THE ANNALS
of
APPLIED
STATISTICS
AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Fast dimension-reduced climate model calibration and the effect of data aggregation
WON CHANG, MURALI HARAN, ROMAN OLSON and KLAUS KELLER 649

Estimation in the partially observed stochastic Morris–Lecar neuronal model with particle
filter and stochastic approximation methods .......... SUSANNE DITTELYSEN and
ADELINE SAMSON 674

Effect of breastfeeding on gastrointestinal infection in infants: A targeted maximum
likelihood approach for clustered longitudinal data ...... MIREILLE E. SCHNITZER,
MARK J. VAN DER LAAN, ÉRICA E. M. MOODIE and ROBERT W. PLATT
Maximum likelihood and pseudo score approaches for parametric time-to-event analysis
with informative entry times ...... BRIAN D. M. TOM, VERNON T. FAREWELL and
SHEILA M. BIRD 726

Clustering South African households based on their asset status using latent variable
models ..................... DAMIEN MCPARLAND, ISOBEL CLAIRE GORMLEY,
TYLER H. MCCORMICK, SAMUEL J. CLARK, CHODZIWADZIWA WHITESON
KABUDULA and MARK A. COLLINSON 747

Hypothesis setting and order statistic for robust genomic meta-analysis
CHI SONG and GEORGE C. TSENG 777

Testing the disjunction hypothesis using Voronoi diagrams with applications to genetics
DAISY PHILLIPS and DEBASHIS GHOSH 801

Detection boundary and Higher Criticism approach for rare and weak genetic effects
ZHEYANG WU, YIMING SUN, SHIQUAN HE, JUDY CHO, HONGYU ZHAO and
JIASHUN JIN 824

Small area estimation of general parameters with application to poverty indicators:
A hierarchical Bayes approach ...... ISABEL MOLINA, BALGOBIN NANDRAM and
J. N. K. RAO 852

Estimation of nonlinear differential equation model for glucose–insulin dynamics in
type I diabetic patients using generalized smoothing ........... INNA CHERVONEVA,
BORIS FREYDIN, BRIAN HIPSZER, TATIYANA V. APANASOVICH and
JEFFREY I. JOSEPH 886

Pairwise comparison of treatment levels in functional analysis of variance with
application to erythrocyte hemolysis
OLGA VSEVOLOZHSKAYA, MARK GREENWOOD and DMITRI HOLODOV 905

MONEYBaRL: Exploiting pitcher decision-making using
Reinforcement Learning ......................... GAGAN SIDHU and BRIAN CAFFO
Adjusting models of ordered multinomial outcomes for nonignorable nonresponse in the
occupational employment statistics survey ............... NICHOLAS J. HORTON,
DANIEL TOOTH and POLLY PHIPPS 956

Leveraging local identity-by-descent increases the power of case/control GWAS with
related individuals ...... JOSHUA N. SAMPSON, BILL WHEELER, PENG LI and
JIAXIN SHI 974

continued
A Bayesian nonparametric mixture model for selecting genes and gene subnetworks
YIZE ZHAO, JIAN KANG AND TIANWEI YU 999

Statistical calibration of qRT-PCR, microarray and RNA-Seq gene expression data with measurement error models.
ZHAONAN SUN, THOMAS KUCZEK AND YU ZHU 1022

Regularized 3D functional regression for brain image data via Haar wavelets
XUEJING WANG, BIN NAN, JI ZHU, ROBERT KOEPPE AND FOR THE ALZHEIMER’S DISEASE NEUROIMAGING INITIATIVE 1045

Voxel-level mapping of tracer kinetics in PET studies: A statistical approach emphasizing tissue life tables.
FINBARR O’SULLIVAN, MARK MUZI, DAVID A. MANKOFF, JANET F. EARY, ALEXANDER M. SPENCE AND KENNETH A. KROHN 1065

Analysis of multiple sclerosis lesions via spatially varying coefficients.
TIAN GE, NICOLE MÜLLER-LENKE, KERSTIN BENDFELDT, THOMAS E. NICHOLS AND TIMOTHY D. JOHNSON 1095

A statistical approach to the inverse problem in magnetoencephalography
ZHIGANG YAO AND WILLIAM F. EDDY 1119

Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes.
FRANÇOIS CARON, YEE WHYE TEH AND THOMAS BRENDAN MURPHY 1145

Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint.
JING QIN, TANYA P. GARCIA, YANYUAN MA, MING-XIN TANG, KAREN MARDER AND YUANJIA WANG 1182

A new method of peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data.
SEONGHO KIM, MING OUYANG, JAESIK JEONG, CHANGYU SHEN AND XIANG ZHANG 1209

Gene-level pharmacogenetic analysis on survival outcomes using gene-trait similarity regression.
JUNG-YING TZENG, WENBIN LU AND FANG-CHI HSU 1232

