9^{TH} CONFERENCE OF THE EASTERN MEDITERRANEAN REGION AND THE ITALIAN REGION OF THE INTERNATIONAL BIOMETRIC SOCIETY

BOOK OF ABSTRACTS

9th EMR-IBS and Italian Region Conference

The meeting is devoted to the memory of Prof Marvin Zelen

http://stat-athens.aueb.gr/~emribs/page/emr2017.html

Thessaloniki, 8-12 May 2017

Preface

We are very pleased to welcome you at Thessaloniki for the 9th conference of the Eastern Mediterranean Region of the International Biometric Society (EMR-IBS) held jointly with the Italian Region in Thessaloniki, between 8-12 May 2017. The conference is dedicated to the memory of Marvin Zelen. A conference satellite Symposium honoring Professor Marvin Zelen takes place on 7-8 May 2017 in Thessaloniki, Greece. The two-day Symposium is organized by Frontier Science Foundation-Hellas and is co-sponsored by all Frontier offices.

A complete list of all the abstracts of the papers to be presented in the conference can be found in this book. A detailed index of all presenters can be found at the end to facilitate easy search. We hope that you will enjoy the 9th EMR-IBS Conference honoring the memory of Marvin Zelen.

Urania Dafni
Dimitris Karlis
on behalf of the LOC and SC.

Scientific Committee

Co-Chair

Urania Dafni	University of Athens, Greece
Dimitris Karlis	Athens University of Economics, Greece

Members

Ori Davidov	Univesity of Haifa, Israel
Costas Fokianos	University of Cyprus, Cyprus
Laurence Freedman	Bar Ilan University, Israel
Constantine Gatsonis	Brown University, USA
Lupe Gomez	Universitat Politecnica de Catalunya, Spain
Refik Burgut	Cukurova University, Adana, Turkey
Giota Touloumi	Medical School, University of Athens, Greece
Geert Molenberghs	University of Hasselt, Belgium
Ioannis Ntzoufras	Athens University of Economics, Greece
Sharon-Lise Normand	Harvard University, USA
Benjamin Reiser	University of Haifa, Israel
Christos Nakas	University of Thessaly, Greece
Maria-Grazia Valsecchi	University of Milano-Bicocca, Italy
Constantin Yiannoutsos	Indiana University, USA
David Zucker	University of Haifa, Israel
Argyris Ziogas	UCI Irvine, USA
Yoav Benjamini	Tel Aviv University, Israel
Nikos Demiris	Athens University of Economics, Greece
Clelia Di Serio	Universita San Rafaelle, Italy
Zeynep Kalaylioglu	Middel East Technical University, Turkey
KyungMann Kim	University of Wisconsin Madison, USA

Local Organizing Committee

Chair

Dimitris Karlis	Athens University of Economics, Greece

Members

Urania Dafni	University of Athens, Greece
Nikolaos Demiris	Athens University of Economics, Greece
Christos Nakas	University of Thessaly, Greece
Ioannis Ntzoufras	Athens University of Economics, Greece
Anna-Bettina Haidich	Aristotle University of Thessaloniki, Greece
Eleftherios Angelis	Aristotle University of Thessaloniki, Greece
Giota Touloumi	University of Athens, Greece
Paola Rancoita	Università Vita-Salute San Raffaele, Italy

Contents

Data, Statistics, and Inference
Sharon-Lise T. Normand 1

STRATOS and flexible modeling of time-dependent covariates in time-to-event analyses
Michal Abrahamowicz 3

A bi-dimensional finite mixture model for longitudinal data with dropout
Marco Alfò, Maria Francesca Marino, Alessandra Spagnoli 4

Inference in bidirectional multistate models for panel observed data: patterns of observation and flexible methods of estimation
Ahmadou Alioum 5

Modeling the impact of time to the intermediate event in the illness-death model
Elena Tassistro, Davide Paolo Bernasconi, Paola Rebora, Maria Grazia Valsecchi, Laura Antolini 6

Semiparametric regression on cumulative incidence function with interval-censored competing risks data
Giorgos Bakoyannis, Menggang Yu, Constantin T. Yiannoutsos 7

An inverse probability weighting approach to deal with informative censoring with application to childhood leukemia
Davide Paolo Bernasconi, Jessica Blanco Lopez, Emanuela Rossi, Laura Antolini, Maria Grazia Valsecchi 8

Analysis of time-to-event data with time-varying biomarkers measured only at study entry, with applications to Alzheimer's disease
Rebecca A. Betensky, Catherine Lee 9

Towards Computationally Efficient Epidemic Inference
Paul Birrell 10

Comparing multi-state models: some ideas based on dissimilarities
Marco Bonetti 11

Representation and prediction in multistate models
Bendix Carstensen 12

Time varying network models for brain imaging data
Ivor Cribben, Yi Yu 13

Extended Poisson-Tweedie models with some examples
Clarice G.B. Demétrio, John Hinde, Wagner H. Bonat 14

On some Bayesian spatio-temporal epidemic models
Nikos Demiris 15

Dealing with under-reported data through INAR-hidden Markov chains
A. Fernández-Fontelo, A. Cabaña, P. Puig, D. Moriña 16

Tensor-based methods for the analysis of hospital data
Paolo Giordani, Henk A.L. Kiers 17

Cost effective models for prediction and causal effect estimation: the goldmine of disease registers
Els Goetghebeur 18

Combining Item Response Theory with Multiple Imputation to Equate Health Assessment Questionnaires
Roee Gutman, Chenyang Gu 19

Discovery of structural brain imaging markers of HIV-associated outcomes using connectivity-informed regularization approach
Jaroslaw Harezlak, Marta Karas, Damian Brzyski, Mario Dzemidzic, Joaquin Goni, Beau Ances, Timothy W. Randolph 20

A general framework for selection bias due to missing data in EHR-based research
Sebastien Haneuse, Sarah Peskoe, David Arterburn, Michael Daniels 21

Why many researchers misuse variable selection and how to prevent this
Georg Heinze, Daniela Dunkler 22

Over/Under-dispersion, zero-inflation and Poisson-Tweedie models
John Hinde, Clarice Demétrio, Wagner Bonat 23

Causality, Prediction, and Everything in Between: The Promise and Pitfalls of Electronic Health Records Data
Joseph W. Hogan 24

A Bayesian approach to estimate changes in condom use from limited human immunodeficiency virus prevalence data
Kostas Kalogeropoulos 25

Multiple discrete distributions for modelling heaped count data in health insurance
Dimitris Karlis, Lluis Bermùdez, Miguel Santolino 26

A new longitudinal time-varying measurement error model with application to physical activity assessment instruments in a large biomarker validation study.
Victor Kipnis 27

A Bidirectional Multi-State Model for Panel Data on Bone Mineral Density among HIV-Infected Patients
Klaus Langohr, Guadalupe Gómez, Nuria Pérez, Eugenia Negredo, Anna Bonjoch 28

A Nonparametric Algorithm for Independent Component Analysis with Application to fMRI Data
Shanshan Li, Shaojie Chen, Chen Yue, Brian Caffo 29

Evaluating diagnostic accuracy of biomarkers in the presence of missing biomarkers
Shanshan Li 30

Approximate Bayesian Computation for large-scale gene duplication networks
Antonietta Mira, Jukka-Pekka Onnela 31

Time-sequential multiple imputation of missing data in time-dependent covariates used in a Cox model
Havi Murad, Alla Berlin, Rachel Dankner, Liraz Olmer, Paolo Boffetta, Laurence Freedman 32

Power-Expected-Posterior Priors for Generalized Linear Models
Dimitris Fouskakis, Ioannis Ntzoufras, Konstantinos Perrakis 33

Sieve Bootstrap for Functional Time Series
Efstathios Paparoditis 34

Bayesian block-diagonal variable selection and model averaging
Omiros Papaspiliopoulos 35

Machine learning analysis for assessing body composition and bone mass in HIV infected patients
Nuria Perez-Alvarez, Esteban Vegas, Carla Estany 36

Underdispersion in a bivariate framework
Pedro Puig, Jordi Valero, Amanda Fernandez-Fontelo 37

Conflict diagnostics in evidence synthesis: examples from infectious disease models
Anne Presanis 38

Functional principal component analysis and trajectory reconstruction with accelerated longitudinal designs
Philip T. Reiss 39

Bayesian variable selection: what if residuals are non-normal?
David Rossell 40

Functional data analysis of neuroimaging signals on the cerebral cortex
John A.D. Aston, Eardi Lila, Laura M. Sangalli 41

Functional Covariates in Additive Regression Models
Fabian Scheipl, Sonja Greven 42

Measurement Error in Nutritional Epidemiology: Impact, Current Practice for Analysis, and Opportunities for Improvement
Pamela A. Shaw Veronika Deffner, Kevin W. Dodd, Laurence S. Freedman, Ruth H. Keogh, Victor Kipnis, Helmut Kuechenhoff,, Janet A. Tooze for STRATOS TG4 Working Group 43

An improved estimation method for gray matter volume in presence of white matter hyperintesities in Alzheimer's and Down syndrome studies
Dana Tudorascu, Helmet Karim, Bill Klunk, Brad Christian, Ciprian Crainiceanu 44

Inverse probability of censoring weights under missing not at random with applications to long-term clinical outcomes
Constantin T. Yiannoutsos 45

Two-stage semiparametrics analysis of skeletal growth around pubertal growth spurt with interval-censored observations
Chenghao Chu, Ying Zhang, Wanzhu Tu 46

DNA: an economists perspective
Giovanni Barone Adesi 48

Modeling of Biochemical Networks via a New Graphical Approach
Melih Ağraz, Vilda Purutçuoğlu 49

Integrating Latent Classes in the Bayesian Shared Parameter Joint Model of Longitudinal and Survival Outcomes
Eleni-Rosalina Andrinopoulou, Kazem Nasserinejad, Dimitris Rizopoulos 51

On a class of design-based adaptive and sequential sampling strategies
Federico Andreis, Marco Bonetti 52

Summary indicators to assess the performance of risk predictors
Laura Antolini, Elena Tassistro, Davide Paolo Bernasconi, Maria Grazia Valsecchi 53

Joint Models for Longitudinal and Time-to-Event Data in a Case-Cohort Desig
Sara Baart, Dimitris Rizopoulos, Eric Boersma 54

Non-proportional hazards or unobserved heterogeneity in clustered survival data: Can we tell the difference?
Theodor Adrian Balan, Hein Putter 55

Estimation of Smooth ROC Curves for Biomarkers With Limits of Detection
Leonidas E. Bantis Qingxiang Yan John V. Tsimikas, Ziding Feng 56

Estimating metabolite networks subject to dietary preferences in the longitudinal setting
Georgios Bartzis,, Hae-Won Uh, Fred van Eeuwijk, Jeanine J. Houwing-Duistermaat 57

A Comprehensive Simulation Study for Comparison of Statistical Methods in MRMC ROC Studies
Merve Basol, Dincer Goksuluk, A. Ergun Karaagaoglu 58

On the use of routine health insurance data - exemplified by the Austrians disease management program for diabetes mellitus
Andrea Berghold, Regina Riedl 59

The anticipated odds ratios to decide the choice of a primary binary endpoint
Marta Bofill Roig, Guadalupe Gómez Melis 60

Branching processes in continuous time as models of mutations
Maroussia Slavtchova-Bojkova 61

A Bayesian detection model for chronic disease surveillance: application to COPD hospitalisation data
Areti Boulieri, Marta Blangiardo 62

Overviews of Systematic Reviews with drugs, herbal medicine or dietary supplements
Konstantinos Bougioukas, Eirini Pagkalidou, Eleni Avgerinou, Apostolos Tsapas, Evangelia Ntzani, Emmanouil Smyrnakis, Anna-Bettina Haidich 63

MTHFR C677T and multiple health outcomes: An umbrella review.
Emmanouil Bouras, Christos Kotanidis, Anthoula Chatzikyriakidou, Sofia Kouidou, Anna-Bettina Haidich 64

Between-Sample Heterogeneity — Is AIC Really Optimal?
Mark J Brewer, Adam Butler 65

Bayesian Longitudinal Circular Data Modeling
Onur Camli, Zeynep Kalaylioglu 66

Comparative Analysis of Calibration Methods for Microsimulation Models: An application on the MILC model
Stavroula A. Chrysanthopoulou 67

Identifying clusters of physiological response elicited by an emotion recognition task
Federica Cugnata, Manuela Ferrario, Marco Airoldi, Marcello Arcangeli, Riccardo M. Martoni, Clelia Di Serio, Chiara Brombin 69

A two-stage approach for estimating the parameters of an age-group epidemic model from incidence data
Itai Dattner 70

The use of urn models in response-adaptive randomized designs: a simulation study
Valeria Edefonti, Andrea Ghiglietti, Maria Giovanna Scarale, Rosalba Miceli . 71

Combinatorial Mixtures of Multiparameter Distributions: an Application to Prostate Cancer
Valeria Edefonti, Giovanni Parmigiani 72

On detecting the change-points in piecewise regression models
Dimitra Eleftheriou 73

Predicting Genetic Predisposition to Treatment Responsiveness in Randomized Clinical Trials Using Bayesian Whole Genome Regression
Bahar Erar, George D. Papandonatos 74

Bayesian imputation of time-varying covariates in linear mixed models
Nicole S. Erler, Dimitris Rizopoulos, Vincent W.V. Jaddoe, Oscar H. Franco, Emmanuel M.E.H. Lesaffre 75

Simulation of multi-pollutant model results in the presence of measurement error
Dimitris Evangelopoulos, Ruth Keogh, Klea Katsouyanni, Heather Walton 76

Extent, Duration and Predictors of Exclusive Breastfeeding in a Longitudinal Study: Adjusting for missing data using an Accelerated Failure Time model
Samah Hayek, Havi Murad, Anneke Ifrah, Tamy Shohat, Laurence Freedmant . 77

Generalized Chao estimators with external information and measurement error
Alessio Farcomeni, Francesco Dotto 78

Bayesian Permutation Tests
Livio Finos, Florian Klinglmueller 79

From GWAS to personalized disease risk prediction: statistical aspects of genetic risk score development and validation
Krista Fischer, Kristi Läll, Sulev Reisberg 80

Leave-one-out crossvalidation favors inaccurate estimators
Angelika Geroldinger, Lara Lusa, Mariana Nold, Georg Heinze 81

Conditional frailty Marshall-Olkin survival model for bivariate censored failure time data
Giussani Andrea, Bonetti Marco 82

EMcorrProbit R package
Denitsa Grigorova, Nina Daskalova 83

Measuring Inequality from Incomplete Income and Survival Data
Long Hong, Guido Alfani, Chiara Gigliarano, Marco Bonetti 84

A Score Test for Over-Dispersion in Marginalized Zero-Inflated Poisson Regression Models
Gul Inan, John Preisser, Kalyan Das 85

A Quantile Regression Model for Failure Time Data with Time Dependent Covariates
Malka Gorfine, Yair Goldberg, Ya'acov Ritov 86

Adapting censored regression methods to adjust for the limit of detection in the calibration of diagnostic rules for clinical mass spectrometry proteomic data
Alexia Kakourou, Werner Vach, Bart Mertens 87

Targeting Disease Signatures Towards Precision Healthcare
Mira Markus Kalish, Alexis Mitelpunkt, Tal Kozlovski,, Tal Galili, Anat Mirelman, Yoav Benjamini 88

Adjusting for unit nonresponse in EMENO study: inverse probability weighting and multiple imputation methods
Natasa Kalpourtzi, Giota Touloumi 89

Multiple Factor Analysis with Its Basic Properties and An Application
Siddik Keskin 90

The Use of Joint Modeling Approach in Personalized Medicine
Naime Meric Konar, Eda Karaismailoglu, Ergun Karaagaoglu 91

A Modeling Approach for Predicting Disease Status Using Functional Data in the Absence of a Gold Standard
Amita Manatunga, Qi Long, Andrew T. Taylor 92

Cox Regression for Doubly Truncated Data
Micha Mandel, Jacobo de Uña-Álvarez, David K. Simon, Rebecca A. Betensky . 93

Optimal Sampling Designs of Two-Compartment Nonlinear Regression Models
Noa Molshatski, Sandrah P. Eckel 94

Refining prognostication in CLL
Theodoros Moysiadis, Panagiotis Baliakas, Davide Rossi, Achilles Anagnostopoulos, Jonathan C. Strefford, Sarka Pospisilova, David Oscier,

Gianluca Gaidano, Elias Campo, Paolo Ghia, Richard Rosenquist, Kostas Stamatopoulos 95

Nonparametric and parametric confidence intervals for the Youden index and its associated cutoff point
Christos T Nakas, Leonidas E Bantis, Benjamin Reiser 96

The number of strata in propensity score subclassification
Markus Neuhäuser 97

Dynamic borrowing through empirical power priors that control type I error
Stavros Nikolakopoulos, Kit Roes 98

The latent scale covariogram: a tool for exploring the spatial dependence structure of non-normal responses
Samuel D. Oman, Jorge Mateu 99

Fuzzy C Means Algorithm Applications on MiRNA Gene Expression data and evaluation of miRNA pathways
Suriye Ozgur, Bakiye Goker Bagca, Muhterem Duyu, Ozgur Cogulu, Mehmet N. Orman 100

Classifying Resting State Functional Magnetic Resonance Imaging Data
Ebru Ozturk, Ozlem Ilk 101

Three different approaches to diagnose of rheumatic diseases by using quality of life scores
Ozge Pasin, Handan Ankarali, Safinaz Ataoglu 102

Using Pseudo Individual Patient Data in Random-effects Meta-analysis
Katerina Papadimitropoulou, Saskia le Cessie, Olaf Dekkers, Theo Stijnen 103

Individualized Dynamic Prediction of Survival under Time-Varying Treatment Strategies
Papageorgiou G., Rizopoulos D., Mokhles M. M., Takkenberg J. J. M. 104

Attenuated spline reconstruction technique for SPECT/CT
Nicholas E. Protonotarios, Athanassios S. Fokas, George A. Kastis 105

A new measure for prognostic index evaluation
Paola M.V. Rancoita 106

Sequential Cox and marginal structural models to evaluate the role of anticoagulant therapy on mortality in haemodialysis patients
Paola Rebora, Emanuela Rossi, Simonetta Genovesi, Maria Grazia Valsecchi . . 107

Choosing the number of classes in Bayesian finite mixture models using the posterior distribution of the mixing proportions
Kazem Nasserinejad, Joost van Rosmalen, Wim de Kort, Emmanuel Lesaffre . . 108

A new biosurveillance method based on convex hulls
Athanasios Sachlas, Polychronis Economou, Sotiris Bersimis 109

Deterministic Modeling and Inference of Biological Networks
Deniz Seçilmiş, Vilda Purutçuoğlu 110

Comparing Logistic Regression and Decision Tree Analysis Results: A Simulation Study With An Application To Real Data
Yasar Sertdemir, Hacer Y. Yildizdas, Ferda G. Ozlu, Ilker Unal, Adnan Barutcu, Mehmet Satar, Mustafa Akcali 111

Development of high dimensional microbiome biomarkers
Ziv Shkedy, Nolen Joy Perualila, Rudradev Sengupta 113

Bayesian optimal cluster designs
Satya Prakash Singh, Siuli Mukhopadhyay 114

Assessing Variable Selection Uncertainty in Linear Models
Aldo Solari, Ningning Xu, Jelle J. Goeman 115

Estimating the intervention effect and measurement error in calibration sub-studies
Michal Talitman, Malka Gorfine, David M. Steinberg 116

Misspecifying the covariance structure in a linear mixed model under MAR dropout
Christos Thomadakis, Loukia Meligkotsidou, Nikos Pantazis, Giota Touloumi . . 117

Numerical approach to modelling mutation times in continuous time multi-type branching processes
Plamen Trayanov, Maroussia Slavtchova-Bojkova, Stoyan Dimitrov 118

Methodological challenges in the analysis of longitudinal RNAseq data
Roula Tsonaka 119

Comparison of Estimation Methods for Area under the ROC Curve with Skewed Data
Ilker Unal, Yasar Sertdemir 120

