellon University, (3) Biomedical Forensic Sciences, Boston University

160 North Craig Street, Suite 210

Pittsburgh, PA 15213

perlin@cybgen.com

Abstract

DNA typing is a powerful method of comparing biological specimens and establishing genetic identity. However, in actual forensic casework, there is typically considerable uncertainty in the underlying quantitative DNA data. This uncertainty is currently addressed by qualitative "include/exclude" interpretation methods which can discard much of the identity information present in the data. These methods can be improved upon through the use of a quantitative likelihood function that models the data. This poster introduces a general probability framework for describing DNA interpretation methods, and compares the identification power of leading methods on representative uncertain evidence. We demonstrate that better modeling of the data can greatly improve DNA identification.

Bayesian analysis of functional Magnetic Resonance Images

by
Alicia Quirós and Raquel Montes Diez
Departamento de Estadística e Investigación Operativa
Universidad Rey Juan Carlos
Madrid (Spain)
alicia.quiros@urjc.es, raquel.montes@urjc.es
Dani Gamerman
Instituto de Matemática
Universidade Federal do Rio de Janeiro
Rio de Janeiro (Brazil)
dani@im.ufrj.br

Abstract

In this work, we are interested in analysing functional Magnetic Resonance Imaging (fMRI) data, in order to detect active areas into the brain and by making use of Bayesian inference. In a fMRI study, several brain scans are acquired over time, half of which under a certain stimulation paradigm, and the other half during a rest period. In order to detect the activity, we set the basis of our model in the well known fact that active areas emit higher signals in response to an stimulus that non-activated do.

Bayesian inference provides an ideal framework to face this problem, allowing us to include prior information about the model parameters. To model the hemodynamic response, in the temporal dimension, we use a parametrization similar to the empirical curve accepted by the scientific community. In the spatial dimension, for the parameter stating the presence or absence of activity in each pixel and the one modelling the activity magnitude, we choose Markov Random Fields as prior distributions. These powerful tools provide a framework to detect active regions much as a neurologist might as they take into account the magnitude level in a voxel as well as the size of the activity area.

A spatio-temporal model allows us to obtain more information about the Magnetic Resonance study. In spite of the high computational cost, a spatio-temporal model improves the inference ability as it takes into account both the uncertainty in the spatial dimension and in the temporal one.

A Bayesian Model of Interpersonal Transactions between Two Parties

by
Eiki Satake and Philip Amato
Emerson College
eiki satake@emerson.edu

Abstract

This paper presents a paradigm and model developed by the authors that demonstrates the use of Empirical Bayesian Methods in the analysis of interpersonal transactions as an alternative to time series analysis. Specifically, the model uses the Beta-Binomial distribution as predictive distribution to estimate the likelihood of a hypothesis, supported by several incidences or events. The import of prediction in communication theory with the Empirical Bayesian data analysis is illustrated within the framework of an interpersonal communication transaction between two parties.

Multiple Testing and Covariance Selection: A Case Study on Mutual Funds

by
James Scott
Duke University
Box 90251
Durham, NC 27708
james@stat.duke.edu

Abstract

A crucial input for many dynamic portfolio-selection problems is the estimated covariance matrix for a collection of asset returns. Large portfolios can often yield highly variable estimates for this matrix (Polson and Tew, 2000), and graphical models offer a potent tool for regularization and stabilization. In deciding which graph (or set of graphs) to use for imposing graphical constraints, one must inevitably confront the issue of multiple hypothesis testing where each null hypothesis corresponds to the exclusion of a single edge from the graph. We illustrate the importance of fully Bayesian multiplicity correction on an example involving a set 86 monthly returns for 59 mutual funds from a variety of different sectors: 13 U.S. bond bunds, 30 U.S. stock funds, 7 balanced funds investing in both U.S. stocks and bonds, and 9 international stock funds. We study the predictive performance of corrected vs. uncorrected model-selection procedures and show that multiplicity correction yields estimates with superior predictive performance.