VILLE A. SATOPÄÄ, SHANE T. JENSEN, BARBARA A. MELLERS, PHILIP E. TETLOCK AND LYLE H. ÜNGAR 1256

Vol. 8, No. 2—June 2014
FAST DIMENSION-REDUCED CLIMATE MODEL CALIBRATION AND THE EFFECT OF DATA AGGREGATION

BY WON CHANG, MURALI HARAN, ROMAN OLSON AND KLAUS KELLER

Pennsylvania State University

How will the climate system respond to anthropogenic forcings? One approach to this question relies on climate model projections. Current climate projections are considerably uncertain. Characterizing and, if possible, reducing this uncertainty is an area of ongoing research. We consider the problem of making projections of the North Atlantic meridional overturning circulation (AMOC). Uncertainties about climate model parameters play a key role in uncertainties in AMOC projections. When the observational data and the climate model output are high-dimensional spatial data sets, the data are typically aggregated due to computational constraints. The effects of aggregation are unclear because statistically rigorous approaches for model parameter inference have been infeasible for high-resolution data. Here we develop a flexible and computationally efficient approach using principal components and basis expansions to study the effect of spatial data aggregation on parametric and projection uncertainties. Our Bayesian reduced-dimensional calibration approach allows us to study the effect of complicated error structures and data-model discrepancies on our ability to learn about climate model parameters from high-dimensional data. Considering high-dimensional spatial observations reduces the effect of deep uncertainty associated with prior specifications for the data-model discrepancy. Also, using the unaggregated data results in sharper projections based on our climate model. Our computationally efficient approach may be widely applicable to a variety of high-dimensional computer model calibration problems.

REFERENCES


Key words and phrases. Climate model, calibration, Gaussian process, principal components, high-dimensional spatial data.


SIMPSON, D., LINDGREN, F. and RUE, H. (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. Environmetrics 23 65–74. MR2873784


ESTIMATION IN THE PARTIALLY OBSERVED STOCHASTIC MORRIS–LECAR NEURONAL MODEL WITH PARTICLE FILTER AND STOCHASTIC APPROXIMATION METHODS

BY SUSANNE DITLEVSEN*,1 AND ADELINE SAMSON†,‡,2

University of Copenhagen*, Université Paris Descartes†
and Université Grenoble-Alpes‡

Parameter estimation in multidimensional diffusion models with only one coordinate observed is highly relevant in many biological applications, but a statistically difficult problem. In neuroscience, the membrane potential evolution in single neurons can be measured at high frequency, but biophysical realistic models have to include the unobserved dynamics of ion channels. One such model is the stochastic Morris–Lecar model, defined by a nonlinear two-dimensional stochastic differential equation. The coordinates are coupled, that is, the unobserved coordinate is nonautonomous, the model exhibits oscillations to mimic the spiking behavior, which means it is not of gradient-type, and the measurement noise from intracellular recordings is typically negligible. Therefore, the hidden Markov model framework is degenerate, and available methods break down. The main contributions of this paper are an approach to estimate in this ill-posed situation and nonasymptotic convergence results for the method. Specifically, we propose a sequential Monte Carlo particle filter algorithm to impute the unobserved coordinate, and then estimate parameters maximizing a pseudo-likelihood through a stochastic version of the Expectation–Maximization algorithm. It turns out that even the rate scaling parameter governing the opening and closing of ion channels of the unobserved coordinate can be reasonably estimated. An experimental data set of intracellular recordings of the membrane potential of a spinal motoneuron of a red-eared turtle is analyzed, and the performance is further evaluated in a simulation study.

REFERENCES


Key words and phrases. Sequential Monte Carlo, diffusions, pseudo likelihood, Stochastic Approximation Expectation Maximization, motoneurons, conductance-based neuron models, membrane potential.


EFFECT OF BREASTFEEDING ON GASTROINTESTINAL INFECTION IN INFANTS: A TARGETED MAXIMUM LIKELIHOOD APPROACH FOR CLUSTERED LONGITUDINAL DATA

BY MIREILLE E. SCHNITZER1,∗, MARK J. VAN DER LAAN2,†, ERICA E. M. MOODIE3,‡ AND ROBERT W. PLATT4,‡

Université de Montréal∗, University of California, Berkeley†
and McGill University‡

The PROmotion of Breastfeeding Intervention Trial (PROBIT) cluster-randomized a program encouraging breastfeeding to new mothers in hospital centers. The original studies indicated that this intervention successfully increased duration of breastfeeding and lowered rates of gastrointestinal tract infections in newborns. Additional scientific and popular interest lies in determining the causal effect of longer breastfeeding on gastrointestinal infection. In this study, we estimate the expected infection count under various lengths of breastfeeding in order to estimate the effect of breastfeeding duration on infection. Due to the presence of baseline and time-dependent confounding, specialized “causal” estimation methods are required. We demonstrate the double-robust method of Targeted Maximum Likelihood Estimation (TMLE) in the context of this application and review some related methods and the adjustments required to account for clustering. We compare TMLE (implemented both parametrically and using a data-adaptive algorithm) to other causal methods for this example. In addition, we conduct a simulation study to determine (1) the effectiveness of controlling for clustering indicators when cluster-specific confounders are unmeasured and (2) the importance of using data-adaptive TMLE.

REFERENCES


Key words and phrases. Causal inference, G-computation, inverse probability weighting, marginal effects, missing data, pediatrics.


MAXIMUM LIKELIHOOD AND PSEUDO SCORE APPROACHES FOR PARAMETRIC TIME-TO-EVENT ANALYSIS WITH INFORMATIVE ENTRY TIMES

BY BRIAN D. M. TOM, VERNON T. FAREWELL AND AND SHEILA M. BIRD

MRC Biostatistics Unit

We develop a maximum likelihood estimating approach for time-to-event Weibull regression models with outcome-dependent sampling, where sampling of subjects is dependent on the residual fraction of the time left to developing the event of interest. Additionally, we propose a two-stage approach which proceeds by iteratively estimating, through a pseudo score, the Weibull parameters of interest (i.e., the regression parameters) conditional on the inverse probability of sampling weights; and then re-estimating these weights (given the updated Weibull parameter estimates) through the profiled full likelihood. With these two new methods, both the estimated sampling mechanism parameters and the Weibull parameters are consistently estimated under correct specification of the conditional referral distribution. Standard errors for the regression parameters are obtained directly from inverting the observed information matrix in the full likelihood specification and by either calculating bootstrap or robust standard errors for the hybrid pseudo score/profiled likelihood approach. Loss of efficiency with the latter approach is considered. Robustness of the proposed methods to misspecification of the referral mechanism and the time-to-event distribution is also briefly examined. Further, we show how to extend our methods to the family of parametric time-to-event distributions characterized by the generalized gamma distribution. The motivation for these two approaches came from data on time to cirrhosis from hepatitis C viral infection in patients referred to the Edinburgh liver clinic. We analyze these data here.