A weighted composite likelihood estimators through an information criterion
Kostas Florios, Irini Moustaki, Dimitris Rizopoulos, Vassilis Vasdekis 121

Simple bias formulas for mediation analysis with unmeasured confounding
Kai Wang 122

Optimal dose finding for confirmatory studies
Panagiota Zygoura, Thomas Jaki 123

A Non-parametric Bayes Approach for Instrumental Variable Analysis
Samrachana Adhikari, Sharon-Lise Normand 125

A New Hos Index for Academic Performance
Handan Ankarali, Ozge Pasin, Seyit Ankarali 126

Reliability and Validity of Short Forms of Some Scales for Used in Quality of Life in Rheumatic Diseases
Safinaz Ataoglu, Handan Ankarali, Ozge Pasin 127

Developing an R-script for semi-automated statistical analysis
Avramopoulos A, Apostolidou Kiouti F, Goulis DG, Symeonidis A, Haidich AB 128

MTHFR C677T and multiple health outcomes: An umbrella review.
Emmanouil Bouras, Christos Kotanidis, Anthoula Chatzikyriakidou, Sofia Kouidou, Anna-Bettina Haidich 129

Big Data and Usage in Data Mining Applications on Medicine
BAKIRARAR Batuhan, KAR Irem, ELHAN Atilla Halil, KOSE Serdal Kenan 130

Random Forest application on DNA methylation microarray data in CLL
Nestoras Chalkidis, Maria Tsagiopoulou, Nikos Papakonstantinou, Theodoros Moysiadis 131

Algorithmic Construction of Optimal Block Designs for Two-Colour cDNA Microarray Experiments Using the Linear Mixed Effects Model
Legesse Kassa Debusho, Dibaba Bayisa Gemechu, Linda M. Haines 132

Computer-intensive methods for Model Selection in Animal Breeding
Nikos Demiris 133

How to Overcome Class Imbalance Problem
Duygu Aydin Hakli, Dincer Goksuluk and Erdem Karabulut 134

Estimation of blood pressure percentiles in healthy children using polynomial regression
Pembe Keskinoglu, Suriye Ozgur, Timur Kose, Ahmet Keskinoglu 135

Evaluation of the performance of some data mining methods for hardly discriminated clinical diagnosis
Ahmet Keskinoglu, Suriye Ozgur, Caner Alparslan, Onder Yavascan, Burak Ordin, Pembe Keskinoglu 136

A meta-analysis and meta-regression approach to Obsessive-Compulsive Disorder
Laura Marciano, Clelia Di Serio 138

Perceived barriers of colorectal cancer screening in Northern Greece; a population survey of healthy adults.
Mastrokostas A, Gavana M, Touloumi G, Benos A, Haidich AB 139

Statistical evaluation of HPV-E7 oncoprotein detection as a triage method to colposcopy
Theodoros Moysiadis, Theodoros Agorastos, Kimon Chatzistamatiou, Andreas M. Kaufmann, Alkmini Skenderi, Irini Lekka, Isabel Koch, Erwin Soutschek, Oliver Boecher, Vasilis Kilintzis, Stamatia Angelidou, Evangelia Katsiki, Ingke Hagemann, Eleonora Boschetti-Gruetzmacher, Athena Tsertanidou, Eleftherios Angelis, Nikolaos Maglaveras, Pidder Jansen-Duerr 141

Assessment of diagnostic test results detected in independent samples, for correspondence with the Suzuki and standards methods
Mustafa S. Senocak Hayriye Ertem Vehid Nurgul Bulut, Gokalp Eral 142

A review of Bayesian methods in meta-analysis of diagnostic accuracy test studies
Eirini Pagkalidou, Eleni Verykouki, Anna-Bettina Haidich 143

Analyzing Data from BCI Control of Robotic Arms
N. Pandria, A. Athanasiou, G. Arfaras, P.D. Bamidis, ... 144

Accounting for heterogeneity in a sparse-event meta-analysis of a few small trials
Konstantinos Pateras, Stavros Nikolakopoulos, Kit Roes 145

Evaluation of Different Statistical Methods Used to Test for Agreement between the Results of Continuous Measurements
Ozge Pasin, Ahmet Dirican, Basak Gurtekin, Sevda Ozel Yildiz, Rian Disci . . . 146

Methods for estimating the cumulative effect of correlated exposures: an application on the health effects of air pollution in Athens, Greece
Sophia Rodopoulou, Klea Katsouyanni, Pagona Lagiou, Evangelia Samoli 147

Non-orthogonal multi-stratum design of experiments
Luzia A. Trinca 149

Differences in correlation structure of gene expression data, used for variable selection in Support Vector Machines algorithm
Athina Tsanousa, Lefteris Angelis 150

Joint Modeling of Longitudinal Measurements and Survival Time Data- a Simulation Study of Endogenous Time Dependent Covariates in JM versus in Extended Cox Model
Ebru Turgal, Beyza Doğanay Erdoğan 151

Discovery Of Ensemble Methods On A Real Delirium Data From Hospital Database
Merve Gulsah Ulusoy, Cigdem Dinckal Pinar Tosun Tasar, Suriye Ozgur, Nur Ozge Akcam, Ozan Fatih Sarikaya, Aysin Noyan, Soner Duman 152

Impact of outliers in Meta-Analysis
Mutlu Umaroglu, Pinar Ozdemir 153

Index 154

Marvin Zelen Keynote Lecture

Data, Statistics, and Inference

Sharon-Lise T. Normand[1]

Department of Health Care Policy, Harvard Medical School and Department of Biostatistics, Harvard T.H. Chan School of Public Health

Abstract. Marvin Zelen pioneered and advocated the use of quantitatively rigorous methodology in statistical science. With increased access to electronic health information, ambitious attempts to understand the effect caused by new medical interventions in usual care populations have intensified. Moreover, global connectedness of health information has informed country-specific public health policy decisions. By conditioning on rich confounding information, utilizing larger populations, and multiple sources of information, researchers aim to comply with key principles underpinning causal inference. During this talk, an examination of current (and future) substantive and methodological problems will be discussed. Funded by R01- GM111339 and U01-FDA004493.

Invited papers

STRATOS and flexible modeling of time-dependent covariates in time-to-event analyses

Michal Abrahamowicz[1]

Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada

Abstract. Prospective or retrospective observational studies of human health increasingly explore associations between repeated-over-time measurements of time-dependent covariates and hazard of a clinical endpoint of interest. One analytical challenge specific to modeling the effect of a time-dependent covariate concerns the need to specify a time-varying 'exposure metric' that defines how its past and current values jointly affect the current hazard [1]. Most recent epidemiological and clinical studies rely on simple *a*d hoc exposure metrics such as current value of $X(t)$ or any exposure ($X(t) \neq 0$) in the past year. Yet, an arbitrary choice of the exposure metric may largely reduce the power for detecting an association, lead to biased estimates and incorrect conclusions [1]. To simultaneously account for the intensity, duration and timing of past exposures, we proposed a more general, flexible Weighed Cumulative Exposure (WCE) model for time-to-event analyses [2]: $WCE(\tau|x(t), t < \tau) = \sum w(\tau - t)[x(t)]$ where τ is the current time when the hazard is evaluated;x(t) represents value of the time-dependent covariate (e. daily dose of a drug dose) observed at time t ($t < \tau$) in the past; and the function $w(\tau - t)$ assigns relative importance weights to past doses, depending on the time elapsed since the dose was taken ($\tau - t$). Thus, the WCE metric is defined as the weighted sum of past doses, with weights determined by $w(\tau - t)$. The weight function $w(\tau - t)$ is modelled using un-penalized cubic regression B-splines, avoiding the need to specify its analytical form. The estimated $WCE(\tau)$ is then included as a time-varying covariate in the Cox's PH model [2].The R program, that implements the WCE model in Cox regression analyses is available on the free-access CRAN website [3]. Recently, we have extended the WCE modeling to Marginal Structural Cox model (MSM Cox) with inverse probability of treatment (IPT) weights [4].The accuracy of the WCE estimates will be evaluated in simulations. To illustrate pharmaco-epidemiological applications, WCE model will be used to re-assess the associations of (a) oral glucocorticoids and infections, (b) antiretroviral treatmentand cardiovascular risks in HIV(MSM analysis).

References

Abrahamowicz M, Beauchamp M-E, Sylvestre M-P. Comparison of alternative models for linking drug exposure with adverse effects. *Stat Med. 2012;31:1014–1030.*.

Sylvestre M.P. & Abrahamowicz M.Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Stat Med. 2009;28:3437–3453.*.

Xiao Y., Abrahamowicz M., Moodie EM., et al. Flexible marginal structural models for estimating the cumulative effect *JASA.2014;109:455–464.*

A bi-dimensional finite mixture model for longitudinal data with dropout

Marco Alfò[1], Maria Francesca Marino[2], and Alessandra Spagnoli[1]

[1] Sapienza Università di Roma, Rome, Italy marco.alfo@uniroma1.it, alessandra.spagnoli@uniroma1.it
[2] Università di Perugia, Perugia, Italy mariafrancesca.marino@unipg.it

Abstract. In longitudinal studies, subjects may be lost to follow up, and therefore individual-specific response sequences may be incomplete. When the probability of nonresponse, conditional on available covariates and responses, still depends on unobservables, the dropout mechanism can not be *ignoed*. Insuch a case, he aim is at defining a potentially reliable association structure that may account for dependence between the longitudinal and the dropout processes. We discuss a random coefficient based dropout model where latent effects follow an unknown discrete distribution which describes heterogeneity in the univariate profiles, with possibly different numbers of locations in each margin. Dependence between profiles is introduced by using a bi-dimensional representation for the corresponding distribution, with a full association structure connecting each location in a margin to each location in the other one. Unlike standard (unidimensional) finite mixture models, the non ignorable dropout model properly nests its ignorable counterpart. We detail the proposed modelling approach by analysing data from a longitudinal study on the dynamics of cognitive functioning in the elderly, and propose measures for parameters sensitivity.

Keywords

FINITE MIXTURES, RANDOM COEFFICIENT DROP-OUT MODELS, BI-DIMENSIONAL ASSOCIATION

References

ALFÓ, M. and ROCCHETTI, I (2013) A flexible approach to finite mixture regression models for multivariate mixed responses, *Stat. & Prob. Lett.*, 83:1754–1758.

CREEMERS, A., HENS, N., AERTS, M., MOLENBERGHS, G., VERBEKE, G. and KENWARD, M. (2010) A sensitivity analysis for shared-parameter models for incomplete longitudinal data, *Biom. Journ.*, 52:111–125.

DUNSON, D. and XING, C.J. (2009) Nonparametric bayes modeling of multivariate categorical data, *Journ. Amer. Stat. Ass.*, 104:1042–1051.

Inference in bidirectional multistate models for panel observed data: patterns of observation and flexible methods of estimation

Ahmadou Alioum[1]

Inserm, Bordeaux Population Health Research Center, UMR 1219, Univ. Bordeaux, ISPED, F-33000 Bordeaux, France. ahmadou.alioum@u-bordeaux.fr

Abstract. Multistate models are very useful for modeling the occurrence of several events over time in longitudinal epidemiological studies. The definition of states (often based on plausible clinical or biological conditions) and possible transitions between states depends on the problem under consideration and can lead to very complex models with many states and backward transitions. If, in addition, the observations are made in discrete times so that the exact transition times are unknown, inference for such models becomes very complex. This explains why the time homogeneous Markov process is very often assumed for inference in such models in the presence panel data. Even if extensions based on models with piecewise constant transition intensities or time transformation models have been proposed and used, there is a need for more flexible estimation methods for fitting nonhomogeneous Markov models. These methods rely on the use of penalized likelihood or B-spline functions and are however more computationally time-consuming. The objective of this talk is to review flexible methods of estimation and discuss the impact of observation schemes on inference in bidirectional multistate models for panel data.

Modeling the impact of time to the intermediate event in the illness-death model

Elena Tassistro[1], Davide Paolo Bernasconi[1], Paola Rebora[1], Maria Grazia Valsecchi[1] and Laura Antolini[1]

School of Medicine and Surgery, University of Milano-Bicocca, Via Cadore 48, 20900 Monza, Italy laura.antolini@unimib.it

Abstract. The illness-death model is the simplest multistate model where the transition from the initial state to the final state involves an intermediate state (illness). The analysis of the impact of the transition to illness and that of the time to transition on the hazard of failure has a key role in gaining insights into the dynamic of disease.

The standard approach is the joint model of both hazards including illness as time-varying covariate and measuring time on the original scale (from initial state). The hazard of failure on the subsample of ill patients can be modelled including time to illness as a covariate, measuring the time on the clock reset scale (from illness). A limitation of this approach is that time from start is accounted only through the time to illness, and not as a time scale. A recently proposed approach addressed these issues by a Poisson regression model that include both time scales and is applied to all patients from initial state. A further possibility we propose for consideration is a hazard based model where time is measured in the original scale before the transition to illness and on the clock reset scale after transition.

In this presentation we show through a simulation protocol that the clock reset approach is the most appropriate to deal with semi-Markov and extended semi-Markov scenarios.

Keywords

Illness-death model; Poisson model; time scales; transition hazard

References

Iacobelli S, Carstensen B. (2013). Multiple time scales in multi-state models. *Statistics in Medicine* 32:5315-5327

Eulenburg C, Mahner S, Woelber L, Wegscheider K. (2015). A systematic model specification procedure for an illness-death model without recovery. *PLoS ONE* 10(4):e0123489

Semiparametric regression on cumulative incidence function with interval-censored competing risks data

Giorgos Bakoyannis[1], Menggang Yu[1], and Constantin T. Yiannoutsos[1]

Indiana University R.M. Fairbanks School of Public Health, Indianapolis, IN, USA
gbakogia@iu.edu

Abstract. Many biomedical and clinical studies with time-to-event outcomes involve competing risks data subject to interval censoring. Interval censoring is the situation where the failure time is not precisely observed, but is only known to lie between two observation times such as clinical visits in a cohort study. Not taking into account the interval censoring may result in biased estimation of the cause-specific cumulative incidence function, an important quantity in the competing risk framework, used for studying the prognosis of various diseases, for evaluating interventions in populations, and for prediction and implementation science purposes. In this work we consider the class of semiparametric generalized odds-rate transformation models in the context of sieve maximum likelihood estimation based on B-splines. This large class of models includes both the proportional odds and the proportional subdistribution hazard models (i.e., the Fine-Gray model) as special cases. The estimator for the regression parameter is shown to be consistent, semiparametrically efficient and asymptotically normal. Simulation studies suggest that the method performs well even with small sample sizes. As an illustration we use the proposed method to analyze data from HIV-infected individuals obtained from a large cohort study in sub-Saharan Africa. The proposed methods can easily be performed using the R function ciregic that can be provided by the authors.

An inverse probability weighting approach to deal with informative censoring with application to childhood leukemia

Davide Paolo Bernasconi[1,2], Jessica Blanco Lopez[1], Emanuela Rossi[1], Laura Antolini[1], and Maria Grazia Valsecchi[1]

[1] School of Medicine and Surgery, University of Milano-Bicocca, Via Cadore 48, 20900 Monza, Italy
[2] davide.bernasconi@unimib.it

Abstract. An important cause of poor outcomes in children of low-income countries affected by Acute Lymphoblastic Leukemia (ALL) is abandonment of treatment, a factor associated with both biologic and socioeconomic factors. Estimation of the relapse-free survival using traditional methods (e.g. Kaplan-Meier curves, Cox model) is based on the assumption that abandon causes non informative censoring, which is not appropriate.

Marginal structural models based on the inverse probability of treatment and censoring (IPTC) weighting, under specific assumptions, allow the estimation of potential outcomes as if no informative censoring occurred and as if the whole cohort was exposed, or not, to a factor (Robins et al. 2000). We compared the outcomes of children enrolled in two subsequent protocols for ALL treatment (2000-2007 and 2008-2015) in Central America. We adopted an IPTC weight-adjusted Kaplan-Meier method to assess the potential relapse-free survival under no abandonment and given all patients exposed to both protocols. The evaluation of the time-dependent censoring (abandon) weights was carried out using the Aalen additive model, while logistic regression was adopted for the treatment (protocol) weights. Normalization and truncation of the weights was also considered. Pointwise 95% confidence intervals for the relapse-free survival were computed using bootstrap.

Keywords

Pediatric Leukemia; Treatment Abandon; Marginal Structural Models; IPW

References

Robins, J.; Hernan, M.; Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 11(5):550-60.

Analysis of time-to-event data with time-varying biomarkers measured only at study entry, with applications to Alzheimer's disease

Rebecca A. Betensky[1] and Catherine Lee[1]

Harvard T.H. Chan School of Public Health
betensky@hsph.harvard.edu
catherinelee@fas.harvard.edu

Abstract. Relating time-varying biomarkers of Alzheimer's disease (AD) to time-to-event using a Cox model is complicated by the fact that AD biomarkers are sparsely collected, typically only at study entry; this is problematic since Cox regression with time-varying covariates requires observation of the covariate process at all failure times. The analysis might be simplified by treating the time-varying covariate collected only at study entry as a fixed baseline covariate, however in longitudinal AD studies, the time at study entry tends to be arbitrary and the validity of this approach is questionable. In this paper, we investigate the validity of using study entry as the time origin in a Cox model with time-varying covariates and of treating a time-varying covariate as fixed at its value at study entry, assuming the true origin precedes study entry. We first derive conditions under which using an incorrect time origin of study entry results in consistent estimation of regression parameters when the time-varying covariate is continuous and fully observed. We then derive conditions under which treating the time-varying covariate as fixed results in consistent estimation. Then, assuming the biomarker follows a specified functional form, we provide methods for estimating the regression parameter in the setting of delayed entry where the time-varying predictor is measured only at study entry and an appropriate time origin precedes study entry. Our analytical results and methods are supported through a simulation study. Finally, we apply our methods to data from the Rush Religious Orders Study and Memory and Aging Project.

Keywords

survival analysis; Cox model; time-dependent covariates; choice of time origin; sparsely collected covariates

Towards Computationally Efficient Epidemic Inference

Paul Birrell[1]

MRC Biostatistics Unit, University of Cambridge paul.birrell@mrc-bsu.cam.ac.uk

Abstract. In a pandemic where infection is widespread, there is no direct observation of the infection processes. Instead information comes from a variety of surveillance data schemes that are prone to noise, contamination, bias and sparse sampling. To form an accurate impression of the epidemic and to be able to make forecasts of its evolution, therefore, as many of these data streams as possible need to be assimilated into a single integrated analysis. The result of this is that the transmission model describing the infection process and the linked observation models can become computationally demanding, limiting the capacity for statistical inference in real-time.

I will discuss some of our attempts at making the inferential process more efficient, with particular focus on dynamic emulation, where the computationally expensive epidemic model is replaced by a more readily evaluated proxy, a time-evolving Gaussian process trained on a (relatively) small number of model runs at key input values, training that can be done a priori. It appears, however, that algebraically convenient methods to subsequently calibrate the model can damage the inference and some compromises need to be made.

Comparing multi-state models: some ideas based on dissimilarities

Marco Bonetti[1]

Bocconi University

Abstract. Multi-state models are used to describe the occurrence over time of events of different kinds, with the inclusion of explanatory variables and with particular emphasis on microsimulation-based prediction. Such situation occurs commonly in a variety of setting, both in biomedical studies and increasingly in the social sciences.

We discuss some distance-based criteria that can be used to compare the ability of two or more competing models to fit and predict sequence data. We suggest some possibilities in this direction, and apply them to data collected as part of the Fertility and Family Surveys study.

Representation and prediction in multistate models

Bendix Carstensen[1]

Steno Diabetes Center Copenhagen, DK-2820 Gentofte, Denmark
`bcar0029@regionH.dk, b@bxc.dk;  http://bendixcarstensen.com`

Abstract. Life history studies will typically require that you set up a multistate model and define models for all transitions. These will be the basis for calculation of quantities such as the life-time risk of a particular event or the expected sojourn time in a given state.

If the underlying data collection is from a panel study or from clinical records where state is recorded at different times, we do not have the exact transition (event) times between states. This can be regarded as a missing data problem, that can be fixed by multiple imputation.