Modelling Birthweight in the Presence of Gestational Age Measurement Error - A Semi-parametric Approach

by

Russell Steele[1], Robert Platt[2], Michelle Ross[3]

[1] Dept. of Mathematics and Statistics, McGill University

[2] Dept. of Epidemiology and Biostatistics, McGill University

[3] Dept. of Biostatistics, University of Washington

805 Rue Sherbrooke O.

Montréal, Québec Canada

Email: steele@math.mcgill.ca

Abstract

Gestational age is an important variable in perinatal research, as it is a strong predictor of mortality and other adverse outcomes, and is also a component of measures of fetal growth. Recently, several authors have proposed the use of gestational age as a time axis rather than a covariate. However, gestational ages measured using the date of the last menstrual period (LMP) are prone to substantial errors. These errors are apparent in most population-based data sources, which often show such implausible features as a bimodal distribution of birth weight at early preterm gestational ages ($\leq$ 34 weeks) and constant or declining mean birth weight at postterm gestational ages ($\geq$ 42 weeks). These features are likely consequences of errors in gestational age. Gestational age plays a critical role in measurement of outcome (preterm birth, small for gestational age) and is an important predictor of subsequent outcomes. It is important in the development of fetal growth standards. Therefore, accurate measurement of gestational age, or, failing that, a reasonable understanding of the structure of measurement error in the gestational age variable, is critical for perinatal research.

In this paper, we will present a straightforward approach for adjusting for gestational age measurement error via a semi-parametric approximation to the functional relationship between gestational age and birthweight. Certain assumptions about the distribution of true gestational ages and the structure of the problem allow us to compare several reasonable models. A Bayesian approach to estimating the measurement error distribution improves the analysis, in particular, by incorporating prior information about the probability

Computational Approaches for Parameter Uncertainty Estimation in Climate Models

by

Alejandro Villagran, Gabriel Huerta, Charles S. Jackson and Mrinal K. Sen

Department of Mathematics and Statistics, The University of New Mexico and Institute for Geophysics, The University of Texas at Austin.

Department of Mathematics and Statistics, The University of New Mexico

Albuquerque, NM 87131

ghuerta@stat.unm.edu

Abstract

Intensive computational methods have been used by Earth scientists for more than 30 years. They have been applied to a wide range of problems, from Earthquake epicenter location to Climate prediction modeling. To quantify the uncertainties resulting from a range of model configurations is necessary to estimate a multidimensional probability distribution. The computational cost of evaluating a multidimensional probability distribution for a climate model is impractical using traditional methods (Metropolis/Gibbs algorithms). Multiple Very Fast Simulated Annealing (MVFSA) is an efficient but approximate algorithm to tackle this problem. This article describes the development and application of different intensive computation techniques used in geophysical inverse problems. Besides MVFSA, we apply Adaptive Metropolis methods to a surrogate climate model that is able to approximate the noise and response behavior of a realistic atmospheric general circulation model (GCM). This AGCM models the surface air temperature of the Earth by considering the variability of its three orbital forcing parameters (Obliquity, Longitude of Perihelion and Eccentricity). In this case the geophysical inversion problem is to find the posterior probability distribution of the parameters given the observed data (temperature) at different latitudes, longitudes and seasons.

We compare the performance of each algorithm by introducing a new measure called the Root Mean Squared (RMS) probability error which combines the CPU time spent of each algorithm and its convergence to a target distribution. We analyze both advantages and drawbacks in each method, but mainly we focus on optimal uncertainty estimation, from a Bayesian point of view.

Assessing Convergence of MCMC using Validation Techniques

by
Dawn Woodard
Duke University Dept. Stat. Sci.
Box 90251, Duke University
Durham, NC 27708
dawn@stat.duke.edu

Abstract

Computation in Bayesian statistical models is often performed using Markov chain Monte Carlo (MCMC) methods. The convergence of the Markov chain is typically assessed using a set of standard diagnostics; recent draft FDA guidelines for the use of Bayesian statistics in medical device trials, for instance, advocate this approach for validating computations.

We give several examples showing that this approach may be insufficient when the posterior distribution is multimodal–that lack of convergence due to posterior multimodality can go undetected using the standard convergence diagnostics. We detect the problem by applying validation techniques that were originally proposed for detecting coding and analytical errors in MCMC software (Cook, Gelman and Rubin, 2006 and Geweke, 2004). We then argue that these validation techniques should be widely applied when using MCMC for computation in models where posterior unimodality is not guaranteed. In applications where accuracy is imperative, such as medical device trials, such measures would be most important.