REFERENCES


Key words and phrases. Biased data, generalized gamma distribution, outcome-dependent sampling, pseudo score, robust standard error, survival analysis, Weibull distribution.


TOM, B. D. M., FAREWELL, V. T. and BIRD, S. M. (2014). Supplement to “Maximum likelihood and pseudo score approaches for parametric time-to-event analysis with informative entry times.” DOI:10.1214/14-AOAS725SUPP.


CLUSTERING SOUTH AFRICAN HOUSEHOLDS BASED ON THEIR ASSET STATUS USING LATENT VARIABLE MODELS

BY DAMIEN MCPARLAND1,*, ISOBEL CLAIRE GORMLEY1,*, TYLER H. MCCORMICK3,†, SAMUEL J. CLARK2,‡,§,¶, CHODZIWADZIWA WHITESON KABUDULA‡ AND MARK A. COLLINSON‡,¶

University College Dublin*, University of Washington†, University of the Witwatersrand‡, University of Colorado at Boulder§ and INDEPTH Network¶

The Agincourt Health and Demographic Surveillance System has since 2001 conducted a biannual household asset survey in order to quantify household socio-economic status (SES) in a rural population living in northeast South Africa. The survey contains binary, ordinal and nominal items. In the absence of income or expenditure data, the SES landscape in the study population is explored and described by clustering the households into homogeneous groups based on their asset status.

A model-based approach to clustering the Agincourt households, based on latent variable models, is proposed. In the case of modeling binary or ordinal items, item response theory models are employed. For nominal survey items, a factor analysis model, similar in nature to a multinomial probit model, is used. Both model types have an underlying latent variable structure—this similarity is exploited and the models are combined to produce a hybrid model capable of handling mixed data types. Further, a mixture of the hybrid models is considered to provide clustering capabilities within the context of mixed binary, ordinal and nominal response data. The proposed model is termed a mixture of factor analyzers for mixed data (MFA-MD).

The MFA-MD model is applied to the survey data to cluster the Agincourt households into homogeneous groups. The model is estimated within the Bayesian paradigm, using a Markov chain Monte Carlo algorithm. Intuitive groupings result, providing insight to the different socio-economic strata within the Agincourt region.

REFERENCES


Key words and phrases. Clustering, mixed data, item response theory, Metropolis-within-Gibbs.


Rutstein, S. O. and Johnson, K. (2004). The DHS wealth index. DHS comparative Reports No. 6, ORC Macro, Calverton, MD.


HYPOTHESIS SETTING AND ORDER STATISTIC FOR ROBUST GENOMIC META-ANALYSIS

BY CHI SONG AND GEORGE C. TSENG

University of Pittsburgh

Meta-analysis techniques have been widely developed and applied in genomic applications, especially for combining multiple transcriptomic studies. In this paper we propose an order statistic of p-values (rth ordered p-value, rOP) across combined studies as the test statistic. We illustrate different hypothesis settings that detect gene markers differentially expressed (DE) “in all studies,” “in the majority of studies” or “in one or more studies,” and specify rOP as a suitable method for detecting DE genes “in the majority of studies.” We develop methods to estimate the parameter r in rOP for real applications. Statistical properties such as its asymptotic behavior and a one-sided testing correction for detecting markers of concordant expression changes are explored. Power calculation and simulation show better performance of rOP compared to classical Fisher’s method, Stouffer’s method, minimum p-value method and maximum p-value method under the focused hypothesis setting. Theoretically, rOP is found connected to the naïve vote counting method and can be viewed as a generalized form of vote counting with better statistical properties. The method is applied to three microarray meta-analysis examples including major depressive disorder, brain cancer and diabetes. The results demonstrate rOP as a more generalizable, robust and sensitive statistical framework to detect disease-related markers.

REFERENCES


Key words and phrases. Genomics, meta-analysis, order statistic, p-value.


Testing the disjunction hypothesis is appropriate when each gene or location studied is associated with multiple \( p \)-values, each of which is of individual interest. This can occur when more than one aspect of an underlying process is measured. For example, cancer researchers may hope to detect genes that are both differentially expressed on a transcriptomic level and show evidence of copy number aberration. Currently used methods of \( p \)-value combination for this setting are overly conservative, resulting in very low power for detection. In this work, we introduce a method to test the disjunction hypothesis by using cumulative areas from the Voronoi diagram of two-dimensional vectors of \( p \)-values. Our method offers much improved power over existing methods, even in challenging situations, while maintaining appropriate error control. We apply the approach to data from two published studies: the first aims to detect periodic genes of the organism Schizosaccharomyces pombe, and the second aims to identify genes associated with prostate cancer.

REFERENCES


Key words and phrases. Multiple testing, false discovery rates, Voronoi tessellations, empirical null distributions.