I shall describe the philosophy behind the Lexis machinery for representation of multistate data on multiple time scales [1] as implemented in the `Epi` package for R [2,3], with focus on the practical use.

I will present examples of the application of this from the clinical literature, including a practical approach to interval censoring. I will present some practical advice on how modeling of transitions in multistate models should be approached and what to expect in clinical studies. I will demonstrate the use of `simLexis` to estimate otherwise intractable quantities from complex multistate models, through simulation of transitions based on parametric models for transition intensities.

Keywords

MULTISTATE DATA REPRESENTATION, PREDICTION, SIMULATION

References

1.IACOBELLI, S. and CARSTENSEN, B. (2013): Multiple time scales in multi-state models. *Statistics in Medicine, 32(30), 5315–5327.*

2.PLUMMER, M. and CARSTENSEN, B. (2011): Lexis: An R class for epidemiological studies with long-term follow-up. *Journal of Statistical Software, 38(5), 1–12.*

3.CARSTENSEN, B. and PLUMMER, M. (2011): Using Lexis objects for multi-state models in R. *Journal of Statistical Software, 38(6), 1–18.*

Time varying network models for brain imaging data

Ivor Cribben[1] and Yi Yu[2]

[1] Alberta School of Business, University of Alberta cribben@ualberta.ca
[2] University of Bristol y.yu@bristol.ac.uk

Abstract. In functional magnetic resonance imaging (fMRI) studies, the networks between brain regions are assumed to be stationary over time. However, there is now more evidence that the network is changing over time even when the subjects are at rest. In the first part of this talk, we formulate the problem in a high-dimensional time series framework and introduce a data-driven method which detects change points in the network structure of a multivariate time series, with each component of the time series represented by a node in the network. In the second part of this talk, we introduce a new time varying approach that is model-free, data-adaptive, and is applicable in situations where the (global) stationarity of the time series from the brain regions fails, such as the cases of local stationarity and/or change points. We apply both new methods to simulated data and to a resting-state fMRI data set.

Keywords

Spectral clustering; Binary Segmentation; Wild Binary Segmentation; Network change points; Stationary bootstrap; fMRI.

Extended Poisson-Tweedie models with some examples

Clarice G.B. Demétrio[1], John Hinde[2], and Wagner H. Bonat[3]

[1] ESALQ/USP, Piracicaba, Brazil clarice.demetrio@usp.br
[2] National University of Ireland, Galway, Ireland john.hinde@nuigalway.ie
[3] Paraná Federal University, Curitiba wbonat@ufpr.br

Abstract. The standard Poisson and binomial generalized linear models are inadequate to analyse count and proportion data when there is evidence of some form of over/under-dispersion or zero-inflation, see Nelder and McCullagh (1989) and Hinde and Demétrio (1999). The extended Poisson-Tweedie models proposed by Bonat et al (2016) is a new class of models to analyse count data, with variance $\mu + \phi\mu^p$, where μ is the mean, ϕ and p are the dispersion and Tweedie power parameters, respectively. This class of models provides a flexible and comprehensive family including many standard discrete models. The family provides for modelling of overdispersed count data, including Neyman Type A, Polya-Aeppli, negative binomial, Poisson-inverse Gaussian and Hermite distributions, and can also accommodate zero-inflation and underdispersion. We provide here a set of examples illustrating the extended Poisson-Tweedie modelling approach for over and under-dispersion.

Keywords

Count data; overdispersion; underdispersion; zero-inflation; extended Poisson-Tweedie model.

References

Bonat, W.H.; Jrgensen, B.; Kokonendji, C.C., Hinde, J.; Demt́rio, C.G.B. (2016) Extended Poisson-Tweedie: properties and regression models for count data (submmitted).

Hinde, J.; Demt́rio, C.G.B. (1998) Overdispersion: Models and estimation. *Computational Statistics and Data Analysis.* **27**, 151–170.

McCullagh, P.; Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall.

On some Bayesian spatio-temporal epidemic models

Nikos Demiris[1]

Athens University of Economics and Business nikos@aueb.gr

Abstract. Epidemic data often possess certain characteristics, such as the presence of many zeros, the spatial nature of the disease spread mechanism and environmental noise. This work addresses these issues via suitable Bayesian modelling. In doing so we utilise a general class of stochastic regression models appropriate for spatio-temporal count data with an excess number of zeros. The developed regression framework does incorporate serial correlation and time varying covariates through an Ornstein Uhlenbeck process formulation. In addition, we explore the effect of different priors, including default options and techniques based upon variations of mixtures of g-priors. The effect of different distance kernels for the epidemic model component is investigated. We proceed by developing branching process-based methods for testing scenarios for disease control. This link between traditional epidemiological models and stochastic epidemic processes, useful in policy-focused decision making, is discussed in detail. Model selection is determined by taking a predictive view with different scoring rules being explored. The approach is illustrated through application to data from foot and mouth and sheep pox outbreaks.

Dealing with under-reported data through INAR-hidden Markov chains

A. Fernández-Fontelo[1], A. Cabaña[1], P. Puig[1] and D. Moriña[2]

[1] Departament de Matemàtiques, Universitat Autònoma de Barcelona, Bellaterra, Spain. amanda@mat.uab.cat, acabana@mat.uab.cat, ppuig@mat.uab.cat

[2] Unit of Infections and Cancer (UNIC), Cancer Epidemiology Research Program (CERP), Catalan Institute of Oncology (ICO) - IDIBELL, Spain.dmorina@iconcologia.net

Abstract. Since the introduction of the Integer-Valued AutoRegressive (INAR) models (Al-Osh and Alzaid, 1987), the interest in the analysis of count series has been growing. The main reason for this increasing popularity is the limited performance of the classical series analysis approach when dealing with discrete valued series. With the introduction of discrete time series analysis techniques, several challenges appeared such as unobserved heterogeneity, periodicity, under-reporting, ... Many efforts have been devoted to introduce seasonality in these models (Moriña et al., 2011) and also coping with unobserved heterogeneity. However, the problem of under-reported data is still in a quite early stage of study in many different fields, leading to potentially biased inference and also invalidating the main assumptions of the classical models. The model we will present considers the observed discrete series of counts Y_t which may be under-reported, and the hidden discrete series X_t with an INAR(1) structure $X_t = \alpha \circ X_{t-1} + W_t$, where $0 < \alpha < 1$ is a fixed parameter and W_t is Poisson(λ). The *binomial thinning* $\circ$ operator is defined as $\alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} Z_i$, where Z_i are i.i.d Bernoulli r.v. with probability α. The way we allow Y_t to be under-reported is by defining that Y_t is X_t with probability $1-\omega$ or is $q \circ X_t$ with probability ω. Several examples of application of the model in the field of public health will be discussed, using data regarding incidence and mortality attributable to diseases related to occupational or environmental exposures, ... Full details in Fernández-Fontelo et al. (2016).

Keywords

binomial sub-sampling, forward probabilities, under-recorded data

References

AL-OSH, M. A. and ALZAID, A. A. (1987). First-order integer-valued autoregressive INAR(1) process. *Journal of Time Series Analysis*, **8**, 261–275.

MORIÑA, D., PUIG, P., RÍOS, J., VILELLA, A. and TRILLA, A. (2011). A statistical model for hospital admissions caused by seasonal diseases. *Statistics in Medicine*, **30**, 3125–3136.

FERNÁNDEZ-FONTELO, A., CABAÑA, A., PUIG, P. AND MORIÑA, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, **35**, 4875-4890.

Tensor-based methods for the analysis of hospital data

Paolo Giordani[1] and Henk A.L. Kiers[2]

[1] Sapienza University of Rome, Rome, Italy paolo.giordani@uniroma1.it
[2] University of Groningen, Groningen, The Netherlands h.a.l.kiers@rug.nl

Abstract. In several situations the research interest leads to the analysis of a collection of observations on which a set of variables are registered. However, in life sciences, it is very frequent that the information is replicated in different occasions. The occasions can be time-varying or refer to different conditions. In such cases the data can be stored in a three-way array or tensor. The Candecomp/Parafac (CP) and Tucker3 (T3) models represent the most common methods for analyzing three-way tensors. In this work these methods are discussed from a practical point of view and applied in order to study a three-way data set concerning the admissions to a hospital in Rome (Italy) during fifteen years distinguished in three groups of consecutive years (1892–1896, 1940–1944, 1968–1972). The analysis allows us to discover the evolution of health condition in Rome.

Keywords

THREE-WAY DATA, TENSOR-BASED METHODS, TUCKER3, HOSPITAL DATA

References

CARROLL, J.D. and CHANG, J.J. (1970): Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35, 283–319.*

GIORDANI, P., KIERS, H.A.L and DEL FERRARO, M.A. (2014): Three-way component analysis using the R package ThreeWay. *Journal of Statistical Software, 57 (7), 1–23.*

HARSHMAN, R.A. (1970): Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics, 16, 1–84.*

TUCKER, L.R (1966): Some mathematical notes on three-mode factor analysis. *Psychometrika, 31, 279–311.*

Cost effective models for prediction and causal effect estimation: the goldmine of disease registers

Els Goetghebeur[1]

Ghent University, Belgium

Abstract. Ever expanding clinical registers call for efforts to learn about treatment effectiveness in practice. This can happen by avoiding unmeasured confounders when adjusting outcome regression on exposure. Embracing too many covariates may however hamper performance with higher measurement costs and reduced durability, leading to registration fatigue with less precise and more missing data. Against the background of a potentially high dimensional covariate space, we propose a frequentist approach for cost-efficient selection of variables.

From patient-specific baseline covariates with added hospital effects in generalized linear models, we estimate individual mortality risks first and derive directly standardized risks. Missing data are handled by complete case analysis or multiple imputation. For both targets a minimum error criterion is offset by cost constraints on covariate measurements via stochastic search algorithms like the basic hill-climber or parallel tempering, possibly sped up through an initial generalized LASSO screen. Predicting 30-day mortality for pneumonia patients thus lands with smaller error than the Bayesian population RJMCMC approach in a fraction of the search time. Retained covariates in the Swedish Riksstroke register depend in part on the target of estimation. Unexpectedly, the more involved NIHSS measure of baseline disease severity was outperformed by a simple consciousness indicator.

Combining Item Response Theory with Multiple Imputation to Equate Health Assessment Questionnaires

Roee Gutman[1] and Chenyang Gu[2]

[1] Department of Biostatistics, Brown University, Providence, RI, USA
roee_gutman@brown.edu
[2] Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
gu@hcp.med.harvard.edu

Abstract. The assessment of patients functional status across the continuum of care requires a common patient assessment tool. However, assessment tools that are used in various health care settings differ and cannot be easily contrasted. For example, the Functional Independence Measure (FIM) is used to evaluate the functional status of patients who stay in inpatient rehabilitation facilities, the Minimum Data Set (MDS) is collected for all patients who stay in skilled nursing facilities, and the Outcome and Assessment Information Set (OASIS) is collected if they choose home health care provided by home health agencies. All three instruments or questionnaires include functional status items, but the specific items, rating scales, and instructions for scoring different activities vary between the different settings. We consider equating different health assessment questionnaires as a missing data problem, and propose a variant of predictive mean matching method that relies on Item Response Theory (IRT) models to impute unmeasured item responses. Using real data sets, we simulated missing measurements and compared our proposed approach to existing methods for missing data imputation. We show that, for all of the estimands considered, and in most of the experimental conditions that were examined, the proposed approach provides valid inferences, and generally has better coverages, relatively smaller biases, and shorter interval estimates. The proposed method is further illustrated using a real data set.

Keywords

IRT, Missing Data, Predictive Mean Matching, Equating

Discovery of structural brain imaging markers of HIV-associated outcomes using connectivity-informed regularization approach

Jaroslaw Harezlak[1], Marta Karas[1], Damian Brzyski[1], Mario Dzemidzic[2], Joaquin Goni[3], Beau Ances[4], and Timothy W. Randolph[5]

[1] Indiana University, Bloomington, IN, USA harezlak@iu.edu
[2] Indiana University School of Medicine, Indianapolis, IN, USA
[3] Purdue University, West Lafayette, IN, USA
[4] Washington University, Saint Louis, MO, USA
[5] Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Abstract. Study of multimodal brain imaging biomarkers of disease is frequently performed by analyzing each modality separately. In our work, we use a recently proposed regularization method, riPEER (ridge-identity partially empirical eigenvectors for regression), to discover early biomarkers of HIV-associated outcomes including CD4 count and HIV RNA plasma level. Specifically, we incorporate information arising from the functional and structural connectivity in the penalized generalized linear model framework to inform the associations between the brain cortical features and disease outcomes. Penalty terms are defined as a combination of Laplacian matrices arising from the functional and structrucal connectivity adjacency matrices. We study the advantages of employing different measures of connectivity as well as synergistic functional and structural information. Finally, we address the issue of using different cerebral cortex parcellations, from the common FreeSurfer parcellations (68 and 148 cortical areas) to a novel multi-modal parcellation into 360 cortical areas, in discovering global and local biomarkers.

Keywords

Linear Regression, Penalized methods, Structured penalties, Laplacian matrix, Brain connectivity, Brain structure

References

Randolph, T. W., Harezlak, J., and Feng, Z. (2012). Structured penalties for functional linear models – partially empirical eigenvectors for regression. *Electronic Journal of Statistics, 6:323–353.*

Karas, M., Brzyski, D., Dzemidzic, M., Goni, J., Kareken, D.A., Randolph, T.W. and Harezlak J. (2017). Brain connectivity-informed regularization methods for regression *bioRxiv 117945*; doi: https://doi.org/10.1101/117945

A general framework for selection bias due to missing data in EHR-based research

Sebastien Haneuse[1], Sarah Peskoe[1], David Arterburn[2], and Michael Daniels[3]

[1] Harvard T.H. Chan School of Public Health
[2] Group Health Research Institute
[3] University of Texas - Austin

Abstract. Electronic health records (EHR) data provide unique opportunities for public health research in part because they typically contain rich information on large populations. Notwithstanding their benefits, however, EHR-based studies may suffer from a number of sources of bias. Among these selection bias due to incomplete data is an underappreciated source of bias in analyzing EHR data. When framed as a missing-data problem, standard methods are often applied to control for selection bias. In EHR-based studies, however, the provenance of the observed data generally involves the interplay of many clinical decisions made by patients, health care providers, and the health system; thus standard methods fail to capture the complexity of the mechanism that give rise to the observed data. In this work we use a novel framework for selection bias in EHR-based research that allows for a hierarchy of missingness mechanisms to inform an inverse-probability weighted estimator that better aligns with the complex nature of EHR data. We show that this estimator is consistent and asymptotically normal. Based off extensive simulations, a key insight is the bias-variance trade-off in using this framework when the data provenance is functionally misspecified. We use this approach to adjust for selection in an on-going, multi-site EHR-based study of bariatric surgery on BMI.

Why many researchers misuse variable selection and how to prevent this

Georg Heinze[1][2] and Daniela Dunkler[1]

[1] Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Austria
[2] STRATOS TG2
georg.heinze@meduniwien.ac.at

Abstract. Statistical models are handy tools for empirical medical research. They facilitate individualized outcome prognostication conditional on covariates as well as adjustments of estimated effects of risk factors on the outcome by covariates. Theory of statistical models is well-established if the set of covariates to consider is fixed and small, such that we can assume that effect estimates are unbiased and the usual methods for confidence interval estimation are valid. In routine work, however, it is not known a priori which covariates should be included in a model, and often we are confronted with the number of candidate variables in the range 10-30. This number is often too large to be considered in a statistical model.

We reviewed the practice of variable selection in several medical journals, which revealed that variable selection methods were often applied inappropriately. We attribute this to five myths about variable selection circulating among applied researchers which we will briefly discuss (HEINZE and DUNKLER, 2017).

Using real examples and simulated data, we will discuss implications of variable selection, e.g., on bias and uncertainty of a model. Currently, routine software implementations largely ignore instability issues caused by variable selection. We demonstrate how resampling methods can help to quantify instability issues in real data analyses. We advocate the implementation of these methods in routine software such that applied researchers can get aware of instabilities in the finally selected model, in particular if they attempt to apply variable selection to too small data sets.

Keywords

EPIDEMIOLOGICAL METHODS, MODEL UNCERTAINTY AND DIAGNOSTICS, SURVIVAL ANALYSIS

References

HEINZE, G. and DUNKLER, D. (2017): Five myths about variable selection. *Transpl Int, 30, 6–10.*

Over/Under-dispersion, zero-inflation and Poisson-Tweedie models

John Hinde[1], Clarice Demétrio[2], and Wagner Bonat[3]

[1] National University of Ireland, Galway, Ireland john.hinde@nuigalway.ie
[2] ESALQ/USP, Piracicaba, Brazil clarice.demetrio@usp.br
[3] Paraná Federal University, Curitiba wbonat@ufpr.br

Abstract. The standard distributions for the analysis of count and proportion data are the Poisson and binomial distributions. Frequently, in practice they are too restrictive in that the variability in the data is either significantly greater (overdispersed) or less (underdispersed) than that implied by the model's variance function. For the analysis of count data, Nelder and McCullagh (1989) says that overdispersion is the norm and not the exception and this has been well studied, see Hinde and Demetrio (1999) and many subsequent articles presenting a wide range of distributions. Although less common, underdispersion can arise, typically from dependent responses. For instance, when there is competition between plants and animals this can induce negative correlation in temporal and spatial counting processes. Here we will also consider how underdispersion can occur as a result of features of the underlying counting, or data collection, process. The range of distributions for modelling underdispersed count data is relatively limited, although models can be derived in specific situations.

A class of general models is presented based on Poisson-Tweedie factorial dispersion models with variance $\mu + \phi\mu^p$, where μ is the mean, ϕ and p are the dispersion and Tweedie power parameters, respectively. This class of models provides a flexible and comprehensive family including many standard discrete models. The family provides for modelling of overdispersed count data, including Neyman Type A, Polya-Aeppli, negative binomial, Poisson-inverse Gaussian and Hermite distributions, and can also accommodate zero-inflation and underdispersion. For a general approach we consider an extended version of the Poisson-Tweedie model and discuss estimation of regression, dispersion and Tweedie power (variance function) parameters.

Keywords

Count data; overdispersion; underdispersion; zero-inflation; Poisson-Tweedie model.

References

Hinde, J.; Demetrio, C.G.B. (1998) Overdispersion: Models and estimation. *Computational Statistics and Data Analysis.* **27**, 151–170.

McCullagh, P.; Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall.

Causality, Prediction, and Everything in Between: The Promise and Pitfalls of Electronic Health Records Data

Joseph W. Hogan[1]

Center for Statistical Sciences, Brown University

Abstract. Electronic health records (EHR) provide rich information on healthcare delivery and health outcomes. The uses of EHR information include billing, patient monitoring, and storage of complex information such as images and physician notes. More recent interest has focused on the utility of EHR for development of prediction rules, clinical decision support, and comparative effectiveness studies. However, unlike data from clinical trials and cohort studies, data from EHR are not collected according to a protocol or study design; in that sense they are 'experiential' or 'found' data. This talk illustrates some of the challenges of using EHR for drawing inference about causal effects, and describes statistical methods needed to address them. We illustrate using two case studies from a large HIV care program in Kenya: one examining timing of antiviral treatment initiation among those with HIV/TB coinfection, and another looking at the causal effect of recently recommended 'test-and-treat' policy.