Modeling epilepsy disparities among ethnic groups in Philadelphia, PA

by

David C. Wheeler[1]*, Lance A. Waller*[1]*, and John O. Elliott*[2]

[1]*Department of Biostatistics, Rollins School of Public Health, Emory University;*

[2]*Comprehensive Epilepsy Center, Department of Neurology, The Ohio State University*

Abstract

The Centers for Disease Control and Prevention (CDC) defined epilepsy as an emerging public health issue in a recent report and emphasized the importance of epilepsy studies in minorities and people of low socioeconomic status. Previous research has suggested that the incidence rate for epilepsy is positively associated with various measures of social and economic disadvantage. In response, we utilize and compare several different hierarchical Bayesian models to analyze health disparities in epilepsy and seizure risks among multiple ethnicities in the city of Philadelphia, Pennsylvania, while modeling shared risk between ethnicities. The goals of the analysis are to highlight any overall significant disparities in epilepsy risks between the populations of Caucasians, African Americans, and Hispanics in the study area during the years 2002-2004 and to visualize the spatial pattern of epilepsy risks by ethnicity to indicate where certain ethnic populations were most adversely effected by epilepsy within the study area. Results of the Bayesian model indicate that Hispanics have the highest epilepsy risk overall, followed by African Americans, and then Caucasians. Also, there are significant increases in relative risk for both African Americans and Hispanics when compared with Caucasians. The results demonstrate that using a Bayesian analysis in combination with geographic information system technology can reveal spatial patterns in patient data and highlight areas of disparity in epilepsy risk among subgroups of the population. The results also show that a model that explicitly models shared risk between ethnicities for epilepsy fits the data the best when compared with other models.

Keywords: hierarchical Bayesian regression, health disparities, spatial epidemiology, MCMC, spatial statistics, epilepsy

On Combining Computer Model Output and Ground Level Ozone Concentration Data for Improved Forecasts

by
Chun-yin Yip, Sujit Sahu
University of Southampton
School of Mathematics, University of Southampton,
Southampton, ENGLAND SO17 1BJ
cyy102@soton.ac.uk

Abstract

Ground level ozone is a common air pollutant which is formed indirectly by car engines, industrial boilers, power plants, refineries; these sources emit hydrocarbons and nitrogen oxides(NO_x) that react chemically in the presence of sunlight. The Environmental Protection Agency of the United States have designed the Community Multi-scale Air Quality modelling system (CMAQ) to forecast levels of various air pollutants such as ground level ozone. The CMAQ forecast is not a statistical model but rather a numerical deterministic differential equations model based on emission, meteorology, transportation dynamics and ground characteristics affecting the level of the pollutant. It is well known, however, these forecasts are biased. In this talk, we present a set of models to incorporate information from the CMAQ to obtain improved forecasts. Throughout the investigation, we often deal with some tradeoffs between a simple model and a model with higher computational burden. Different modelling strategies will also be presented.

Nonparametric Hierarchical Bayes Analysis of Binomial Data Via Bernstein Polynomial Priors

by
Tingting Zhang, Jun Liu
Harvard University
1 Oxford Street 906
Cambridge, MA
tzhang@fas.harvard.edu

Abstract

Since the influential work of Ferguson (1973), the term "nonparametric Bayes (NB) analysis" is almost synonymous of assuming a Dirichlet process prior on an unknown probability distribution. The recent work of Petrone (1999a,b,2002) introduces Bernstein polynomials into the NB framework and opens up new possibilities. We here examine Bernstein-Dirichlet process-based nonparametric hierarchical Bayes procedures for analyzing binomial data. Under this setting, we find that the predictive density of a future binomial observation can be expressed as a weighted mixture of beta densities, which is absolutely continuous. We compare performances of the new NB approach with a previous method based purely on the Dirichlet process through examples, and show that the Bernstein polynomial-based estimates are more robust to the sample variation and smoother than estimates with the Dirichlet process prior.

Statistical Model for Detecting Multiple eQTLs

by
Wei Zhang, Jun Zhu, Jun Liu
Statistics Department, Harvard University
1 Oxford St
Cambridge, MA 02138
weizhang@fas.harvard.edu

Abstract

Treating mRNA transcript abundances as quantitative traits and mapping gene expression quantitative trait loci for these traits has been studied in many organisms. In various data analysis, people have identified significant associations between gene expression and genetic markers and the results can help discover important regulation pathways. Due to the large number of gene expression values and genetic markers, it is still a challenging question to researchers where these associations are and how eQTLs affect expression levels. We will present a statistical model to describe the associations between gene expression and genetic markers. Unlike tradition eQTL analysis, this method treats genes with similar expression values and linked with similar markers as a module. Information from genes within the same module can be borrowed from each other to enhance the detection of the linked markers. The linkage is defined in a statistical way such that marker interactions are automatically considered. We use the MCMC method and some advanced technique to search over the space of all possible gene-marker partitions. We show from simulation studies that this method achieves better power compared to traditional methods. Further research will be focused on efficient sampling schemes and application to real data.