DETECTION BOUNDARY AND HIGHER CRITICISM APPROACH FOR RARE AND WEAK GENETIC EFFECTS

BY ZHEYANG WU*, YIMING SUN*, SHIQUAN HE*, JUDY CHO†, HONGYU ZHAO† AND JIASHUN JIN‡

Worcester Polytechnic Institute*, Yale University†
and Carnegie Mellon University‡

Genome-wide association studies (GWAS) have identified many genetic factors underlying complex human traits. However, these factors have explained only a small fraction of these traits’ genetic heritability. It is argued that many more genetic factors remain undiscovered. These genetic factors likely are weakly associated at the population level and sparsely distributed across the genome. In this paper, we adapt the recent innovations on Tukey’s Higher Criticism (Tukey [The Higher Criticism (1976) Princeton Univ.]; Donoho and Jin [Ann. Statist. 32 (2004) 962–994]) to SNP-set analysis of GWAS, and develop a new theoretical framework in large-scale inference to assess the joint significance of such rare and weak effects for a quantitative trait. In the core of our theory is the so-called detection boundary, a curve in the two-dimensional phase space that quantifies the rarity and strength of genetic effects. Above the detection boundary, the overall effects of genetic factors are strong enough for reliable detection. Below the detection boundary, the genetic factors are simply too rare and too weak for reliable detection. We show that the HC-type methods are optimal in that they reliably yield detection once the parameters of the genetic effects fall above the detection boundary and that many commonly used SNP-set methods are sub-optimal. The superior performance of the HC-type approach is demonstrated through simulations and the analysis of a GWAS data set of Crohn’s disease.

REFERENCES


Key words and phrases. Multiple hypotheses testing, large-scale inference, detection boundary, Higher Criticism, rare and weak effects, statistical power, genome-wide association studies, SNP-set methods.


Poverty maps are used to aid important political decisions such as allocation of development funds by governments and international organizations. Those decisions should be based on the most accurate poverty figures. However, often reliable poverty figures are not available at fine geographical levels or for particular risk population subgroups due to the sample size limitation of current national surveys. These surveys cannot cover adequately all the desired areas or population subgroups and, therefore, models relating the different areas are needed to “borrow strength” from area to area. In particular, the Spanish Survey on Income and Living Conditions (SILC) produces national poverty estimates but cannot provide poverty estimates by Spanish provinces due to the poor precision of direct estimates, which use only the province specific data. It also raises the ethical question of whether poverty is more severe for women than for men in a given province. We develop a hierarchical Bayes (HB) approach for poverty mapping in Spanish provinces by gender that overcomes the small province sample size problem of the SILC. The proposed approach has a wide scope of application because it can be used to estimate general nonlinear parameters. We use a Bayesian version of the nested error regression model in which Markov chain Monte Carlo procedures and the convergence monitoring therein are avoided. A simulation study reveals good frequentist properties of the HB approach. The resulting poverty maps indicate that poverty, both in frequency and intensity, is localized mostly in the southern and western provinces and it is more acute for women than for men in most of the provinces.

REFERENCES


Key words and phrases. Hierarchical Bayes, mixed linear model, nested error linear regression model, noninformative priors, poverty mapping, small area estimation.


ESTIMATION OF NONLINEAR DIFFERENTIAL EQUATION MODEL FOR GLUCOSE–INSULIN DYNAMICS IN TYPE I DIABETIC PATIENTS USING GENERALIZED SMOOTHING

BY INNA CHERVONEVA∗, BORIS FREYDIN∗, BRIAN HIPSZER∗, TATIYANA V. APANASOVICH† AND JEFFREY I. JOSEPH∗

Thomas Jefferson University* and George Washington University†

In this work we develop an ordinary differential equations (ODE) model of physiological regulation of glycemia in type 1 diabetes mellitus (T1DM) patients in response to meals and intravenous insulin infusion. Unlike for the majority of existing mathematical models of glucose–insulin dynamics, parameters in our model are estimable from a relatively small number of noisy observations of plasma glucose and insulin concentrations. For estimation, we adopt the generalized smoothing estimation of nonlinear dynamic systems of Ramsay et al. [J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (2007) 741–796]. In this framework, the ODE solution is approximated with a penalized spline, where the ODE model is incorporated in the penalty. We propose to optimize the generalized smoothing by using penalty weights that minimize the covariance penalties criterion (Efron [J. Amer. Statist. Assoc. 99 (2004) 619–642]). The covariance penalties criterion provides an estimate of the prediction error for nonlinear estimation rules resulting from nonlinear and/or nonhomogeneous ODE models, such as our model of glucose–insulin dynamics. We also propose to select the optimal number and location of knots for B-spline bases used to represent the ODE solution. The results of the small simulation study demonstrate advantages of optimized generalized smoothing in terms of smaller estimation errors for ODE parameters and smaller prediction errors for solutions of differential equations. Using the proposed approach to analyze the glucose and insulin concentration data in T1DM patients, we obtained good approximation of global glucose–insulin dynamics and physiologically meaningful parameter estimates.

REFERENCES


Key words and phrases. Generalized profiling, covariance penalties, parameter cascading, penalized smoothing, profiled penalty estimation, prediction error.


PAIRWISE COMPARISON OF TREATMENT LEVELS IN FUNCTIONAL ANALYSIS OF VARIANCE WITH APPLICATION TO ERYTHROCYTE HEMOLYSIS

BY OLGA VSEVOLOZHSKAYA*, MARK GREENWOOD* AND DMITRI HOLODOV†

Montana State University* and Voronezh State University†

Motivated by a practical need for the comparison of hemolysis curves at various treatment levels, we propose a novel method for pairwise comparison of mean functional responses. The hemolysis curves—the percent hemolysis as a function of time—of mice erythrocytes (red blood cells) by hydrochloric acid have been measured among different treatment levels. This data set fits well within the functional data analysis paradigm, in which a time series is considered as a realization of the underlying stochastic process or a smooth curve. Previous research has only provided methods for identifying some differences in mean curves at different times. We propose a two-level follow-up testing framework to allow comparisons of pairs of treatments within regions of time where some difference among curves is identified. The closure multiplicity adjustment method is used to control the family-wise error rate of the proposed procedure.