Joint work with Liangyuan Hu, Hana Lee, Becky Genberg, Paula Braitstein, Yizhen Xu, Michael Daniels, Rami Kantor, Ann Mwangi

A Bayesian approach to estimate changes in condom use from limited human immunodeficiency virus prevalence data

Kostas Kalogeropoulos[1]

London School of Economics K.Kalogeropoulos@lse.ac.uk

Abstract. Evaluation of large-scale intervention programmes against human immunodeficiency virus (HIV) is becoming increasingly important, but impact estimates frequently hinge on knowledge of changes in behaviour such as the frequency of condom use over time, or other self-reported behaviour changes, for which we generally have limited or potentially biased data. We employ a Bayesian inference methodology that incorporates an HIV transmission dynamics model to estimate condom use time trends from HIV prevalence data. Estimation is implemented via particle Markov chain Monte Carlo methods, applied for the first time in this context. The preliminary choice of the formulation for the time varying parameter reflecting the proportion of condom use is critical in the context studied, because of the very limited amount of condom use and HIV data available. We consider various novel formulations to explore the trajectory of condom use over time, based on diffusion-driven trajectories and smooth sigmoid curves. Numerical simulations indicate that informative results can be obtained regarding the amplitude of the increase in condom use during an intervention, with good levels of sensitivity and speci- ficity performance in effectively detecting changes. The application of this method to a real life problem demonstrates how it can help in evaluating HIV interventions based on a small number of prevalence estimates, and it opens the way to similar applications in different contexts. Joint work with: J. Dureau, P. Vickerman, M. Pickles and M.C. Boily

Multiple discrete distributions for modelling heaped count data in health insurance

Dimitris Karlis[1], Lluis Bermùdez[2], and Miguel Santolino[2]

[1] Department of Statistics, Athens University of Economics and Business, Greece, karlis@aueb.gr
[2] University of Barcelona, Spain

Abstract. In many circumstances, when working with count data, we observe a large number of spikes in the frequency distribution. This phenomenon occurs in several disciplines, like demography for example, known as heaping, while it is also known as digit preference and some other names depending on the application domain. In this talk, a new modelling approach, based on finite mixtures of multiple discrete distributions of different multiplicities, is proposed to fit data with a lot of periodic spikes in certain values. An EM algorithm is provided in order to ensure the models' ease-of-fit and then a simulation study is presented to show its efficiency. A numerical application with a real data set involving the length, measured in days, of inability to work after an accident occurs is treated. The main finding is that the model provides a very good fit when working week, calendar week and month multiplicities are taken into account.

Keywords

Multiple Poisson, Heaped data, Work Disability Days, EM algorithm

References

Bermùdez, Lluis, Dimitris Karlis, and Miguel Santolino. (2017) A finite mixture of multiple discrete distributions for modelling heaped count data. *Computational Statistics & Data Analysis*, 112, 14–23.

A new longitudinal time-varying measurement error model with application to physical activity assessment instruments in a large biomarker validation study.

Victor Kipnis[1]

Biometry, National Cancer Institute, USA

Abstract. Systematic investigations into the structure of measurement error of different physical activity instruments are lacking. Whether existing instruments consist of objective measurements made by accelerometers or involve self-report on questionnaires or recalls, their measurement errors may contain bias as well as random variation. In lieu of observed true physical activity levels, to estimate those different error components, it is necessary to have some unbiased biomarker measurements such as those made by doubly labeled water (DLW). Existing measurement error models treat an individuals level of physical activity as a fixed quantity over a long period of time.However, physical activity involves both short-term (e.g., month-to-month) and long-term (over years) variation over time. We describe a longitudinal measurement error model that accounts for such variationand apply it to the analysis ofdata on physical activity energy intake from a largevalidation study of different physical activity instruments using DLW as reference measurements. We show that this time-varying measurement error model fits the data better than the fixed long-term physical activity assumption.Accountingfor the time element in physical activityassessment is crucial to avoid biases in evaluation of the effects of measurement error.

A Bidirectional Multi-State Model for Panel Data on Bone Mineral Density among HIV-Infected Patients

Klaus Langohr[1], Guadalupe Gómez[1], Nuria Pérez[2], Eugenia Negredo[2], and Anna Bonjoch[2]

[1] Universitat Politècnica de Catalunya/BARCELONATECH, Barcelona, Spain
klaus.langohr@upc.edu, lupe.gomez@upc.edu
[2] Fundació Lluita contra la Sida, Badalona, Spain nperez@flsida.org,
enegredo@flsida.org, abonjoch@flsida.org

Abstract. Bone mineral density (BMD) measurements are used to determine bone health and can help to identify the risk of fracture. The most widely recognized BMD scan, which measures bone density at different parts of the body, is called dual-energy x-ray absorptiometry (DXA). The DXA measures are compared to the BMD of a healthy 30-year-old adult of the same gender and are converted into T-scores: T-scores above -1 are considered normal, values between -1 and -2.5 indicate low bone mass (osteopenia), and values below -2.5 indicate osteoporosis. The main goals of the present study are to study the evolution of BMD over time in a cohort of more than 700 HIV-infected persons with at least two DXA scans and to determine the risk factors for the progression of bone loss.

For this purpose, a bidirectional multi-state model with states normal BMD, osteopenia, and osteoporosis is fitted to the data. The model considers four possible transitions —normal BMD to osteopenia, osteopenia to normal BMD, osteopenia to osteoporosis, and osteoporosis to osteopenia— which are studied as a function of age and antiretroviral treatment. Due to the nature of the panel data available, all transition times are interval-censored.

This multi-state model allows us to estimate the transition probabilities and predict the percentages of patients in every health state as a function of age and treatment. The clinical relevance of building such a model is to guide the clinical practice and to rationalize DXA scans measurements.

Keywords

BONE MINERAL DENSITY, MULTI-STATE MODEL, PANEL DATA

A Nonparametric Algorithm for Independent Component Analysis with Application to fMRI Data

Shanshan Li[1], Shaojie Chen[2], Chen Yue[2], and Brian Caffo[2]

[1] Indiana University R.M. Fairbanks School of Public Health, Indianapolis, IN, USA
[2] Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Abstract. Independent Component analysis (ICA) is a widely used technique for separating signals that have been mixed together. In this work, we propose a novel ICA algorithm using density estimation and maximum likelihood, where the densities of the signals are estimated via p-spline based histogram smoothing and the mixing matrix is simultaneously estimated using an optimization algorithm. The algorithm is exceedingly simple, easy to implement and blind to the underlying distributions of the source signals The performance of the proposed algorithm is evaluated in different simulation settings. For illustration, the algorithm is applied to a research investigation with a large collection of resting state fMRI datasets. The results show that the algorithm successfully recovers the established brain networks.

Keywords

Blind source separation, density estimation, functional MRI, p-spline bases

References

Li, S., Chen, S., Yue, C. and Caffo, B. (2016): A Parcellation Based Nonparametric Algorithm for Independent Component Analysis with Application to fMRI Data *Front Neurosci, 10, 15.*

Evaluating diagnostic accuracy of biomarkers in the presence of missing biomarkers

Shanshan Li[1]

Indiana University R.M. Fairbanks School of Public Health, Indianapolis, IN, USA
sl50@iu.edu

Abstract. The Receiver Operating Characteristic (ROC) curve is a common tool for evaluating diagnostic accuracy of biomarkers. The existing work on estimating ROC functions are mostly developed for data collected under ideal settings. In many medical studies, measurements of biomarkers are subject to missingness due to high cost or limitation of technology. To deal with the missing data problem in biomarker studies, we propose an augmented weighted distribution model that incorporates information from subjects with incomplete data. The resulting estimator enjoys the double-robustness property in the sense that it remains consistent if either the missing data process or the conditional distribution of the missing data given the observed data is correctly specified. We derive the asymptotic properties of the proposed estimators and evaluate their performances using extensive numerical studies.

Approximate Bayesian Computation for large-scale gene duplication networks

Antonietta Mira[1] and Jukka-Pekka Onnela[2]

[1] Università della Svizzera Italiana, Lugano and Università degli Studi dell' Insubria, Como.
[2] Department of Biostatistics , Harvard University

Abstract. Many systems of scientific interest can be investigated as networks, where network nodes correspond to the elements of the system and network edges to interactions between the elements. Increasing availability of large-scale biological data and steady improvements in computational capacity are continuing to fuel the growth of this field. Network models are now used commonly to investigate biological complexity at the systemic level. Gene duplication is one of the main drivers of the evolution of genomes, and network models based on gene duplication were one of the first large-scale models used in systems biology. An attractive feature of some of these so-called duplication-divergence models is their analytical tractability, but there is typically no statistically principled way to estimate their model parameters from empirical data. This is a reflection of a more general divide between the two prominent paradigms to the modeling of networks, which are the approach of mechanistic networks models and the approach of statistical network models. Mechanistic network models assume that the microscopic mechanisms governing network formation and evolution at the level of individual nodes are known, and questions often focus on understanding macroscopic features that emerge from repeated application of these known mechanisms. The statistical approach, in contrast, often starts from observed network structures and attempts to infer some aspects about the underlying data generating process. Mechanistic network models provide insight into how the network is formed and how it evolves at the level of individual nodes, but as mechanistic rules typically lead to complex network structures, it is difficult to assign a probability to any given network realizations that a mechanistic model may generate. Because of this difficulty, there is typically no closed form expression for likelihood for these models and, consequently, likelihood based inference for learning from data is not possible. We have developed a principled statistical framework, based on Approximate Bayesian Computation, to bring some of the mechanistic network models into the realm of statistical inference. This approach is feasible because given a set of parameter values, it is easy to sample network configurations from most mechanistic models. I will introduce this general framework and demonstrate its application to large-scale gene duplication networks, where it can be used to infer model parameters, and their associated uncertainties, for mechanistic network models from empirical data.

Time-sequential multiple imputation of missing data in time-dependent covariates used in a Cox model

Havi Murad[1], Alla Berlin[2], Rachel Dankner[2,3], Liraz Olmer[1], Paolo Boffetta[4], Laurence Freedman[1,3]

[1] Biostatistics Unit, Gertner Institute, Tel-Hashomer, ISRAEL
[2] Unit for Cardiovascular Epidemiology, Gertner Institute, Tel-Hashomer, ISRAEL
[3] School of Public Health, Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, ISRAEL ,
[4] School of Medicine, Mount Sinai Hospital, New York, New York, United States

Abstract. We describe a project to explore the links between diabetes and cancer through analyzing data from the Clalit Health Maintenance Organization Database in Israel. The database includes up to eleven years longitudinal data on 567,347 persons diagnosed with diabetes (2002-2012). Information on cancer incidence was added from the Israel Cancer Registry that we linked to the Clalit file. In this talk we will focus on the following question: among persons with diabetes are those with higher glucose levels or HbA1c levels at a higher (or lower) risk of developing cancer? The Cox model with time-dependent covariates is appropriate for answering this question. However, since at any given time-point there is a large proportion of missing data in some time-dependent covariates (30%-50% in HbA1c; 20%-40% in glucose), we developed a procedure for handling the missing values. White and Royston1 proposed a method for imputing covariates under the Cox proportional hazards model. We generalize their method to time-dependent covariates, and develop a procedure for time-sequential multiple imputation at each time-point for the missing HbA1c and glucose values using the chained equations method, based on completed variables from previous time-points. Simulations under different missing data structures to examine the performance of this procedure are presented, together with the results of the analysis of the data for different cancers.

Keywords

Missing data, Chained Equations Method, Multiple Imputation (MI), Time-sequential MI, Cox model, Time-dependent covariates

References

White and Royston (2009). Imputing missing covariate values for the COX model. *Stat in Med. 28:1982-98.*

Power-Expected-Posterior Priors for Generalized Linear Models

Dimitris Fouskakis[1], Ioannis Ntzoufras[2], and Konstantinos Perrakis[2]

[1] Department of Mathematics, National Technical University of Athens, 15780, Athens, Greece fouskakis@math.ntua.gr
[2] Department of Statistics, Athens University of Economics and Business, 10434 Athens, Greece ntzoufras@aueb.gr,kperrakis@aueb.gr

Abstract. The power-expected-posterior (PEP) prior developed for variable selection in normal regression models provides an objective, automatic, consistent and parsimonious model selection procedure. At the same time it resolves the conceptual and computational problems due to the use of imaginary data. These attributes allow for large sample approximations, when needed, in order to reduce the computational burden under more complex models. In this work we generalize the applicability of the PEP methodology, focusing on the framework of generalized linear models (GLMs), by introducing two new PEP definitions which are in effect applicable to any general model setting. Hyper prior extensions for the power-parameter that regulates the contribution of the imaginary data are further considered. Under these approaches the resulting PEP prior can be asymptotically represented as a double mixture of g-priors. For estimation of posterior model and inclusion probabilities we introduce a tuning-free Gibbs-based variable selection sampler. Several simulation scenarios and one real data example are considered in order to evaluate the performance of the proposed methods compared to other commonly used approaches based on mixtures of g-priors. Empirical results indicate that the GLM-PEP adaptations are more effective when the aim is parsimonious inference.

Keywords

Expected-posterior prior, g-prior, Generalized linear models, hyper-g priors, Imaginary data, Objective Bayesian model selection, Power-prior.

References

BAIER, D. and GAUL, W. (1999): Optimal Product Positioning Based on Paired Comparison Data. *Journal of Econometrics, 89, 365–392.*

BOCK, H.H. (1974): *Automatische Klassifikation.* Vandenhoeck & Ruprecht, Göttingen.

BRUSCH, M. and BAIER, D. (2002): Conjoint Analysis and Stimulus Presentation: a Comparison of Alternative Methods. In: K. Jajuga, A. Sokołowski and H.H. Bock (Eds.): *Classification, Clustering, and Analysis.* Springer, Berlin, 203–210.

Sieve Bootstrap for Functional Time Series

Efstathios Paparoditis

University of Cyprus, Department of Mathematics and Statistics

Abstract. A bootstrap procedure for functional time series is proposed which exploits a general vector autoregressive representation of the time series of Fourier coefficients appearing in the Karhunen-Loève expansion of the functional process. A double sieve-type bootstrap method is developed which avoids the estimation of process operators and generates functional pseudo-time series that appropriately mimic the dependence structure of the functional time series at hand. The method uses a finite set of functional principal components to capture the essential driving parts of the infinite dimensional process and a finite order vector autoregressive process to imitate the temporal dependence structure of the corresponding vector time series of Fourier coefficients. By allowing the number of functional principal components as well as the autoregressive order used to increase to infinity (at some appropriate rate) as the sample size increases, a basic bootstrap central limit theorem is established which shows validity of the bootstrap procedure proposed for functional finite Fourier transforms. Simulations illustrate the good finite sample performance of the new bootstrap method proposed.

Keywords

Fourier transform, Principal components, Karhunen-Loève expansion, Spectral density operator

References

E. Paparoditis (2016): Sieve Bootstrap for Functional Time Series. *Submitted.*

Bayesian block-diagonal variable selection and model averaging

Omiros Papaspiliopoulos[1]

Universitat Pompeu Fabra omiros.papaspiliopoulos@upf.edu

Abstract. We propose a scalable algorithmic framework for exact Bayesian variable selection and model averaging in linear models under the assumption that the Gram matrix is block-diagonal, and as a heuristic for exploring the model space for general designs. In block-diagonal designs our approach returns the most probable model of any given size without resorting to numerical integration. The algorithm also provides a novel and efficient solution to the frequentist best subset selection problem for block-diagonal designs. Posterior probabilities for any number of models are obtained by evaluating a single one-dimensional integral and other quantities of interest such as variable inclusion probabilities and model averaged regression estimates are obtained by an adaptive, deterministic one-dimensional numerical integration. The overall computational cost scales linearly with the number of blocks, which can be processed in parallel, and exponentially with the block size, rendering it most adequate in situations where predictors are organized in many moderately-sized blocks. For general designs, we approximate the Gram matrix by a block-diagonal matrix using spectral clustering and propose an iterative algorithm that capitalizes on the block-diagonal algorithms to explore efficiently the model space. All methods proposed in this article are implemented in the R library mombf. Joint work with Daivd Rossell.

Machine learning analysis for assessing body composition and bone mass in HIV infected patients

Nuria Perez-Alvarez[1], Esteban Vegas[2], and Carla Estany[3]

[1] Department of Statistics and Operations Research,Technical University of Catalonia-Barcelona Tech nuria.perez@upc.edu
[2] Department of Statistics, University of Barcelona evegas@ub.edu
[3] Fight against AIDS Foundation, HUGTIP cestany@flsida.org

Abstract. Osteoporosis is widely recognised as an important public health problem because of the significant morbidity, mortality and costs associated with its complications (fractures of the hip, spine, forearm and other skeletal sites). The body weight is associated with the bone mass, however the mediation by lean or fat body mass in of bone density is uncertain. Bone DXA examinations are used to classify the patients into normal, osteopenia and osteoporosis status by means of the World Health Organization criteria, which has a proven ability to predict fracture risk.

The data set contains the DXA scans, clinical and demographic variables related to a HIV infected cohort of patients (1475 patients and 89 variables). The goals are to evaluate the relationship among the lean, fat and bone mass parameters and to identify the fat, lean and demographic variables that can inform the bone disease classification. The techniques used were scatterplots and calculation of correlation coefficients for the concordance assessment and multivariate analysis to identify outliers and to determine the profile of patients with bone injury. Non supervised and supervised multivariate techniques, such as principal component analysis (PCA), kernel PCA, random forests, CART and support vector machine techniques, were applied.

The machine learning approach can help in a variety of clinical settings to elucidate patterns and characteristics of patients presenting a specific disease. In our case, few differences among the body composition in normal, osteopenia and osteoporosis diagnosed patients were found. The results obtained by the different techniques are displayed and compared.

Keywords

MACHINE LEARNING, BODY COMPOSITION, BONE MASS, OSTEOPOROSIS

Underdispersion in a bivariate framework

Pedro Puig[1], Jordi Valero[2], and Amanda Fernandez-Fontelo[1]

[1] Departament de Matematiques, Universitat Autonoma de Barcelona
ppuig@mat.uab.cat, amanda@mat.uab.cat

[2] Escola Superior d'Agricultura de Barcelona, Universitat Politecnica de Catalunya
jordi.valero@upc.edu

Abstract. In studies of porcine reproductive and respiratory syndrome, reproduc- tive records of farms are analysed by considering the number of live-born (X) and dead-born (Y) piglets along births. This bivariate count data has interesting properties: negative correlation, under/over-dispersion of X and/or Y depending on the parity number, and under/over-dispersion of the total number of piglets $Z = X + Y$ (Fernandez et. al, 2017). Let Z be a count random variable, and let $\xi_1; \xi_2; \ldots$ be iid Bernoulli variables with probability of success $\alpha \in (0; 1]$, all them independent of Z. The count variable, $Z_\alpha = \sum_{i=1}^{Z} \xi_i$, $Z_\alpha = 0$ if $Z = 0$); is called an independent binomial α-thinning or binomial subsampling of Z. Each α-thinning induces a natural bivariate decomposition of Z of the form $(Z_\alpha; Z - Z_\alpha)$, so that the number of live-born piglets could be interpreted as a binomial subsampling $X = Z_\alpha$ of the total number of piglets Z. Unfortunately, this nice representation does not allow rich dispersion and correlation patterns such that those observed in the records of farms. In this talk we will introduce some new mechanisms leading to under-dispersion based on the binomial thinning operation (Puig et al., 2017). It is known that the Poisson distribution is closed under α-thinnings, but if α depends of the number of Poisson realizations the resulting distribution can be underdispersed. Moreover, in this case the bivariate decomposition is richer in dispersion and correlation structures, allowing to describe the behaviour of the birth records.

Keywords

UNDERDISPERSION, BIVARIATE DISTRIBUTION, COUNT DATA

Conflict diagnostics in evidence synthesis: examples from infectious disease models

Anne Presanis[1]

MRC Biostatistics Unit, University of Cambridge anne.presanis@mrc-bsu.cam.ac.uk

Abstract. Bayesian evidence synthesis, where multiple independent data sources contribute to the likelihood, is becoming increasingly employed in various fields, including infectious disease epidemiology. Evidence synthesis methods are most useful for estimating quantities which can't be directly observed, but for which indirect evidence is available: for example, prevalence of undiagnosed HIV infection or the case-fatality risk for influenza. However, the use of multiple sources informing common parameters entails the potential for different datasets to provide conflicting or inconsistent inferences about the common parameters. The detection and measurement of such conflict is therefore a crucial step in the model criticism process. Cross-validatory posterior predictive methods have previously been proposed for conflict assessment (”node-splitting”) in graphical models. However, the systematic assessment of conflict, at multiple locations throughout a graphical model, provokes the multiple testing problem. We therefore present a framework for systematic conflict diagnostics, accounting for the multiple null hypothesis tests of no conflict. We illustrate the method through syntheses to estimate HIV prevalence in Poland and influenza severity in the UK.

Functional principal component analysis and trajectory reconstruction with accelerated longitudinal designs

Philip T. Reiss[1]

University of Haifa and New York University reiss@stat.haifa.ac.il

Abstract. This talk will consider the use of functional principal component analysis (FPCA) for reconstructing growth trajectories when the data arise from an accelerated longitudinal design. Such a design entails following individuals of varying ages for a short period in order to study development over a wide age range. In functional data analysis terminology, the resulting data are sparse as opposed to dense; and a standard approach to FPCA for sparse data is to smooth the covariance surface and then estimate its eigenfunctions (Staniswalis and Lee, 1998; Yao et al., 2005). But in this setting there are no data pairs far from the main diagonal, so when the smoothing step is performed with penalized splines (e.g., Goldsmith et al., 2013), the penalization strategy can have an inordinate influence on the results. This problem and proposed solutions will be illustrated with cortical thickness data derived from a longitudinal magnetic resonance imaging study.

Keywords

cortical thickness, eigenfunctions, functional covariance, penalized splines

References

GOLDSMITH, J., GREVEN, S., and CRAINICEANU, C. (2013). Corrected Confidence Bands for Functional Data Using Principal Components. *Biometrics, 69, 41–51. STANISWALIS, J. G., and LEE, J. J. (1998). Nonparametric Regression Analysis of Longitudinal Data.* Journal of the American Statistical Association, 93, 1403–1418.

YAO, F., MÜLLER, H.-G., and WANG, J.-L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association, 100, 577–590.*

Bayesian variable selection: what if residuals are non-normal?

David Rossell[1]

Universitat Pompeu Fabra david.rossell@upf.edu

Abstract. A main challenge in high-dimensional variable selection is enforcing sparsity. Because of theoretical and computational considerations most research are based on linear regression with Normal errors, but in actual applications errors may not be Normal, which can have a particularly marked effect on Bayesian inference. We extend the usual Bayesian variable selection framework to consider more flexible errors that capture asymmetry and heavier-than-normal tails. The error structure is learnt from the data, so that the model automatically reduces to Normal errors when the flexibility is not needed. We show convenient properties (log-likelihood concavity, simple computation) that render the approach practical in high dimensions. Further, although the models are slightly non-regular we show that one can obtain asymptotic sparsity rates under model misspecification. We also shed some light on an important consequence of model misspecification on Bayesian variable selection, namely a potential for a marked drop in power to detect truly active coefficients. This is confirmed in our examples, where we also illustrate computational advantages of inferring the residual distribution from the data.

Functional data analysis of neuroimaging signals on the cerebral cortex

John A.D. Aston[1], Eardi Lila[1], and Laura M. Sangalli[2]

[1] University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK
j.aston@statslab.cam.ac.uk, e.lila@maths.cam.ac.uk
[2] Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy
laura.sangalli@polimi.it

Abstract. Motivated by the analysis of high-dimensional neuroimaging signals over the cerebral cortex, we introduce a principal component analysis technique that can be used for exploring the variability and performing dimensional reduction of signals observed over two-dimensional manifolds. The proposed method is based on a PDE regularization approach, involving the Laplace-Beltrami operator associated to the manifold domain. It can be applied to data observed over any two-dimensional Riemannian manifold topology. The proposed method is applied to the study of main connectivity patterns of neural activity in the cortex, based on the analysis of a dataset made available by Human Connectome Project and consisting of resting state functional magnetic resonance imaging scans from about 500 healthy volunteers.

Keywords

Functional data analysis, principal component analysis, data on bi-dimensional manifold

References

LILA, E., ASTON, J.A.D., SANGALLI, L.M. (2016): Smooth Principal Component Analysis over two-dimensional manifolds with an application to Neuroimaging, *Annals of Applied Statistics, 10 (4), 1854–1879.*

Functional Covariates in Additive Regression Models

Fabian Scheipl[1] and Sonja Greven[1]

Institute for Statistics, Ludwig-Maximilians-Universität München
{fabian.scheipl,sonja.greven}@stat.uni-muenchen.de

Abstract. We describe and evaluate estimators of nonparametric and semiparametric linear and non-linear effects of functional covariates on functional responses that bridge spline-based and functional principal component-based approaches to functional data regression models. Our implementation is embedded in an extensive framework for mixed additive regression models for correlated functional responses. We provide easy-to-use open source software in the `pffr()` function for the R-package refund.

Keywords

functional data analysis, penalized likelihood, functional principal components, splines, generalized additive model

References

SCHEIPL, F. and GREVEN, S. (2016). Identifiability in penalized function-on-function regression models. *Electronic Journal of Statistics, 10:1, 495–526.*

SCHEIPL, F. and GERTHEISS, J. and GREVEN, S. (2016) Generalized functional additive mixed models. *Electronic Journal of Statistics, 10:1, 1455–1492.*

GOLDSMITH, J. and others (2016) refund: Regression with Functional Data. R package version 0.1-16, `https://CRAN.R-project.org/package=refund`

Measurement Error in Nutritional Epidemiology: Impact, Current Practice for Analysis, and Opportunities for Improvement

Pamela A. Shaw[1] Veronika Deffner, Kevin W. Dodd, Laurence S. Freedman, Ruth H. Keogh, Victor Kipnis, Helmut Kuechenhoff, and Janet A. Tooze for STRATOS TG4 Working Group[2]

[1] Department of Biostatistics and Epidemiology University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA, shawp@upenn.edu

[2] STRengthening Analytical Thinking for Observational Studies (STRATOS) Initiative, http://www.stratos-initiative.org

Abstract. Dietary exposures are variable, complex and difficult to measure precisely. The impact of such errors is often not well understood or is ignored by many researchers. As part of the STRengthening Analytical Thinking for Observational Studies (STRATOS) Initiative, a Task Group on measurement error and misclassification (TG4) is currently engaged in several activities to increase the awareness of this problem among biostatisticians and epidemiologists and point to methods to address it. This presentation focuses on nutritional epidemiology, though the issues discussed apply more broadly.

Studies in nutritional epidemiology usually rely on self-reported dietary intake data, though these are known to be subject to random error as well as intake-related and person-specific systematic errors. The potential impact of different types of measurement error will be illustrated. Measurement error reduces power to detect diet-disease associations and can result in biased estimates of such associations. A recent survey by TG4 revealed a lack of understanding that this bias is not always in the form of an attenuation. Corrections for measurement error for estimating associations can be made if the nature of the error can be ascertained; doing so requires additional information, for example from objective biomarkers.

We describe the analytical challenges of measurement error that arise in nutritional epidemiology and discuss some practical design and statistical methods to address them, focusing on regression calibration. Finally, we highlight challenges for improving practice and summarize some initial recommendations for investigators, as well as for editors and reviewers.

Keywords

Measurement error, Misclassification, Nutritional Epidemiology, STRATOS

An improved estimation method for gray matter volume in presence of white matter hyperintesities in Alzheimer's and Down syndrome studies

Dana Tudorascu[1], Helmet Karim[1], Bill Klunk[1], Brad Christian[2], and Ciprian Crainiceanu[3]

[1] University of Pittsburgh, Pittsburgh, PA dlt30@pitt.edu
[2] University of Wisconsin, Madison, WI bchristian@wisc.edu
[3] Johns Hopkins University, Baltimore, MD ccraini1@jhu.edu

Abstract. Automated segmentation of the brain is a challenging task in the presence of brain pathologies such as white matter hyperintensities (WMH). WMHs appear as hyperintense areas in magnetic resonance imaging (MRI) and are frequently found in Alzheimer's disease (AD) and Down syndrome (DS) population's brain. WMHs present a challenge for standard segmentation algorithms that misclassify WMHs as gray matter (GM). An improved segmentation method is presented based on filling (1,2) the WMH areas on MRI scans with normal appearing white matter (NAWM) values sampled from the distribution of NAWM. An automated algorithm (3) was first used to detect WMH that were later filled with NAWM intensities on the MRI scans. The images were then segmented again and the resulting GM volume was compared with that computed from a standard approach. Repeated measures models and neurological case studies were used for the analyses. The fill-in method showed significant improvement in tissue classification for AD and DS.

Keywords

Gray matter segmentation, WMH, MRI, Alzheimer, Down Syndrome

References

Eloyan A et al. (2014). Health Effects of Lesion Localization in Multiple Sclerosis:Spatial Registration and Confounding Adjustment. *PLoS ONE,doi:10.1371/journal.pone.0107263*

Karim HT et al. (2016). The effects of white matter disease on the accuracy of automated segmentation. *Psychiatry Res, 14, 253–257.*

Wu M et al. (2006). A fully automated method for quantifying and localizing white matter hyperintensities on MR images. *Psychiatry Res, 144, 133–142.*

Inverse probability of censoring weights under missing not at random with applications to long-term clinical outcomes

Constantin T. Yiannoutsos[1]

Indiana University R.M. Fairbanks School of Public Health, Indianapolis, IN, USA
cyiannou@iupui.edu

Abstract. Background. Estimate the response to antiretroviral therapy (ART) among HIV-positive patients who start ART in sub-Saharan Africa. Dealing with death when estimating longitudinal measurements and produce a survival-adjusted longitudinal measure (e.g., median CD4 count, percent of patients with perfect ART adherence) and address counterfactual scenarios (e.g., What is the value of the measure had everyone remained on f/u but not necessarily in care). Methods. Inverse Probability of Censoring Weighting (IPCW) methods readily available to estimate median CD4 count over time but MAR assumption likely not applicable in our setting. Use patient tracing (double-sampling) data and modify IPCW for the MNAR setting (MNAR-IPCW). Results. Both longitudinal measures considered (median CD4 count, perfect adherence) were overestimated, even compared to the best-case scenario of everyone having remained under observation and in care. Conclusions. Ignoring biases resulting from non-random losses to follow-up may results in significant biases when estimating longitudinal measurements. These results have broad application on a number of longitudinal biomarkers in this setting, particularly those related to the long-term viral suppression necessary to maintain good outcomes among people living with HIV around the world.

Two-stage semiparametrics analysis of skeletal growth around pubertal growth spurt with interval-censored observations

Chenghao Chu[1], Ying Zhang[1], and Wanzhu Tu[1]

Indiana University R.M. Fairbanks School of Public Health, Indianapolis, IN, USA
yz73@iu.edu

Abstract. Human individuals acquire their adult body shapes through vigorous physical growth in the first two decades of life. Many of the somatic characteristics that define our physical appearance in adulthood take shape around the time of pubertal growth spurt (PGS). An analytical challenge to quantify growth rates before and after PGS is the lack of direct observation of the anchoring PGS event. We propose a two-stage semiparametric analysis to assess the rates of skeletal changes around the PGS with interval-censored observation on the PGS. The first stage is the nonparametric maximum likelihood estimation for the distribution of PGS timing. In the second stage, a least-squares based method is used to estimate the model parameters, including the pre and post-PGS growth rates with latent time of PGS. We show that under mild regularity conditions, the estimators are consistent and asymptotically normal. Statistical inference ensues from the large sample theory. We conduct a simulation study to evaluate the operating characteristics of the proposed method. Analysis of growth data from an observational cohort shows that in comparison to girls, boys tend to have a more sustained skeletal growth after PGS, as evidenced by the greater post-PGS growth rates in the upper body. The findings suggest that strong and sustained post-PGS skeletal growth contributes to the sexual dimorphism in human body.

Contributed papers

DNA: an economists perspective

Giovanni Barone Adesi[1]

Dep. Of Finance and SFI, Faculty of Economics, USI, Lugano, Switzerland baroneg@usi.ch

Abstract. Why does the DNA code use four elements? It is well known from information theory that the most efficient code to process information should have a number of elements equal to e, Nepers number. Computers use binary codes, rounding e down. The choice of a ternary code would be a bit more efficient than the binary, but it is generally not implemented. The DNA sequence uses four bases in two spirals, with apparent redundancy and some constraints. Redundancy in any alphabet is necessary to allow for the possibility of error correction. In the case of the DNA, its two-spiral, four-letter structure is shown to allow for an error-correction mechanism twenty-one times better than an ordinary computer, which just relies on one parity bit. The DNA mechanism reduces the error rate at the cost of substantial redundancy. Therefore it becomes apparent that the DNA has been designed to ensure long-term memory, at the cost of a substantial reduction of its capacity to store information. It is shown that four is indeed the most efficient number of alphabet elements to achieve that.

The elegant solution found for the design logic of the DNA cannot be extended to the protein alphabet, which has twenty-one letters, possibly because of the much more varied functions proteins fulfill.

Modeling of Biochemical Networks via a New Graphical Approach

Melih Ağraz[1] and Vilda Purutçuoğlu[2]

[1] Middle East Technical University, Department of Statistics, 06800, Ankara, Turkey
agraz@metu.edu.tr
[2] Middle East Technical University, Department of Statistics, 06800, Ankara, Turkey
vpurutcu@metu.edu.tr

Abstract. The Gaussian graphical model is one of the well-known graphical approaches which describes the interactions between the genes in a biochemical system via the inverse of the covariance matrix, also called the precision matrix, when the states are represented by the multivariate normal distribution in a lasso regression. In inference of this model, various approaches have been suggested from the graphical lasso (Friedman et al., 2008) to the neighbourhood selection method (Meinshausen and Bühlmann, 2006). Although these approaches are successful in accuracy and computational efficiency if the system has small or moderate dimensions, their performances decrease if the states are non-normal and the system has high dimension. Hereby, in order to ameliorate these challenges, initially, we consider to use the multivariate student-t distribution in the description of the states due to the fact that it can be more robust choice for the states and its convergent distribution also covers the normality. Furthermore, we apply the modified maximum likelihood estimation method (MMLE) in the estimation of the model parameters. MMLE (Tiku, 1967) is a modified version of the ordinary MLE in the sense that it can smooth the nonlinearity in the likelihood function, which causes multiple roots in the solution set, by means of the first order Taylor series expansion and order statistics. In order to evaluate the performance of our model and inference strategy, we use bench-marks real datasets and compare our results with the GGM outputs based on accuracy and computational time. The findings show that our model indicate better accuracy over GGM without losing the computational demand in the construction of the actual biochemical systems.

Keywords

Gaussian graphical model, modified maximum likelihood estimation, lasso regression, biochemical networks.

Acknowledgements

The author would like to thank the BAP project at the Middle East Technical University (Project no: BAP-01-09-2017-002) for their supports.

References

TIKU, M. (1967): Estimating the mean and standard deviation from a censored normal sample, *Biometrika, 54, 155–165.*

Integrating Latent Classes in the Bayesian Shared Parameter Joint Model of Longitudinal and Survival Outcomes

Eleni-Rosalina Andrinopoulou[1], Kazem Nasserinejad[2], and Dimitris Rizopoulos[3]

[1] Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands
e.andrinopoulou@erasmusmc.nl
[2] Department of Hematology, Erasmus MC, Rotterdam, The Netherlands
k.nasserinejad@erasmusmc.nl
[3] Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands
d.rizopoulos@erasmusmc.nl

Abstract. When patients are monitored after a kidney transplantation, it is of clinical interest to investigate the association between the longitudinal biomarker protein-to-creatinine ratio and time-to kidney failure. A feature of this data set is that different sub-populations exhibit different longitudinal profiles. Patients can be categorized in several sup-groups (latent classes) with different trajectories. Therefore, to better model the association between the longitudinal and the survival outcome these latent classes should be taken into account.

The joint model of longitudinal and survival data constitutes a popular framework to analyze such data sets. In particular, two paradigms within this framework are the shared parameter joint models and the joint latent class models. The former paradigm allows to quantify the strength of the association between the longitudinal and survival outcomes but does not allow for latent sub-populations. On the other hand, the latter paradigm explicitly postulates the existence of sub-populations but does not directly quantify the strength of the association.

To answer our motivating research question we propose to integrate latent classes in the shared parameter joint model in a fully Bayesian approach. Specifically, the model allows us to investigate the association between protein-to-creatinine ratio and time-to kidney failure within each latent class. We, furthermore, focus on the selection of the true number of latent classes.

Keywords

JOINT MODEL, LONGITUDINAL OUTCOME, SURVIVAL OUTCOME, LATENT CLASS MODEL

On a class of design-based adaptive and sequential sampling strategies

Federico Andreis[1,2] and Marco Bonetti[1,2]

[1] Department of Policy Analysis and Public Management Università Commerciale Luigi Bocconi federico.andreis@unibocconi.it
[2] Carlo F. Dondena Centre for Research on Social Dynamics and Public Policy

Abstract. When sampling from a finite population, it is sometimes the case that a specific characteristic of the units composing the population under study is of interest, besides the estimation target; a notable example is the Tubercolosis (TB) Prevalence survey program set up by the World Health Organization (WHO), where together with the estimation of the overall TB prevalence in a country, a relevant objective is the detection of as many cases as possible, so that they can be treated. Classic sampling approaches have proven to be inefficient in carrying out both tasks at the same time, whereas more recent approaches - namely, adaptive designs - suffer from other operationally important drawbacks (such as lack of control over the final sample size and costs). We discuss a novel class of sampling strategies aimed at providing both the ability to oversample prescribed subsets of a finite population, and the possibility to draw inference building on the classic Horvitz-Thompson approach to estimation. The proposed class of methods is design-based, has a sequential and adaptive nature and integrates the use of machine learning algorithms in the sampling procedure. We compare our proposal to WHO's current practice and some state-of-the-art alternatives by means of a simulation study on scenarios inspired by the motivating example, providing empirical evidence concerning the estimation of population parameters and the oversampling feature.

Keywords

FINITE POPULATION SURVEY, π-PS DESIGN, ADAPTIVE SAMPLING, MACHINE LEARNING

References

GLAZIOU, P., VAN DER WERF, M. J., ONOZAKI, I. and DYE, C. (2008): Tuberculosis prevalence surveys: rationale and cost. *International Tuberculosis Lung Disease.* **12(9)**, 1003–1008.

THE WORLD HEALTH ORGANISATION (2011): Tubercoulosis PREVALENCE SURVEYS: a handbook. WHO Press, Geneva

Summary indicators to assess the performance of risk predictors

Laura Antolini[1], Elena Tassistro[1], Davide Paolo Bernasconi[1] and Maria Grazia Valsecchi[1]

School of Medicine and Surgery, University of Milano-Bicocca, Via Cadore 48, 20900
Monza, Italy laura.antolini@unimib.it

Abstract. The availability of novel biomarkers opens room for refining prognosis by adding factors on top of those having an established role. It is accepted that the impact of novel factors should not rely solely on regression coefficients and their significance but on predictive power measures. However, novel factors who are promising at the explorative stage often results in disappointingly low impact on the predictive power measures. This motivated the proposal of the net reclassification index and the integrated discrimination improvement, as direct measures of gain due to additional factors. This measures became extremely popular however, recent contributions in the biostatistical literature enlightened strong limitations. A further measure proposed a decade ago, namely the net benefit, appears to be promising in assessing the consequences in terms of costs and benefits when using a risk predictor in practice for classification. This presentation reviews the conceptual formulations and interpretations of the available graphical methods and summary measures for evaluating risk predictor models.

Keywords

Risk predictor; performance measure; biomarker; classification; net benefit

References

Lee M-LT, Gail M, Pfeiffer R, Satten G, Cai T, Gandy A. (2013). Risk assessment and evaluation of predictions. *Springer* New York

Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. (2014). Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology* 25(1):114-121

Vickers AJ, Van Calster B,Steyerberg EW. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 352:i6

Joint Models for Longitudinal and Time-to-Event Data in a Case-Cohort Desig

Sara Baart[1], Dimitris Rizopoulos[2], and Eric Boersma[3]

[1] Erasmus MC, Rotterdam, the Netherlands s.baart@erasmusmc.nl
[2] Erasmus MC, Rotterdam, the Netherlands d.rizopoulos@erasmusmc.nl
[3] Erasmus MC, Rotterdam, the Netherlands h.boersma@erasmusmc.nl

Abstract. Longitudinal measurements are becoming increasingly more popular in clinical research. By repeatedly measuring a patient, their progress is monitored more closely and temporal patterns can be estimated leading to improved prediction of outcomes. A popular approach in combining longitudinal and time-to-event data is the joint modelling approach. Often a set of multiple biomarkers is measured to discover new biomarkers predictive for the outcome. Costs associated with assessing all biomarker values, however, can become exceedingly high. If, in addition, event rates in the study are low and most information is to be expected from the patients experiencing the event (cases), it may be more cost efficient not to assess all biomarkers. For this research we are motivated by the BIOMArCS study, where patients admitted for acute coronary syndrome (ACS) are followed for one year to study the relation between temporal patterns of multiple biomarkers and recurring ACS. The BIOMArCS study follows a case-cohort design in which a random subcohort of patients is selected and supplemented with the other cases outside the subcohort. In standard survival models, weighting schemes have been proposed to account for the overrepresentation of cases in such designs. In the framework of joint modelling different approaches are needed. We propose to include survival information and any potential baseline covariate information of all patients in the analysis. The controls outside the subcohort will have missing values for the biomarker measurements. However, since the subcohort was chosen at random, the missingness mechanism is missing at random (MAR), and hence results obtained from the joint model fitted in the constructed data set will remain valid. We evaluate this procedure with simulations and illustrate its use in the BIOMArCS study.