REFERENCES


Key words and phrases. Functional data analysis, FANOVA, multiple comparison, permutation method, pairwise comparison.


MONEYBARL: EXPLOITING PITCHER DECISION-MAKING USING REINFORCEMENT LEARNING

BY GAGAN SIDHU∗,† AND BRIAN CAFFO‡

General Analytics∗1, University of Alberta† and Johns Hopkins University‡

This manuscript uses machine learning techniques to exploit baseball pitchers’ decision making, so-called “Baseball IQ,” by modeling the at-bat information, pitch selection and counts, as a Markov Decision Process (MDP). Each state of the MDP models the pitcher’s current pitch selection in a Markovian fashion, conditional on the information immediately prior to making the current pitch. This includes the count prior to the previous pitch, his ensuing pitch selection, the batter’s ensuing action and the result of the pitch.

The necessary Markovian probabilities can be estimated by the relevant observed conditional proportions in MLB pitch-by-pitch game data. These probabilities could be pitcher-specific, using only the data from one pitcher, or general, using the data from a collection of pitchers.

Optimal batting strategies against these estimated conditional distributions of pitch selection can be ascertained by Value Iteration. Optimal batting strategies against a pitcher-specific conditional distribution can be contrasted to those calculated from the general conditional distributions associated with a collection of pitchers.

In this manuscript, a single season of MLB data is used to calculate the conditional distributions to find optimal pitcher-specific and general (against a collection of pitchers) batting strategies. These strategies are subsequently evaluated by conditional distributions calculated from a different season for the same pitchers. Thus, the batting strategies are conceptually tested via a collection of simulated games, a “mock season,” governed by distributions not used to create the strategies. (Simulation is not needed, as exact calculations are available.)

Instances where the pitcher-specific batting strategy outperforms the general batting strategy suggests that the pitcher is exploitable—knowledge of the conditional distributions of their pitch-making decision process in a different season yielded a strategy that worked better in a new season than a general batting strategy built on a population of pitchers. A permutation-based test of exploitability of the collection of pitchers is given and evaluated under two sets of assumptions.

To show the practical utility of the approach, we introduce a spatial component that classifies each pitcher’s pitch-types using a batter-parameterized spatial trajectory for each pitch. We found that heuristically labeled “nonelite” batters benefit from using the exploited pitchers’ pitcher-specific strategies, whereas (also heuristically labeled) “elite” players do not.

Key words and phrases. Markov, baseball, sports, simulation, algorithmic statistics.
REFERENCES


ADJUSTING MODELS OF ORDERED MULTINOMIAL OUTCOMES FOR NONIGNORABLE NONRESPONSE IN THE OCCUPATIONAL EMPLOYMENT STATISTICS SURVEY

BY NICHOLAS J. HORTON*,1, DANIEL TOTH† AND POLLY PHIPPS †

Amherst College* and Bureau of Labor Statistics†

An establishment’s average wage, computed from administrative wage data, has been found to be related to occupational wages. These occupational wages are a primary outcome variable for the Bureau of Labor Statistics Occupational Employment Statistics survey. Motivated by the fact that nonresponse in this survey is associated with average wage even after accounting for other establishment characteristics, we propose a method that uses the administrative data for imputing missing occupational wage values due to nonresponse. This imputation is complicated by the structure of the data. Since occupational wage data is collected in the form of counts of employees in predefined wage ranges for each occupation, weighting approaches to deal with nonresponse do not adequately adjust the estimates for certain domains of estimation. To preserve the current data structure, we propose a method to impute each missing establishment’s wage interval count data as an ordered multinomial random variable using a separate survival model for each occupation. Each model incorporates known auxiliary information for each establishment associated with the distribution of the occupational wage data, including geographic and industry characteristics. This flexible model allows the baseline hazard to vary by occupation while allowing predictors to adjust the probabilities of an employee’s salary falling within the specified ranges. An empirical study and simulation results suggest that the method imputes missing OES wages that are associated with the average wage of the establishment in a way that more closely resembles the observed association.

REFERENCES


Key words and phrases. Administrative data, auxiliary data, categorical outcome, establishment survey, missing data, imputation, survival analysis, regression trees.


LEVERAGING LOCAL IDENTITY-BY-DESCENT INCREASES THE POWER OF CASE/CONTROL GWAS WITH RELATED INDIVIDUALS

BY JOSHUA N. SAMPSON*,1, BILL WHEELER‡, PENG LI* AND JIANXIN SHI*,1

National Cancer Institute* and Information Management Services‡

Large case/control Genome-Wide Association Studies (GWAS) often include groups of related individuals with known relationships. When testing for associations at a given locus, current methods incorporate only the familial relationships between individuals. Here, we introduce the chromosome-based Quasi Likelihood Score (cQLS) statistic that incorporates local Identity-By-Descent (IBD) to increase the power to detect associations. In studies robust to population stratification, such as those with case/control sibling pairs, simulations show that the study power can be increased by over 50%. In our example, a GWAS examining late-onset Alzheimer’s disease, the p-values among the most strongly associated SNPs in the APOE gene tend to decrease, with the smallest p-value decreasing from $1.23 \times 10^{-8}$ to $7.70 \times 10^{-9}$. Furthermore, as a part of our simulations, we reevaluate our expectations about the use of families in GWAS. We show that, although adding only half as many unique chromosomes, genotyping affected siblings is more efficient than genotyping randomly ascertained cases. We also show that genotyping cases with a family history of disease will be less beneficial when searching for SNPs with smaller effect sizes.

REFERENCES


*Key words and phrases.* cQLS, GWAS, related individuals, case–control.


A BAYESIAN NONPARAMETRIC MIXTURE MODEL FOR SELECTING GENES AND GENE SUBNETWORKS

BY YIZE ZHAO¹, JIAN KANG¹,² AND TIANWEI YU³

Emory University

It is very challenging to select informative features from tens of thousands of measured features in high-throughput data analysis. Recently, several parametric/regression models have been developed utilizing the gene network information to select genes or pathways strongly associated with a clinical/biological outcome. Alternatively, in this paper, we propose a non-parametric Bayesian model for gene selection incorporating network information. In addition to identifying genes that have a strong association with a clinical outcome, our model can select genes with particular expressional behavior, in which case the regression models are not directly applicable. We show that our proposed model is equivalent to an infinity mixture model for which we develop a posterior computation algorithm based on Markov chain Monte Carlo (MCMC) methods. We also propose two fast computing algorithms that approximate the posterior simulation with good accuracy but relatively low computational cost. We illustrate our methods on simulation studies and the analysis of Spellman yeast cell cycle microarray data.

REFERENCES


STATISTICAL CALIBRATION OF QRT-PCR, MICROARRAY AND RNA-SEQ GENE EXPRESSION DATA WITH MEASUREMENT ERROR MODELS

BY ZHAONAN SUN, THOMAS KUCZEK AND YU ZHU

Purdue University

The accurate quantification of gene expression levels is crucial for transcriptome study. Microarray platforms are commonly used for simultaneously interrogating thousands of genes in the past decade, and recently RNA-Seq has emerged as a promising alternative. The gene expression measurements obtained by microarray and RNA-Seq are, however, subject to various measurement errors. A third platform called qRT-PCR is acknowledged to provide more accurate quantification of gene expression levels than microarray and RNA-Seq, but it has limited throughput capacity. In this article, we propose to use a system of functional measurement error models to model gene expression measurements and calibrate the microarray and RNA-Seq platforms with qRT-PCR. Based on the system, a two-step approach was developed to estimate the biases and error variance components of the three platforms and calculate calibrated estimates of gene expression levels. The estimated biases and variance components shed light on the relative strengths and weaknesses of the three platforms and the calibrated estimates provide a more accurate and consistent quantification of gene expression levels. Theoretical and simulation studies were conducted to establish the properties of those estimates. The system was applied to analyze two gene expression data sets from the Microarray Quality Control (MAQC) and Sequencing Quality Control (SEQC) projects.

REFERENCES


Key words and phrases. Transcriptome profiling, gene differential expression, comparative calibration, functional and structural parameters.
REGULARIZED 3D FUNCTIONAL REGRESSION FOR BRAIN IMAGE DATA VIA HAAR WAVELETS

BY XUEJING WANG, BIN NAN, JI ZHU, ROBERT KOEPPE AND FOR THE ALZHEIMER’S DISEASE NEUROIMAGING INITIATIVE

University of Michigan

The primary motivation and application in this article come from brain imaging studies on cognitive impairment in elderly subjects with brain disorders. We propose a regularized Haar wavelet-based approach for the analysis of three-dimensional brain image data in the framework of functional data analysis, which automatically takes into account the spatial information among neighboring voxels. We conduct extensive simulation studies to evaluate the prediction performance of the proposed approach and its ability to identify related regions to the outcome of interest, with the underlying assumption that only few relatively small subregions are truly predictive of the outcome of interest. We then apply the proposed approach to searching for brain subregions that are associated with cognition using PET images of patients with Alzheimer’s disease, patients with mild cognitive impairment and normal controls.

REFERENCES


Key words and phrases. Alzheimer’s disease, brain imaging, functional data analysis, Haar wavelet, Lasso, PET image, variable selection.


VOXEL-LEVEL MAPPING OF TRACER KINETICS IN PET STUDIES: A STATISTICAL APPROACH EMPHASIZING TISSUE LIFE TABLES

BY FINBARR O’SULLIVAN*, MARK MUZI‡, DAVID A. MANKOFF†, JANET F. EARY‡, ALEXANDER M. SPENCE‡,‡,‡ AND KENNETH A. KROHN‡

University College Cork*, University of Pennsylvania‡ and University of Washington‡

Most radiotracers used in dynamic positron emission tomography (PET) scanning act in a linear time-invariant fashion so that the measured time-course data are a convolution between the time course of the tracer in the arterial supply and the local tissue impulse response, known as the tissue residue function. In statistical terms the residue is a life table for the transit time of injected radiotracer atoms. The residue provides a description of the tracer kinetic information measurable by a dynamic PET scan. Decomposition of the residue function allows separation of rapid vascular kinetics from slower blood-tissue exchanges and tissue retention. For voxel-level analysis, we propose that residues be modeled by mixtures of nonparametrically derived basis residues obtained by segmentation of the full data volume. Spatial and temporal aspects of diagnostics associated with voxel-level model fitting are emphasized. Illustrative examples, some involving cancer imaging studies, are presented. Data from cerebral PET scanning with \(^{18}\text{F}\) fluoro-deoxyglucose (FDG) and \(^{15}\text{O}\) water (H2O) in normal subjects is used to evaluate the approach. Cross-validation is used to make regional comparisons between residues estimated using adaptive mixture models with more conventional compartmental modeling techniques. Simulations studies are used to theoretically examine mean square error performance and to explore the benefit of voxel-level analysis when the primary interest is a statistical summary of regional kinetics. The work highlights the contribution that multivariate analysis tools and life-table concepts can make in the recovery of local metabolic information from dynamic PET studies, particularly ones in which the assumptions of compartmental-like models, with residues that are sums of exponentials, might not be certain.