Keywords

CASE-COHORT DESIGN, LONGITUDINAL BIOMARKERS, SURVIVAL ANALYSIS, JOINT MODELS

Non-proportional hazards or unobserved heterogeneity in clustered survival data: Can we tell the difference?

Theodor Adrian Balan[1] and Hein Putter[1]

Leiden University Medical Center, The Netherlands t.a.balan@lumc.nl

Abstract. In survival analysis, shared frailty (random effect) models are often used to model heterogeneous survival data, such as clustered failures or recurrent events. In Hougaard (2000) a large family of infinitely divisible frailty distributions were proposed, including the positive stable and compound Poisson distributions with a probability mass at 0. The estimation of semiparametric frailty models with these distributions has proved challenging.

The assumption of proportional hazards, usually made conditional on the frailty, does not carry over to the marginal model for most random effect distributions. It has been shown that, for univariate data, this makes it impossible to distinguish between the presence of a frailty or marginal non-proportional hazards. Difficulties also arise when the data consists of small sized clusters. When modelling the effects of covariates on the hazard, if proportional hazards are assumed conditional on the frailty, this assumption does not carry over to the marginal model for most random effect distributions. For univariate data, this implies that it is impossible to distinguish between the presence of a frailty or non-proportional hazards at marginal level.

We discuss the results of a simulation study carried out in the situation where the clusters have a small size or individuals have few recurrent events. For a large number of frailty distributions, we analyze the behaviour of test statistics for the presence of the frailty and for the proportional hazards assumption. With a novel software implementation for estimating semiparametric shared frailty models, we discuss the situations when the unobserved heterogeneity can be distinguished from the non-proportional hazards. The practical implications are illustrated in real-world data analysis examples. We study the behaviour of the test statistics for the presence of the random effects and sensitivity to the proportional hazard assumptions for a large number of frailty distributions. For this, we introduce a novel software implementation for estimating semiparametric shared frailty models.

Keywords

SURVIVAL ANALYSIS, SHARED FRAILTY, PROPORTIONAL HAZARDS

References

YASHIN, AI., IACHINE, IA., BEGUN, AZ. and VAUPEL, JW. (2001): Hidden frailty: myths and reality. *Research report, Odense University.*

HOUGAARD, P. (2000): *Analysis of Multivariate Survival Data.* Springer-Verlag, New York

Estimation of Smooth ROC Curves for Biomarkers With Limits of Detection

Leonidas E. Bantis[1] Qingxiang Yan[1] John V. Tsimikas[2] and Ziding Feng[1]

[1] Dept. of Biostatistics, The University of Texas MD Anderson Cancer Center.
lebantis@mdanderson.org
[2] Dept. of Mathematics, Division of Statistics and Actuarial-Financial Mathematics, University of the Aegean.

Abstract. Protein biomarkers found in plasma are commonly used for cancer screening and early detection. Measurements obtained by such markers are often based on different assays that may not support detection of accurate measurements due to a limit of detection (LOD). The ROC curve is the most popular statistical tool for the evaluation of a continuous biomarker. However, in situations where LODs exist, the empirical ROC curve fails to provide a valid estimate for the whole spectrum of the false positive rate (FPR). Hence, crucial information regarding the performance of the marker in high sensitivity and/or high specificity values is not revealed. In this paper, we address this problem and propose methods for constructing ROC curve estimates for all possible FPR values. We explore flexible parametric methods, transformations to normality, robust kernel-based and spline-based approaches. We evaluate our methods though simulations and illustrate them in colorectal and pancreatic cancer data.

Keywords

BOX-COX, CENSORING, CLASSIFICATION, ROC, GENERALIZED GAMMA, KERNELS, LIMITS OF DETECTION, SPLINE.

Estimating metabolite networks subject to dietary preferences in the longitudinal setting

Georgios Bartzis[1],[2], Hae-Won Uh[1], Fred van Eeuwijk[2], and Jeanine J. Houwing-Duistermaat[3]

[1] Leiden University Medical Center, Leiden, the Netherlands G.Bartzis@lumc.nl
[2] Wageningen University and Research Center, Wageningen, the Netherlands
[3] University of Leeds, Leeds, the UK

Abstract. The metabolome is the intermediate between DNA variation and clinical phenotypes and is expected to provide a better predictor of phenotypes than DNA variation or gene expression. Although, several studies have examined the interplay between diet and metabolism, to date, no studies have investigated how metabolic patterns are influenced by dietary variation ($\boldsymbol{F}$) while genetic contribution ($\boldsymbol{G}$) and time ($\boldsymbol{T}$) have been modeled in the context of network analysis. We propose using linear mixed effect models for capturing the correlation coming from repeated measurements and decomposing the total metabolite information into parts relevant to $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{T}$; the part relevant to $\boldsymbol{F}$ instead of the original values should further be used for network estimation. Dietary information resulted from Food Frequency Questionnaires is typically summarized by using Exploratory Factor Analysis (*EFA*). Genetic information is quantified by using Polygenic Risk Scores (*PRS*).

For estimating, describing and visualizing networks, a correlation-based network estimation method, i.e. Weighted Gene Co-expression Network Analysis (WGCNA) is used. In the cohort considered here (DILGOM study) all sources of metabolic variation were measured. In the resulting networks, several groups of biologically associated metabolites (VLDL, HDL, AA/BCAA, *omega*-3 FA) were clustered together based on their association to 6 known diets. The novelty of our method is on taking into account all sources of metabolic variation and resulting in networks with higher interconnectedness and interpretability, meaning identifying meaningful metabolite groups sharing similar association to $\boldsymbol{F}$.

Keywords

Dietary Preferences; Longitudinal Measurements; Metabolite Networks

References

BARTZIS, G., HOUWING-DUISTERMAAT, J.J., EEUWIJK, F.v. and UH, H.W.: Estimation of metabolite networks with regard to a specific covariable. *Metabolomics, in preparation.*

A Comprehensive Simulation Study for Comparison of Statistical Methods in MRMC ROC Studies

Merve Basol[1*], Dincer Goksuluk[1], and A. Ergun Karaagaoglu[1]

Department of Biostatistics, Hacettepe University - Turkey
merve.basol@hacettepe.edu.tr

Abstract. ROC analysis is often used to determine performance of a diagnostic test or compare performances of two or more diagnostic tests whose results are either numerical measurements or the readers' interpretations. In radiology, it is preferred that diagnostic tests to be interpreted by two or more readers rather than one reader since readers' interpretations are subjective. Furthermore, multiple tests might be performed on the same case in order to obtain more accurate results. This approach is called as multi-reader multi-case (MRMC) studies. In such studies, full factorial experimental design are commonly preferred study design. However, in this design there might be a correlation both between readers and between tests. Some statistical methods which consider these correlation structures are developed for comparing performances of diagnostic tests. In this study, we focused on some of these methods such as DBM (Dorfman-Berbaum-Metz), OR (Obuchowski-Rockette), BWC (Beiden-Wagner-Campbell) and MM (Marginal Model) [1,2]. The performances of each method is compared using a comprehensive simulation study. Continuous rating data was generated under null hypothesis. Type-I error rates and coverages of AUC differences are used to evaluate simulation results.

Keywords

MRMC, ROC curve, DBM, OR, diagnostic tests

References

1. ZHOU, X.H., OBUCHOWSKI, N.A., McCLISH, D.K. (2011). *Statistical Methods in Diagnostic Medicine.* John Wiley & Sons, Inc., New York.
2. SKARON, A., LI, K., ZHOU, X.H. (2012). Statistical Methods for MRMC ROC Studies. *Acad Radiol*, 19:1499–1507.

On the use of routine health insurance data - exemplified by the Austrians disease management program for diabetes mellitus

Andrea Berghold[1] and Regina Riedl[1]

Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2, 8036 Graz, Austria andrea.berghold@medunigraz.at, regina.riedl@medunigraz.at

Abstract. Evaluation of effectiveness of primary care-based disease management programs (DMPs) is essential, but rarely integrated in the roll-out of a program. Routine health insurance data provide a useful tool for evaluation of implemented DMPs. However, due to the observational design and the use of routine data, special care must be given to the design, analysis and interpretation. Strengths and limitations of secondary use of routine health insurance data and adequate methods for analysis will be discussed and illustrated by the Austrian DMP for type II diabetes mellitus. A population-based retrospective cohort study was conducted considering patient-relevant outcomes (overall mortality, cardiovascular disease) and economic impact over a four years follow-up. The DMP-group consisted of 7181 participants enrolled in the program during 2008-2009. In the routine health insurance database 208.532 controls with DM type 2 were identified based on antidiabetic drug therapy. A comparable control group was derived using propensity score matching (PSM) taking demographics, antidiabetic drug therapy, prescriptions, hospital admissions and days, main discharge diagnoses and costs at baseline into account.

By using PSM, we were able to ensure comparable groups for a large number of measured confounders, however we cannot rule out an influence by unmeasured confounding. Despite these limitations our results indicate a survival benefit and an average reduction of costs for participants in the DMP compared with the controls.

Keywords

secondary use of health data, propensity score, retrospective cohort study

References

RIEDL, R., ROBAUSCH, M. and BERGHOLD, A. (2016): The Evaluation of the Effectiveness of Austrians Disease Management Program in Patients with Type 2 Diabetes Mellitus - A Population-Based Retrospective Cohort Study. *PLoS ONE 11(8):e0161429*.

The anticipated odds ratios to decide the choice of a primary binary endpoint

Marta Bofill Roig[1] and Guadalupe Gómez Melis[2]

[1] Universitat Politècnica de Catalunya (UPC), Spain marta.bofill.roig@upc.edu
[2] Universitat Politècnica de Catalunya (UPC), Spain lupe.gomez@upc.edu

Abstract. Composite binary endpoints are widely chosen as primary endpoint in clinical trials. The use of composite endpoints entails difficulties in the interpretation of the results since the composite effect might not reflect the effect of its components. We propose a methodology to quantify the gain in efficiency of using the composite binary endpoint instead of its most relevant component as primary endpoint to lead the trial. The method, based on the Asymptotic Relative Efficiency (ARE), depends on six parameters including the degree of association between components, the event proportion and the effect of therapy given by the corresponding odds ratio of the single endpoints. We apply the ARE method to several scenarios defined by different values of the anticipated parameters and conclude with recommendations for discerning which could be the best suited primary endpoint given anticipated parameters.

Keywords

Asymptotic Relative Efficiency; Binary Endpoint; Clinical Trial; Composite Endpoint; Statistical Guidelines.

Branching processes in continuous time as models of mutations

Maroussia Slavtchova-Bojkova[1]

Faculty of Mathematics and Informatics, Sofia University St. Kliment Ohridski, 5, J. Bourchier Blvd, 1164 Sofia, Bulgaria bojkova@fmi.uni-sofia.bg

Abstract. The appearance of mutations in cancer development plays a crucial role in the disease control and its medical treatment. Motivated by the practical significance, it is of interest to model the event of occurrence of mutant cells that will possibly lead to a path of indefinite survival. A multi-type branching process model in continuous time is proposed for describing the relationship between the waiting time till the first escaping extinction mutant cell is born and the lifespan distribution of different types of cells, which due to the applied treatment have small reproductive ratio. A numerical method and related algorithm for solving the integral equations is developed, in order to estimate the distribution of the waiting time to the escaping extinction mutant cell is born.

Keywords

Decomposable branching processes, Continuous time, Mutations, Waiting time to escape mutant, Hazard function, Attaining high levels

References

Slavtchova-Bojkova, M., Trayanov, P., Dimitrov, S. (2017): Branching processes in continuous time as models of mutations: Computational approaches and algorithms. Computational Statistics and Data Analysis. http://dx.doi.org/10.1016/j.csda.2016.12.013

A Bayesian detection model for chronic disease surveillance: application to COPD hospitalisation data

Areti Boulieri[1] and Marta Blangiardo[2]

[1] Department of Epidemiology and Biostatistics, Imperial College London, UK
a.boulieri@imperial.ac.uk
[2] Department of Epidemiology and Biostatistics, Imperial College London, UK
m.blangiardo@imperial.ac.uk

Abstract. Disease surveillance is an important public health practice, as it provides information which can be used to make successful interventions and improve population health. In this work, we propose an extension of a Bayesian hierarchical model introduced by Li et al.(2012), which is appropriate for chronic disease surveillance. The model is able to describe spatial and temporal patterns of the disease, and also to detect areas that exhibit unusual temporal trends compared to the national one, which can be indicative of an emerged localised factor, a policy impact etc. In order to assess the performance of the model we carry out a simulation study considering a number of scenarios. The model is applied to a set of chronic obstructive pulmonary disease (COPD) hospitalisation data in England at clinical commissioning group (CCG) level, from April 2010 to March 2011. Finally, a web-based application that integrates the developed methodology is presented.

Keywords

Spatio-temporal data, Bayesian modelling, Chronic disease surveillance

References

Besag, J., York, J. and Mollie, A. (1991): Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics 43, 1–59.*

Boulieri, A., Hansell, A. and Blangiardo, M. (2016): Investigating trends in asthma and COPD through multiple data sources: a small area study. *Spatial and Spatio-Temporal Epidemiology, doi:10.1016/j.sste.2016.05.004.*

Li, G., Best, N., Hansell, A., Ahmed, I. and Richardson, S. (2012): BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice. *Biostatistics 13(4), 695–710.*

Overviews of Systematic Reviews with drugs, herbal medicine or dietary supplements

Konstantinos Bougioukas[1], Eirini Pagkalidou[1], Eleni Avgerinou[1], Apostolos Tsapas[1], Evangelia Ntzani[2], Emmanouil Smyrnakis[1], Anna-Bettina Haidich[1]

[1] School of Health Sciences, Department of Medicine, Aristotle University of Thessaloniki, Greece. mpougioukas@auth.gr
[2] University of Ioannina School of Medicine, Greece

Abstract. An overview of systematic reviews (OoSRs) is a study designed to synthesize the multiple evidence from existing systematic reviews (SRs) on a topic area. The objective of this study was to describe the basic characteristics of health-related OoSRs. We searched the Medline via Ovid, the Evidence-based Child Health journal and reference lists up to 31 December 2015 for eligible OoSRs with drags, herbal medicine or dietary supplements and harms-related content. The included studies were assessed with our pilot version of a preferred reporting checklist. We analyzed 54 OoSRs that were published between 2001 to 2015. The majority (34 [63.0%]) of corresponding authors were affiliated with institutions from the United Kingdom and Canada. A typical OoSRs included a median of 6 SRs and 77 primary studies involving more than 10,000 participants. The most common health problems examined in our sample were diseases of the respiratory system (15 [27.8%]) and mental and behavioural disorders (9 [16.7%]). The majority (35 [64.8%]) of the OoSRs did not report any information about a formal protocol and 15 (27,8%) studies did not present results for harms in the abstract. Most of OoSRs (39 [72,2%]) typically searched only one electronic database. Forty-three (79.6%) studies did not report language restrictions while methods for data extraction were reported only in 24 (44.4%) articles. Quality assessment of the included SRs was performed in 20 (37.0%) overviews and quality of evidence was presented only in 13 (24.1%) studies. Almost half of the OoSRs (24 [44.4%]) provided a quality sythesis while the remaining studies conducted a meta-analysis. Few studies (17 [31,5%]) considered about overlapping and publication bias was discussed only in 18 (33.3%) papers. The ratio of the number of studies with information for adverse events cited within the text to the total number of references of a typical OoSRs was one-fifth. Funding sources were not reported in almost half of the studies (24 [44.4%]). This study shows that OoSRs often lack completeness in reporting and methodological rigor. Strategies and guidelines are needed to improve this new type of study.

Keywords

overview of reviews, completeness in reporting, adverse events reporting

MTHFR C677T and multiple health outcomes: An umbrella review.

Emmanouil Bouras[1], Christos Kotanidis[1], Anthoula Chatzikyriakidou[1], Sofia Kouidou[1], Anna-Bettina Haidich[1]

Aristotle University of Thessaloniki, Health Sciences School, Faculty of Medicine.
ebouras@auth.gr

Abstract. MTHFR C677T (rs1801133) is a common variant affecting a key enzyme in one carbon metabolism. As such it has been implicated in the pathogenesis of numerous different health outcomes, over the years. Our aim is to examine the strength of each unique association between polymorphism MTHFR C677T and different health outcomes and provide an overview of potential biases. We searched Pubmed and Scopus from January 1st 1990 to December 22nd 2016 to identify systematic reviews and meta-analyses of observational studies. For each meta-analysis odds ratios (OR), 95% confidence intervals (CI) and 95% prediction intervals were calculated using random and fixed effects models. Between-study heterogeneity was assessed with I^2. Overall, we examined 81 unique meta-analyses that synthesized data from 1444 studies on different outcomes. Almost half of the outcomes (37 out of 81 meta analyses with random effects model) showed that the T allele of rs1801133 was associated with increased risk of developing a disease ($p<0.05$). A suggestive evidence of class II (more than 1000 cases, $p<0.001$ by random-effects model, heterogeneity <50%, primary studies in Hardy Weinberg Equilibrium) was only found for gastric non-cardia cancer, ischemic stroke, epilepsy and Down Syndrome. There is substantial evidence linking MTHFR to several health outcomes, but a considerable number of them may reflect, residual confounding, information bias, gene-gene and gene-environment interactions.

Keywords

polymorphism MTHFR C677T, systematic review, umbrella review, health outcome

Between-Sample Heterogeneity — Is AIC Really Optimal?

Mark J Brewer[1] and Adam Butler[1]

Biomathematics and Statistics Scotland, Craigiebuckler, ABERDEEN, AB15 8QH, UK.
Mark.Brewer@bioss.ac.uk

Abstract. Model selection is difficult, even in the apparently straightforward case of choosing between linear regression models. There has been a lively debate in the statistical ecology literature in recent years, where some authors have sought to evangelise AIC in this context while others have disagreed strongly.

A series of discussion articles in the journal Ecology in 2014 (e.g. Murtaugh, 2014; Burnham and Anderson, 2014) dealt with part of the issue: the distinction between AIC and p-values. But within the family of information criteria, is AIC always the best choice?

Theory suggests that AIC is optimal in terms of prediction, in the sense that it will minimise out-of-sample root mean square error of prediction. Earlier simulation studies have largely borne out this theory. However, we argue that since these studies have almost always ignored between-sample heterogeneity, the benefits of using AIC have been overstated.

Via a novel simulation framework, we show that relative predictive performance of model selection by different information criteria is heavily dependent on the degree of unobserved heterogeneity between data sets.

Acknowledgement This work was funded by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS).