REFERENCES


Key words and phrases. Kinetic analysis, life-table, mixture modeling, PET.


ANALYSIS OF MULTIPLE SCLEROSIS LESIONS VIA SPATIALLY VARYING COEFFICIENTS

BY TIAN GE*,†, NICOLE MÜLLER-LENKE‡, KERSTIN BENDFELDT‡, THOMAS E. NICHOLS†,1 AND TIMOTHY D. JOHNSON§,1

Fudan University*, University of Warwick‡, University Hospital Basel‡ and University of Michigan§

Magnetic resonance imaging (MRI) plays a vital role in the scientific investigation and clinical management of multiple sclerosis. Analyses of binary multiple sclerosis lesion maps are typically "mass univariate" and conducted with standard linear models that are ill suited to the binary nature of the data and ignore the spatial dependence between nearby voxels (volume elements). Smoothing the lesion maps does not entirely eliminate the non-Gaussian nature of the data and requires an arbitrary choice of the smoothing parameter. Here we present a Bayesian spatial model to accurately model binary lesion maps and to determine if there is spatial dependence between lesion location and subject specific covariates such as MS subtype, age, gender, disease duration and disease severity measures. We apply our model to binary lesion maps derived from $T_2$-weighted MRI images from 250 multiple sclerosis patients classified into five clinical subtypes, and demonstrate unique modeling and predictive capabilities over existing methods.

REFERENCES


**Key words and phrases.** Image analysis, multiple sclerosis, magnetic resonance imaging, lesion probability map, Markov random fields, conditional autoregressive model, spatially varying coefficients.


A STATISTICAL APPROACH TO THE INVERSE PROBLEM IN MAGNETOENCEPHALOGRAPHY

BY ZHIGANG YAO†,* AND WILLIAM F. EDDY†

Ecole Polytechnique Fédérale de Lausanne* and Carnegie Mellon University†

Magnetoencephalography (MEG) is an imaging technique used to measure the magnetic field outside the human head produced by the electrical activity inside the brain. The MEG inverse problem, identifying the location of the electrical sources from the magnetic signal measurements, is ill-posed, that is, there are an infinite number of mathematically correct solutions. Common source localization methods assume the source does not vary with time and do not provide estimates of the variability of the fitted model. Here, we reformulate the MEG inverse problem by considering time-varying locations for the sources and their electrical moments and we model their time evolution using a state space model. Based on our predictive model, we investigate the inverse problem by finding the posterior source distribution given the multiple channels of observations at each time rather than fitting fixed source parameters. Our new model is more realistic than common models and allows us to estimate the variation of the strength, orientation and position. We propose two new Monte Carlo methods based on sequential importance sampling. Unlike the usual MCMC sampling scheme, our new methods work in this situation without needing to tune a high-dimensional transition kernel which has a very high cost. The dimensionality of the unknown parameters is extremely large and the size of the data is even larger. We use Parallel Virtual Machine (PVM) to speed up the computation.

REFERENCES


Key words and phrases. Ill-posed problem, sequential importance sampling, state space model, parallel computing, source localization.


BAYESIAN NONPARAMETRIC PLACKETT–LUCE MODELS FOR
THE ANALYSIS OF PREFERENCES FOR
COLLEGE DEGREE PROGRAMMES

BY FRANÇOIS CARON∗, YEE WHYE TEH∗ AND THOMAS BRENDAN MURPHY †

University of Oxford* and University College Dublin †

In this paper we propose a Bayesian nonparametric model for clustering partial ranking data. We start by developing a Bayesian nonparametric extension of the popular Plackett–Luce choice model that can handle an infinite number of choice items. Our framework is based on the theory of random atomic measures, with the prior specified by a completely random measure. We characterise the posterior distribution given data, and derive a simple and effective Gibbs sampler for posterior simulation. We then develop a Dirichlet process mixture extension of our model and apply it to investigate the clustering of preferences for college degree programmes amongst Irish secondary school graduates. The existence of clusters of applicants who have similar preferences for degree programmes is established and we determine that subject matter and geographical location of the third level institution characterise these clusters.

REFERENCES


Key words and phrases. Ranking data, permutations, gamma process, Dirichlet process, mixture models.


COMBINING ISOTONIC REGRESSION AND EM ALGORITHM TO PREDICT GENETIC RISK UNDER MONOTONICITY CONSTRAINT

BY JING QIN*, TANYA P. GARCIA†, YANYUAN MA‡, MING-XIN TANG§, KAREN MARDER2,§ AND YUANJIA WANG3,§

National Institute of Allergy and Infectious Diseases*, Texas A&M Health University Science Center†, Texas A&M University‡ and Columbia University§

In certain genetic studies, clinicians and genetic counselors are interested in estimating the cumulative risk of a disease for individuals with and without a rare deleterious mutation. Estimating the cumulative risk is difficult, however, when the estimates are based on family history data. Often, the genetic mutation status in many family members is unknown; instead, only estimated probabilities of a patient having a certain mutation status are available. Also, ages of disease-onset are subject to right censoring. Existing methods to estimate the cumulative risk using such family-based data only provide estimation at individual time points, and are not guaranteed to be monotonic or nonnegative. In this paper, we develop a novel method that combines Expectation–Maximization and isotonic regression to estimate the cumulative risk across the entire support. Our estimator is monotonic, satisfies self-consistent estimating equations and has high power in detecting differences between the cumulative risks of different populations. Application of our estimator to a Parkinson’s disease (PD) study provides the age-at-onset distribution of PD in PARK2 mutation carriers and noncarriers, and reveals a significant difference between the distribution in compound heterozygous carriers compared to noncarriers, but not between heterozygous carriers and noncarriers.