Keywords

MODEL SELECTION, AIC, BIC, ECOLOGICAL APPLICATIONS

References

BURNHAM, K.P. and ANDERSON, D.R. (2004): Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods and Research, 33, 261–304.*

MURTAUGH, P.A. (2014): In defense of P values. *Ecology, 95, 611–617.*

Bayesian Longitudinal Circular Data Modeling

Onur Camli[1] and Zeynep Kalaylioglu[2]

[1] Department of Statistics, Middle East Technical University, Turkey camli@metu.edu.tr,
[2] Department of Statistics, Middle East Technical University, Turkey
kzeynep@metu.edu.tr,

Abstract. This work is motivated by a study that investigates the fetal head rotation trajectory during the first stage of natural labour adjusted for maternal characteristics and environmental factors. The particular challenge with such data is the model selection procedures that objectively asses the models when outcome data are longitudinal and circular/directional. Traditional model selection criteria employed for linear data such as AIC, BIC, and DIC may not be appropriate for circular variables as they do not take the account of directional properties of the data. On one hand, there is very few model selection criteria developed for circular data which are illustrated to crave for improvement. On the other hand, there is an increasing demand for such criteria as the directional data proliferates in many disciplines with the advancing technology particularly in medicine as illustrated in our motivating study in which a primitive measure (cervical dilation) have recently been replaced by an ultrasound technology measuring fetal head rotation to determine whether the birth will be natural. We construct a fully Bayesian random effects circular model and develop circular model selection criteria. One of our proposed criterion is based on angular distance. Extensive simulations evaluate and compare the performances of our model and model selection criteria under various realistic longitudinal settings involving longitudinal circular responses and baseline continuous covariates. Simulations reveal that the proposed model selection criteria have remarkable gain in selecting the appropriate random effects circular model.

Keywords

DIRECTIONAL STATISTICS; RANDOM EFFECTS; MODEL SELECTION; MEDICINE;BIOLOGY

Comparative Analysis of Calibration Methods for Microsimulation Models: An application on the MILC model

Stavroula A. Chrysanthopoulou[1,2]

[1] University of Massachusetts Medical School
Department of Quantitative Health Sciences
Stavroula.Chrysanthopoulou@umassmed.edu
qhs@umassmed.edu
[2] Brown University School of Public Health
Department of Biostatistics
Stavroula_Chrysanthopoulou@brown.edu
public_health@brown.edu

Abstract. Microsimulation Models (MSMs) have proven to be a very useful tool for describing complex disease processes, predicting individual-patient trajectories and assessing the impact of interventions on outcomes of interest. There are several applications of MSMs in Medical Decision Making for simulating intervention scenarios on populations and informing Public Health Policies.

The calibration procedure followed for specifying plausible values of the model parameters plays an essential role in the development of a valid MSM. This study provides a comparative analysis of the two main approaches for calibrating an MSM, a Bayesian and an Empirical technique. Both methods are applied to calibrate the MIcrosimulation Lung Cancer (MILC) model, a new, dynamic, continuous time MSM that describes the natural history of lung cancer, and predicts individual trajectories incorporating information about the age, sex, and smoking habits of the person.

Results from this study show that while empirical techniques are more efficient, Bayesian methods seem to perform better especially when calibration targets involve rare outcomes. Therefore it seems that a combination of the two approaches would be helpful. An empirical method should be applied first for an efficient search of the multidimensional parameter space and the identification of plausible ranges of values for the model parameters. Choosing appropriate starting values from the previously defined ranges a Bayesian method could further provide more accurate parameter values and a better fit of the MSM to available data.

Findings from this study suggest that a combination of an Empirical and a Bayesian method would be advantageous for a more effective calibration procedure of a microsimulation model.

Keywords

Microsimulation Models, Calibration, Medical Decision Making, MILC model

References

Chrysanthopoulou, S. A. (2013): Statistical methods in micro-simulation modeling: Calibration and predictive accuracy (doctoral dissertation). Brown University.

Chrysanthopoulou, S.A. (2014): Milc: Microsimulation lung cancer (milc) model [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=MILC (R package)

Rutter, C. M., Miglioretti, D. L., & Savarino, J. E. (2009): Bayesian calibration of microsimulation models. J Am Stat Assoc, 104(488), 1338-1350.

Rutter, C. M., Zaslavsky, A. M., & Feuer, E. J. (2011). Dynamic microsimulation models for health outcomes. Medical Decision Making, 31(1), 10-18.

Identifying clusters of physiological response elicited by an emotion recognition task

Federica Cugnata[1], Manuela Ferrario[2], Marco Airoldi[2], Marcello Arcangeli[2], Riccardo M. Martoni[3], Clelia Di Serio[1], and Chiara Brombin[1]

[1] Vita-Salute San Raffaele University, CUSSB (University Centre of Statistics in the Biomedical Sciences) cugnata.federica@hsr.it, diserio.clelia@unisr.it, brombin.chiara@hsr.it

[2] Politecnico di Milano, Department of Electronics, Information and Bioengineering (DEIB) manuela.ferrario@polimi.it

[3] San Raffaele Scientific Institute, Department of Clinical Neuroscience riccardo.martoni@hsr.it

Abstract. Stressful and emotion-inducing tasks trigger physiological reactions. In this work, we focus on changes in skin conductance (SC), i.e., the skin's ability to conduct electricity, recorded in 91 healthy individuals while completing the Reading Mind in Eye Test (RMET), a widely used assessment tool of emotion recognition aptitude. A previous exploratory study investigated the feasibility to use the Dynamic Time Warping (DTW) approach to analyze this data. This technique enabled to uncover three different templates (patterns) of the SC response. However, within this framework, it is complex to integrate in the analysis information on respondents' clinical and psychopathological traits, thus obtaining a better characterization of the identified cluster. To overcome this issue, in this work, we propose to apply Latent Class Mixed Models (LCMMs, Proust-Lima *et al.* 2015) which, generalizing traditional Linear Mixed Effects models, allow to identify latent classes (e.g., unobserved sub-populations), characterized by their own mean trajectories.

Keywords

DYNAMIC TIME WARPING, LATENT CLASS MIXED MODELS, SKIN CONDUCTANCE

References

Cecile Proust-Lima, Viviane Philipps, Amadou Diakite and Benoit Liquet (2015) *lcmm: Extended Mixed Models Using Latent Classes and Latent Processes.* R package version 1.7.2. http://CRAN.R-project.org/package=lcmm

A two-stage approach for estimating the parameters of an age-group epidemic model from incidence data

Itai Dattner[1]

Department of Statistics, University of Haifa, 199 Abba Khoushy Ave, Mount Carmel,
Haifa 3498838, Israel idattner@stat.haifa.ac.il

Abstract. Age-dependent dynamics is an important characteristic of many infectious diseases. Age-group epidemic models describe the infection dynamics in different age-groups by allowing to set distinct parameter values for each. However, such models are highly nonlinear and may have a large number of unknown parameters. Thus, parameter estimation of age-group models, while becoming a fundamental issue for both the scientific study and policy making in infectious diseases, is not a trivial task in practice. In this talk, we examine the estimation of the so called next-generation matrix using incidence data of a single entire outbreak, and extend the approach to deal with recurring outbreaks. Unlike previous studies, we do not assume any constraints regarding the structure of the matrix. A novel two-stage approach is developed, which allows for efficient parameter estimation from both statistical and computational perspectives. Simulation studies corroborate the ability to estimate accurately the parameters of the model for several realistic scenarios. The model and estimation method are applied to real data of influenza-like-illness in Israel. The parameter estimates of the key relevant epidemiological parameters and the recovered structure of the estimated next-generation matrix are in line with results obtained in previous studies.

Keywords

direct integral method; estimation; infectious diseases; optimization; smoothing

References

BAIER, D. and GAUL, W. (1999): Optimal Product Positioning Based on Paired Comparison Data. *Journal of Econometrics, 89, 365–392.*

BOCK, H.H. (1974): *Automatische Klassifikation.* Vandenhoeck & Ruprecht, Göttingen.

BRUSCH, M. and BAIER, D. (2002): Conjoint Analysis and Stimulus Presentation: a Comparison of Alternative Methods. In: K. Jajuga, A. Sokołowski and H.H. Bock (Eds.): *Classification, Clustering, and Analysis.* Springer, Berlin, 203–210.

The use of urn models in response-adaptive randomized designs: a simulation study

Valeria Edefonti[1], Andrea Ghiglietti[1], Maria Giovanna Scarale[2], and Rosalba Miceli[3]

[1] University of Milan, Milan, Italy `valeria.edefonti@unimi.it` `andrea.ghiglietti@unimi.it`
[2] IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy `mgiovanna.scarale@operapadrepio.it`
[3] Fondazione IRCCS Istituto Nazionale Tumori, Milan, Italy `rosalba.miceli@istitutotumori.mi.it`

Abstract. Recently, response-adaptive designs have been proposed in randomized trials to achieve ethical and cost advantages by using sequential accrual information collected during the trial to dynamically update the probabilities of treatment assignments. In this context, urn models - where the probability to assign patients to treatments is interpreted as the proportion of balls of different colors available in a virtual urn - have been used as response-adaptive randomization rules [1].

We propose the use of Randomly Reinforced Urn (RRU) models in a simulation study based on a randomized clinical trial on the efficacy of home enteral nutrition in cancer patients after major gastrointestinal surgery. We compare results (number of patients allocated to the inferior treatment and empirical power of the t-test for the treatment coefficient) obtained with the RRU design with those previously published with the non-adaptive approach. In detail, we simulate 10,000 trials based on the RRU model in three set-ups of different total sample sizes.

For each sample size, in approximately 75% of the simulation runs, the number of patients allocated to the inferior treatment by the RRU design is lower. The empirical power of the t-test for the treatment effect is similar in the two designs.

Keywords

RESPONSE-ADAPTIVE RANDOMIZATION, RANDOMLY REINFORCED URN MODEL, RANDOMIZED TRIALS, SIMULATION STUDY

References

1.ATKINSON, A.C. and BISWAS, A. (2014): *Randomised response-adaptive designs in clinical trials*. Chapman and Hall/CRC.

Combinatorial Mixtures of Multiparameter Distributions: an Application to Prostate Cancer

Valeria Edefonti[1] and Giovanni Parmigiani[2]

[1] University of Milan, Milan, Italy valeria.edefonti@unimi.it
[2] Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA gp@jimmy.harvard.edu

Abstract. The term *combinatorial mixtures* refers to a flexible class of parametric models for inference on mixture distributions whose components have multidimensional parameters [1]. The idea behind it is to allow each element of the component-specific parameter vector to be shared by a subset of other components.

We develop Bayesian inference and computational approaches based on Markov Chain Monte Carlo methods for this class of mixture distributions with an unknown number of components. We define the structure for a general prior distribution - a mixture of prior distributions itself - where a positive probability is put on every possible combination of sharing patterns. We illustrate our approach in an application based on the normal mixture model for bivariate data. We assume a decomposition of the covariance matrix which allows to model standard deviations and correlations separately. We also discuss solutions to the 'label switching' problem.

For our application, we use publicly available data on mRNA expression in prostate carcinoma [2], where a two-component 'ellipsoidal, varying volume, shape, and orientation' model has been suggested by a different approach [3].

Keywords

FINITE MIXTURE MODELS, UNKNOWN NUMBER OF COMPONENTS, BAYESIAN METHODS, PROSTATE CANCER, MRNA EXPRESSION DATA

References

1.FRUHWIRTH-SCHNATTER, S. (2006): *Finite mixture and Markov switching models.* Springer Science & Business Media.

2.THE CANCER GENOME ATLAS RESEARCH NETWORK et al. (2015): The molecular taxonomy of primary prostate cancer. *Cell, 163, 1011–1025.*

3.FRALEY, C. and RAFTERY, A.E. (2002): Model-based clustering, discriminant analysis, and density estimation. *JASA, 97, 458, 611–631*

On detecting the change-points in piecewise regression models

Dimitra Eleftheriou[1]

National and Kapodistrian University of Athens deleftheriou@med.uoa.gr

Abstract. In certain circumstances the hazard rate of patients is constant across time but in real life it is more likely to vary over different intervals giving rise to the piecewise exponential and piecewise Weibull models, respectively. One of the major problems in such piecewise models is to determine the points of change of the hazard rate. From the practical point of view this can provide very important information as it may reflect changes in the progress of a disease. The proposed project refers to piecewise regression models with covariates and in particular on methods to identify the change points. Both cases of known number of points and the challenging of unknown have been examined. An example based on herpes zoster data set has been used to demonstrate the developed methodology.

Keywords

Piecewise Regression Modelling, Herpes Zoster, Bootstrap, Simulation, Simulated Annealing, Survival Analysis, Model Selection.

Predicting Genetic Predisposition to Treatment Responsiveness in Randomized Clinical Trials Using Bayesian Whole Genome Regression

Bahar Erar[1] and George D. Papandonatos[1]

Department of Biostatistics, Brown University, Providence, RI, USA

Abstract. A critical aspect of personalized medicine is the development of methods that evaluate genetic predisposition to respond to a treatment. Studies have shown that a small number of genes identified by one-at-a-time testing methods are shown to be inadequate for prediction modeling. Whole genome prediction (WGP) methods have been shown to improve predictive accuracy in complex traits. However, the methodology is not tailored for outcome prediction in human randomized clinical trials (RCTs). We propose a Bayesian WGP method that accounts for the underlying genetic heterogeneity present in populations often targeted in RCTs using a mixed model approach. Under this model, small effects regulated by the treatment that would normally go undetected and disregarded are captured by the unconstrained covariance structure of the genetic random effects. We employ an efficient estimation approach that allows application to large datasets at a reasonable computational cost. Predictive accuracy is evaluated in comparison to existing methods using simulated phenotypes generated from real genotypes under various scenarios. Results demonstrate that the BWGP approach adapted to RCTs performs better or at least as well as stratified application of existing methods, such as BayesC, Bayesian Ridge Regression and Bayesian LASSO. The gain in prediction accuracy is highest for moderately and highly heritable traits under realistic effect regulation scenarios. Finally, we demonstrate how this approach can be integrated into the treatment decision process using data from a real-life behavioral weight loss trial.

Keywords

WHOLE GENOME PREDICTION, CLINICAL TRIALS, BAYESIAN REGRESSION, POPULATION GENETICS

Bayesian imputation of time-varying covariates in linear mixed models

Nicole S. Erler[1,2,*], Dimitris Rizopoulos[1,2], Vincent W.V. Jaddoe[2,3,4], Oscar H. Franco[2] and Emmanuel M.E.H. Lesaffre[1,5]

[1] Department of Biostatistics, Erasmus MC, Wytemaweg 80, 3015CN Rotterdam, The Netherlands (* corresponding author, n.erler@erasmusmc.nl)
[2] Department of Epidemiology, Erasmus MC, Wytemaweg 80, 3015CN Rotterdam, The Netherlands
[3] Department of Pediatrics, Erasmus MC, Wytemaweg 80, 3015CN Rotterdam, The Netherlands
[4] Generation R Study Group, Erasmus MC, University Medical Center, Rotterdam, The Netherlands
[5] L-Biostat, KU Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium

Abstract. Studies involving large observational datasets commonly face the challenge of dealing with multiple missing values. The most popular approach to overcome this challenge is multiple imputation using chained equations. However, it has been shown to be suboptimal in complex settings, specifically in settings with longitudinal outcomes, which cannot be easily and adequately included in the imputation models. Bayesian methods avoid this difficulty by specification of a joint distribution and thus offer an alternative. A popular choice for that joint distribution is the multivariate normal distribution. In more complicated settings, as in our two motivating examples that involve time-varying covariates, additional issues require consideration: the endo- or exogeneity of the covariate and the functional form of the association with the outcome. In such situations, the implied assumptions of standard methods may be violated, resulting in bias. In this work, we extend and study a more flexible, Bayesian, alternative to the multivariate normal approach, to better handle complex incomplete longitudinal data. We discuss and compare assumptions of the two Bayesian approaches about the endo- or exogeneity of the covariates and the functional form of the association with the outcome, and illustrate and evaluate consequences of violations of those assumptions using simulation studies and two real data examples.

Keywords

BAYESIAN, EPIDEMIOLOGY, IMPUTATION, MISSING COVARIATE VALUES, TIME-VARYING COVARIATES

Simulation of multi-pollutant model results in the presence of measurement error

Dimitris Evangelopoulos[1], Ruth Keogh[2], Klea Katsouyanni[1], and Heather Walton[1]

[1] Environmental Research Group and NIHR HPRU in Heath Impact of Environmental Hazards, King's College London, London, U.K. SE1 9NH

[2] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, U.K.

Abstract. Air pollution is a major public health concern. Pollutants investigated are correlated in space or time, as they are determined by common sources. Policy makers are often interested in pollutants' independent effects on health. In this context, multi-pollutant models are very common in air pollution studies. However, effect estimates can be subject to bias due to measurement error. Our goal is to estimate the size of this bias for $PM_{2.5}$ and NO_2.

A systematic review and meta-analysis has provided plausible values for measurement error and possible sources of heterogeneity. These will be used as simulation input variables. Also, we created error-prone exposures of both Classical and Berkson type error based on logical assumptions using proof-of-concept simulations. We illustrate the hypothetical effects of measurement error on model estimates using different correction formulas (Regression Calibration and SIMEX) and simulations.

Preliminary results indicate heterogeneity in the differences between exposures based on the study. Regarding the simulations results, we confirm the findings from the literature. Assuming some true effects for the pollutants, we get biased estimates for all pollutants and effect transfer from poorly measured to better measured ones when using error-prone variables. We will update our results with better informed input variables, in order to get closer to the true independent effects of the pollutants.

Simulations can lead to the quantification of the consequences of measurement error and adjusting for it can result in better model estimates. It may be inferred that certain potential interpretations are more unlikely than others.

Keywords

AIR POLLUTION, MEASUREMENT ERROR, SIMULATIONS

References

CARROLL, RJ., et al. (2006): *Measurement error in nonlinear models: a modern perspective.* CRC press.

Extent, Duration and Predictors of Exclusive Breastfeeding in a Longitudinal Study: Adjusting for missing data using an Accelerated Failure Time model

Samah Hayek[1], Havi Murad[2], Anneke Ifrah[1], Tamy Shohat[1], and Laurence Freedmant[3]

[1] University of Haifa, Israel center for Disease Control,Ministry of Health, Israel. samah.hayek@moh.health. gov.il , Anneke.Ifrah@moh.health.gov.il , Tamy.shohat@moh.health.gov.il

[2] Biostatistics Unit, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Ramat-Gan, Israel HaviM@gertner.health.gov.il

[3] Sheba Medical Center, Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, 52161, Israel lsf@actcom.co.il

Abstract. Background The World Health Organization (WHO) recommends at least 6 months of exclusive breastfeeding (EBF). Longitudinal epidemiological studies facilitate estimation of the duration of EBF, but often suffer fromloss to follow-up and missing information. Objectives While adjusting for missing data, (1)To estimate the proportion of Israeli women who practice EBF. (2) To estimate the distribution of duration of EBF.(3) To identify factors that predict the duration of EBF. Methods: A longitudinal study was carried out including all women who gave birth between September 2009 and February 2010 in selected Israeli hospitals (N=2119). Participantsreported information related to EBF, sociodemographic characteristics, and breastfeeding practices in the hospital and at two-monthly intervals thereafter. Information onEBF status and duration was missing for 35% of women. We imputed EBF practice using logistic regression Multiple Imputation (MI) method with 20 repeats (procedure MI with option FCS, SAS 9.4) and using Rubins rule estimated the probability of practicing EBF. Predicted probabilities of practicing EBF for women with missing information served as weights in the analyses of objectives 2-3.We imputed EBF duration based on an Accelerated Failure Time (AFT) model built on observed duration times, creating five complete data sets. We then estimated the distribution of duration in those practicing EBF using a weighted Kaplan-Meier curve (SAS 9.4) in each completed dataset and used Rubins rule to estimate the time of EBF survival curve and its standard errors. Results: 1: The observed proportion of women practicing EBF (complete case analysis) was 69% (95%CI; 66%-71%).After imputation, the estimated proportion changed to 65% (95%CI; 62%-68%). 2: After imputation, estimated percentiles the time of EBF among women practicing EBF were: 25%:3.0m; 50%: 4.0m; 75%:5.7m. 3: Predictors of EBF duration were: stated intention to BF - 50% increase (p=0.001); religious observance (secular vs. ultra-orthodox) - 22% decrease ($p < 0.001$); giving formula milk in hospital - 11% decrease in EBF duration ($p < 0.001$); using a pacifier in hospital - 10% decrease($p < 0.001$); ethnicity (Arab v Jew) -9% decrease (p=0.06). Conclusions: By imputing missing practice and duration of EBF we obtained estimated proportion and duration of EBF adjusted for the potential bias caused by missing information. Using an AFT model for EBF duration also allows direct interpretation of the impact of various factors on EBF duration.