REFERENCES


Key words and phrases. Binomial likelihood, Parkinson’s disease, pool adjacent violation algorithm, self-consistency estimating equations.


A NEW METHOD OF PEAK DETECTION FOR ANALYSIS OF COMPREHENSIVE TWO-DIMENSIONAL GAS CHROMATOGRAPHY MASS SPECTROMETRY DATA

BY SEONGHO KIM*, MING OUYANG†, JAESIK JEONG‡, CHANGYU SHEN§ AND XIANG ZHANG¶

Wayne State University*, University of Massachusetts Boston†, Chonnam National University‡, Indiana University§ and University of Louisville¶

We develop a novel peak detection algorithm for the analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-TOF MS) data using normal–exponential–Bernoulli (NEB) and mixture probability models. The algorithm first performs baseline correction and denoising simultaneously using the NEB model, which also defines peak regions. Peaks are then picked using a mixture of probability distribution to deal with the co-eluting peaks. Peak merging is further carried out based on the mass spectral similarities among the peaks within the same peak group. The algorithm is evaluated using experimental data to study the effect of different cutoffs of the conditional Bayes factors and the effect of different mixture models including Poisson, truncated Gaussian, Gaussian, Gamma and exponentially modified Gaussian (EMG) distributions, and the optimal version is introduced using a trial-and-error approach. We then compare the new algorithm with two existing algorithms in terms of compound identification. Data analysis shows that the developed algorithm can detect the peaks with lower false discovery rates than the existing algorithms, and a less complicated peak picking model is a promising alternative to the more complicated and widely used EMG mixture models.

REFERENCES


Key words and phrases. Bayes factor, GC×GC-TOF MS, metabolomics, mixture model, normal–exponential–Bernoulli (NEB) model, peak detection.

BMC Bioinformatics 12 392.


BMC Bioinformatics 30 115.


Anal. Chem. 84 2622–2630.


Bioinformatics 24 1407–1413.


BMC Bioinformatics 10 4.
GENE-LEVEL PHARMACOGENETIC ANALYSIS ON SURVIVAL OUTCOMES USING GENE-TRAIT SIMILARITY REGRESSION

BY JUNG-YING TZENG*,‡,1,2, WENBIN LU*,1,3 AND FANG-CHI HSU†,4

North Carolina State University*, Wake Forest University‡ and National Cheng-Kung University†

Gene/pathway-based methods are drawing significant attention due to their usefulness in detecting rare and common variants that affect disease susceptibility. The biological mechanism of drug responses indicates that a gene-based analysis has even greater potential in pharmacogenetics. Motivated by a study from the Vitamin Intervention for Stroke Prevention (VISP) trial, we develop a gene-trait similarity regression for survival analysis to assess the effect of a gene or pathway on time-to-event outcomes. The similarity regression has a general framework that covers a range of survival models, such as the proportional hazards model and the proportional odds model. The inference procedure developed under the proportional hazards model is robust against model misspecification. We derive the equivalence between the similarity survival regression and a random effects model, which further unifies the current variance component-based methods. We demonstrate the effectiveness of the proposed method through simulation studies. In addition, we apply the method to the VISP trial data to identify the genes that exhibit an association with the risk of a recurrent stroke. The TCN2 gene was found to be associated with the recurrent stroke risk in the low-dose arm. This gene may impact recurrent stroke risk in response to cofactor therapy.

REFERENCES


Key words and phrases. Association study, gene/pathway, pharmacogenetics, similarity regression, survival data, proportional odds model, proportional hazards model.


PROBABILITY AGGREGATION IN TIME-SERIES: DYNAMIC HIERARCHICAL MODELING OF SPARSE EXPERT BELIEFS

BY VILLE A. SATOPÄÄ, SHANE T. JENSEN, BARBARA A. MELLERS, PHILIP E. TETLOCK AND LYLE H. UNGAR

University of Pennsylvania

Most subjective probability aggregation procedures use a single probability judgment from each expert, even though it is common for experts studying real problems to update their probability estimates over time. This paper advances into unexplored areas of probability aggregation by considering a dynamic context in which experts can update their beliefs at random intervals. The updates occur very infrequently, resulting in a sparse data set that cannot be modeled by standard time-series procedures. In response to the lack of appropriate methodology, this paper presents a hierarchical model that takes into account the expert’s level of self-reported expertise and produces aggregate probabilities that are sharp and well calibrated both in- and out-of-sample. The model is demonstrated on a real-world data set that includes over 2300 experts making multiple probability forecasts over two years on different subsets of 166 international political events.

REFERENCES


Key words and phrases. Probability aggregation, dynamic linear model, hierarchical modeling, expert forecast, subjective probability, bias estimation, calibration, time series.


Wallace, B. C. and Dahabreh, I. J. (2012). Class probability estimates are unreliable for imbalanced data (and how to fix them). In *Institute of Electrical and Electronics Engineers (IEEE) 12th International Conference on Data Mining (International Conference on Data Mining)* 695–704. IEEE Computer Society, Washington, DC.