Generalized Chao estimators with external information and measurement error

Alessio Farcomeni and Francesco Dotto

Sapienza - University of Rome alessio.farcomeni@uniroma1.it, francesco.dotto@uniroma1.it

Abstract. In population size estimation Chao estimator is widely used for its simplicity and the fact that it asymptotically guarantees a meaningful lower bound. Building on the work of Böhning *et al.* (2013) and Farcomeni (2017) we present a generalized Chao (GC) estimator based on a subject-occasion-specific design matrix. A conditional formulation is used to accommodate behavioural effects. We then extend the GC estimator to (i) external information, in the form of non-linear constraints on subpopulation sizes and (ii) measurement error. For the first, we propose a reparameterization of the estimating equations. As a result, the constrained MLE can be found with no additional computational efforts. For the second we generalize SIMEX procedure to multiple measurement methods. In simulation we show that (even incorrect) external information can substantially decrease the MSE. We illustrate with an application to a whale shark (*Rhincodon typus*) population, where mostly jouvenile males are observed. We use external information on gender ratio of whale sharks to correct for low catchability of females, and our multivariate SIMEX procedure to correct for measurement error in assessment of shark length. The resulting population size estimates are about 60% larger than the unconstrained-uncorrected counterparts. A sensitivity analysis confirms these findings.

Keywords

capture-recapture, SIMEX, constrained inference

References

BÖHNING, D., VIDAL-DIEZ, A., LERDSUWANSRI, R., VIWATWONGKASEM, C. and ARNOLD, M.A. (2013): A generalization of Chao's estimator for covariate information. *Biometrics, 69, 1033–1042.*

FARCOMENI, A. (2017): Fully general Chao and Zelterman estimators with application to a Whale Shark population. *Journal of the Royal Statistical Society (Series C), in press.*

Bayesian Permutation Tests

Livio Finos[1] and Florian Klinglmueller[2]

[1] University of Padua, livio.finos@unipd.it
[2] Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Austria florian.klinglmueller@meduniwien.ac.at

Abstract. A "Bayesian" approach to permutation inference

Permutation tests have many appealing properties - i.e. exact control of the Type I error rate, asymptotic optimality, consistency - while making few distributional assumptions. Especially, for multivariate problems where distributional assumptions can quickly get unwieldy, permutation tests provide a flexible and powerful framework for statistical inference. Consequently, permutation tests are frequently applied in clinical research, biology, neuroscience, and genomics among others.

An open problem in permutation inference is the choice of test statistic, which has a strong implication on the operating characteristics of the test under the alternative hypothesis. In this work we propose a permutation approach that allows for the inclusion of prior information by choosing the test statistic on the basis of the Bayesian posterior probability of the parameter under test. The use of prior knowledge enhances the power of the test in a pre-specified region of the (possibly multivariate) alternative hypothesis. The proposed tests retain all (frequentist) properties of permutation tests mentioned above, even under mispecification of the prior model.

The proposed approach shows its efficacy also in high dimensional settings (e.g. neuroimaging, OMICs). If prior knowledge about proportion of alternative hypotheses is available, or the experiment is expected to produce mostly effects of the same direction, the proposed tests greatly increase the power over conventional permutation tests.

Keywords

PERMUTATION TESTS, BAYESIAN STATISTICS, EXACT INFERENCE, MULTIVARIATE STATISTICS

References

PESARIN, F. (2001) *Multivariate Permutation Tests : With Applications in Biostatistics.* Wiley, New York.

From GWAS to personalized disease risk prediction: statistical aspects of genetic risk score development and validation

Krista Fischer[1], Kristi Läll[12] and Sulev Reisberg[345]

[1] Estonian Genome Center, University of Tartu, krista.fischer@ut.ee
[2] Institute of Mathematics and Statistics, University of Tartu
[3] Institute of Computer Science, University of Tartu
[4] Software Technology and Applications Competence Centre, Tartu, Estonia

Abstract. We will discuss some statistical issues encountered in the process of development and validation of Genetic Risk Scores (GRS), using simulations as well as the data of the Estonian Biobank. Usually the GRS is defined as a linear combination of effect allele counts of several Single Nucleotide Polymorphisms (SNPs), whereas the SNPs and their corresponding weights are based on results of a large-scale meta-analysis of Genome-Wide Association Study (GWAS). The long-term purpose of GRS development is to use them in risk prediction algorithms for complex diseases, to improve the risk stratification in general practice.

First, we demonstrate that a GRS that is based on a large number of SNPs that are weighted in an optimal manner, provides better predictive accuracy than either a GRS that only combines the most significant SNPs or a GRS that is calculated as unweighted sum of the risk alleles or uses regression coefficients as weights.

Second, we discuss the aspects of study design and sample selection when developing GRS-s. As large GWAS meta-analyses are often based on all available genotyped cohorts, it is possible that the validation cohort has been included in the discovery study. We demonstrate that even if the cohort forms no more than 1-2% of the total meta-analysis sample, one could get dramatically misleading conclusions while validating the GRS. Also, one ideally needs two validation samples – one that is used to compare different versions of the GRS and select the optimal one, and another one for final validation of the GRS. We discuss options for the optimal sample selection and compare alternative scenarios.

Third range of problems is associated with ethnic origin of the samples. We show that cohorts of different ethnicity could lead to markedly different risk score distribution and propose some ideas to account for ethnic diversity, when implementing the personalized risk prediction in practice.

Keywords

GENETIC EPIDEMIOLOGY, PERSONALIZED MEDICINE, GENETIC RISK SCORES

Leave-one-out crossvalidation favors inaccurate estimators

Angelika Geroldinger[1], Lara Lusa[2], Mariana Nold[3], and Georg Heinze[1]

[1] Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Austria angelika.geroldinger@meduniwien.ac.at

[2] Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Slovenia

[3] Institute of Medical Statistics, Computer Sciences and Documentation, University Hospital Jena, Germany

Abstract. The c-statistic is a widely used measure to quantify the discrimination ability of logistic and Cox regression models. For a binary outcome it is simply the proportion of all pairs of observations with opposite outcomes which are correctly ranked by the model. Clearly, calculating the c-statistic for the data on which the model was built will often give too optimistic results, especially in the situation of small samples or rare events. Data sampling techniques such as crossvalidation or bootstrap are frequently used to correct for this over-optimism. Leave-one-out (LOO) crossvalidation has the advantage of being applicable even with small samples where 10-fold crossvalidation might not be feasible. However, mostly in the machine learning community, the accuracy of LOO crossvalidated c-indices is under debate since it was shown that they can be severely biased towards 0. We discuss these results and demonstrate by simulations that the bias in LOO crossvalidated c-indices depends strongly on the estimation method. For instance, the negative bias in LOO crossvalidated c-indices was much stronger for ridge regression than for maximum likelihood estimation. Our simulations indicate that leave-pair-out crossvalidation, a method proposed as alternative to LOO crossvalidation, might be a better choice. Finally, we compare these methods using data from a study on arterial closure devices in minimally invasive cardiac surgery.

This work was supported by the Austrian Science Fund (FWF) within project I 2276.

Keywords

AUC, LOGISTIC REGRESSION, RIDGE REGRESSION

Conditional frailty Marshall-Olkin survival model for bivariate censored failure time data

Giussani Andrea[1] and Bonetti Marco[2]

[1] Bocconi University, Department of Decision Sciences, Milan (Italy) andrea.giussani@unibocconi.it
[2] Bocconi University, Department of Policy Analysis and Public Management, Milan (Italy) marco.bonetti@unibocconi.it

Abstract. The aim of this paper is to develop a new frailty survival model for examining the association between paired failure times under the presence of right-censoring. To take into account the correlation between these measurements, the well-known Marshall-Olkin Bivariate Exponential Distribution (MOBVE) is considered for the joint distribution of frailties. The reason is twofold: on the one hand, it allows to model shocks that affect individual-specific frailties; on the other hand, the parameter underlying the Poisson process describing the common shock completely captures the dependence between the pair of lifetimes (T_1, T_2). The proposed methodology is then applied to the investigation of association in disease-free different-sex couples from the Cache County Study on Memory Health and Aging (CCSMHA) data with respect to death.

Keywords

FRAILTY MODELS, MARSHALL-OLKIN DISTRIBUTION, CUMULATIVE HAZARD ORDERING, CACHE COUNTY STUDY

References

AALEN, O.,BORGAN, O. and GJESSING, H. (2007). *Survival and Event History Analysis: a process point of view.* Springer.

MARSHALL, A.W. and OLKIN, I. (1988). Families of Multivariate Distributions.*Journal of American Statistical Association, 83, 834-841.*

OAKES, D. (1982). A model for association in bivariate survival data.*J.R. Stat. Soc., 44, 414-422.*

SHAKED, M. and SHANTIKUMAR, J. (2007). *Stochastic Orders.* Springer.

TSCHANZ, J.T., NORTON, M., ZANDI, P. and LYKETSOS, M.D. (2013). The Cache County Study on Memory in Aging: Factors Affecting Risk of Alzheimer's disease and its Progression after Onset.*Int Rev Psychiatry, 25, 673-685.*

EMcorrProbit R package

Denitsa Grigorova[1] and Nina Daskalova[2]

[1] Sofia University “St. Kliment Ohridski” dgrigorova@fmi.uni-sofia.bg
[2] Sofia University “St. Kliment Ohridski” ninad@fmi.uni-sofia.bg

Abstract. Correlated probit models (CPMs) are widely used for modeling of ordinal data or joint analyses of ordinal and continuous data which are common outcomes in medical studies. When we have clustered or longitudinal data CPMs with random effects are used to take into account the dependence between clustered measurements. When the dimension of the random effects is large, finding of the maximum likelihood estimates (MLEs) of the model parameters via standard numerical approximations is computationally cumbersome or in some cases impossible. EM algorithms for one ordinal longitudinal variable [**?**] and for one ordinal and one continuous longitudinal variable [**?**] are recently developed. The methods developed set the foundations of the EMcorrProbit R package (https://github.com/ninard/EMcorrProbit) which is going to offer also MLEs of CPM for two longitudinal ordinal variables via recently developed ECM algorithm. An application of the algorithm is presented to CPM for the longitudinal ordinal outcomes self-rated health and categorized body mass index from the Health and Retirement Study. We will report results from fitting the model and also some simulation studies.

Keywords

correlated probit model, EM algorithm, random effects, R package

Measuring Inequality from Incomplete Income and Survival Data

Long Hong[1,2], Guido Alfani[1,2], Chiara Gigliarano[2,3], and Marco Bonetti[1,2]

[1] Bocconi University, long.hong@studbocconi.it
[2] Dondena Centre for Research on Social Dynamics and Public Policy
[3] University of Insubria

Abstract. Quite often, observed income and survival data are incomplete due to left- or right- censoring or truncation. Measuring inequality, for instance by the Gini index of concentration, from such incomplete data, can produce biased results. This paper moves in three directions. First, we use a test statistic for the comparison of two (survival) distributions based on the non-parametric restricted Gini index, using both asymptotic and permutation inference. Second, we develop non-parametric bounds for the unrestricted Gini index from censored data. Finally, we apply maximum likelihood estimation for three commonly used parametric models to estimate the unrestricted Gini Index, both from censored and truncated data. We have developed Stata functions that implement these approaches.

Keywords

Gini Index, Censored Data, Survival Analysis

References

Bonetti, M. Gigliarano, C. and Muliere, P. (2009): The Gini Concentration Test for Survival Data. *Lifetime Data Analysis, 453(15).*

Gigliarano, C., Basellini, U., and Bonetti, M. (2016): Longevity and Concentration in Survival Time: The log-scale-location Family of Failure Time Models. *Lifetime Data Analysis, 10(2):1-21.*

Gigliarano, C. and Bonetti, M. (2013): The Gini Test for Survival Data in Presence of Small and Unbalanced Groups. *EBPH Epidemiology, Biostatistics and Public Health, 10(2).*

Pawitan, Y. (2001): *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* New York, NY: Oxford University Press.

Yitzhaki, S. and Schechtman, E. (2013): *The Gini Methodology.* New York, NY: Springer.

A Score Test for Over-Dispersion in Marginalized Zero-Inflated Poisson Regression Models

Gul Inan[1], John Preisser[2], and Kalyan Das[3]

[1] Department of Statistics, Middle East Technical University, Turkey ginan@metu.edu.tr,
[2] Department of Biostatistics, University of North Carolina at Chapel Hill, U.S.A. jpreisse@bios.unc.edu,
[3] Department of Statistics, University of Calcutta, India kalyanstat@gmail.com

Abstract. Long et al. (2014) and Preisser et al. (2016) have recently proposed marginalized zero-inflated Poisson (MZIP) regression models and marginalized zero-inflated negative binomial (MZINB) regression models, respectively, for analysis of zero-inflated count data with population-based inferences. Motivated by Ridout et al. (2001), this study proposes a score test for testing a MZIP regression model against a MZINB regression model to investigate whether the zero-inflated count data can be better represented via MZIP regression or MZINB regression due to possible over-dispersion in zero-inflated count data sets. The sampling distribution and empirical power of the proposed score test are investigated via a Monte Carlo simulation study and the procedure is illustrated by a horticultural data set.

Keywords

COUNT DATA, EXCESS ZEROS, MARGINAL MODELS, OVER-DISPERSION

References

LONG, D.L., PREISSER, J.S., HERRING, A.H. and GOLIN, C.E. (2014): A Marginalized Zero-Inflated Poisson Regression Model with Overall Exposure Effects. *Statistics in Medicine, 33, 5151–5165.*

PREISSER, J.S., DAS, K., LONG, D.L. and DIVARIS, K. (2016): Marginalized Zero-Inflated Negative Binomial Regression with Application to Dental Caries. *Statistics in Medicine, 35, 1722–1735.*

RIDOUT, M., HINDE, J. and DEMÉATRIO, C.G. (2001): A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Binomial Alternatives. *Biometrics, 57, 219–223.*

A Quantile Regression Model for Failure Time Data with Time Dependent Covariates

Malka Gorfine[1], Yair Goldberg[2], and Ya'acov Ritov[3,4]

[1] University of Tel-Aviv, Ramat Aviv, 6997801 Tel Aviv, Israel gorfinem@post.tau.ac.il
[2] University of Haifa, Mount Carmel, 31905 Haifa, Israel, ygoldberg@stat.haifa.ac.il
[3] The Hebrew University of Jerusalem, Mount Scopus, 91905 Jerusalem, Israel,
[4] Department of Statistics, University of Michigan Ann Arbor, MI 48194, USA
yaacov@mscc.huji.ac.il

Abstract. Since survival data occur over time, often important covariates we wish to consider also change over time. Such covariates are referred as time-dependent covariates. Quantile regression offers a flexible survival data modeling by allowing the covariates to vary with quantiles. In this talk, I will present a novel quantile regression model accommodating time-dependent covariates, for analysing survival data subject to right censoring. The simple estimation technique assumes the existence of instrumental variables. In addition, I will present a doubly-robust estimator in the sense of Robins & Rotnitzky (1992). The utility of the proposed methodology will be demonstrated using the Stanford heart transplant dataset

Keywords

Survival analysis, quantile regression, time dependent covariates

References

ROBINS, J. M. AND ROTNITZKY, A. (1992). *Recovery of information and adjustment for dependent censoring using surrogate markers.* In: Jewell, N. P., Dietz, K. and Farewell, V. T. (Eds.), *AIDS Epidemiology.* Birkhuser Boston, 297-331.

Adapting censored regression methods to adjust for the limit of detection in the calibration of diagnostic rules for clinical mass spectrometry proteomic data

Alexia Kakourou[1], Werner Vach[2], and Bart Mertens[1]

[1] Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands a.a.kakourou@lumc.nl

[2] Center for Medical Biometry and Medical Informatics, University of Freiburg, Freiburg, Germany

Abstract. In this work, we consider the problem of calibrating diagnostic rules based on high-resolution mass-spectrometry (MS) data subject to the limit of detection (LOD). The LOD is related to the limitation of instruments in measuring low-concentration proteins. As a consequence, peak intensities below the LOD are often reported as missings. We propose the use of censored data methodology to handle spectral measurements within the presence of LOD, recognizing that those have been left-censored for low-abundance proteins. We replace the set of incomplete spectral measurements with estimates of the expected intensity and use those as input to a prediction model. To correct for lack of information and measurement uncertainty, we combine this approach with borrowing of information through the addition of an individual-specific random effect formulation. We present different modalities of using the above formulation for prediction purposes and show how it may also allow for variable selection. We evaluate the proposed methods by comparing their predictive performance with the one achieved using the complete information as well as alternative methods to deal with the LOD.

Keywords

clinical proteomics, limit of detection, censored regression, prediction

References

Kakourou, A. and Vach, W. and Mertens, B (2016): Adapting censored regression methods to adjust for the limit of detection in the calibration of diagnostic rules for clinical mass spectrometry proteomic data. *Medical Methods in Medical Research*, DOI:10.1177/0962280216685742.

Targeting Disease Signatures Towards Precision Healthcare

Mira Markus Kalish[1], Alexis Mitelpunkt[1], Tal Kozlovski[1],, Tal Galili[1], Anat Mirelman[1], and Yoav Benjamini[1]

Department of Statistics and Operation Research, Tel-Aviv University, Ramat-Aviv, 6997801 Tel-Aviv miram@post.tau.ac.il

Abstract. The premise of personalized and precision medicine depends on the ability to define the broad, comprehensive and reliable signature of a disease.

Thus, our target is a comprehensive definition of Disease Signature that relates to all relevant personal micro and macro environmental features, physical, mental, cultural and environmental of the human body functioning in his surroundings. It might include parameters such as: age, gender, clinical tests, biological markers, medical history, genetics, imaging, lifestyle, physical and sociological environment, etc. New advanced technologies have greatly enhanced our ability to capture, analyse and translate these parameters. Still, integration knowledge and data from different domains encounters many barriers, for example, dealing with various sets of data originating from different sources, missing values, privacy, security and the special concern of dealing with the lack of consistency in the final diagnosis. Involving and combining various hospital data creates additional barriers of concepts, language, modes of treatments, missing values, etc. A Parkinsons disease hospital cohort is one case study we have analysed, including diagnosed patients and their family members, and containing genetic, cognitive and environmental measures. A pre- processing crucial stage in the Parkinson case and an imputation scheme was developed for addressing missing values while segregating missing at random from missing not at random cases and relating to the characteristics of the observations. Statistical analysis could commence only after these important stages of data cleaning. As a second stage the large database is screened while controlling the average false proportion rate over the selected families using the Benjamini and Bogomolov (2013) proposal for multiple testing of families. This yielded new discoveries regarding the association between genotypes and Parkinsons disease clinical data as well as guarantees for replicable results, in spite of the fact that they were discovered after intensive search. Going beyond the discovery of associations A 3C- Categorization, Classification and Clustering- strategy, was developed, as part of the Medical Informatics efforts in the Human Brain Flagship Project. It was applied to the above described Parkinsons disease cohort and to the Alzheimers disease Neuroimaging Initiative (ADNI) cohort. The 3C approach is a stepwise process, based on supervised and unsupervised algorithms, incorporating medical expert knowledge in a structured way into the analysis process of the disease manifestations and potential biomarkers. The preliminary 3C study applied to the ADNI cohort suggests, new sub-classes, with clinical and biomarker characteristics different from those assigned in the ADNI database. We therefore believe it has the potential to move us, toward personalized reliable prediction and treatment.

Adjusting for unit nonresponse in EMENO study: inverse probability weighting and multiple imputation methods

Natasa Kalpourtzi and Giota Touloumi

Department of Hygiene, Epidemiology & Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece

Abstract. The "National Study of Morbidity and Risk factors" (EMENO) is a health examination survey, which took place in Greece between 2014 and 2015. Multistage stratified random sampling based on 2011 census was applied to select the sample. Appropriate weighting which takes into account the complex design, with additional adjustment for noncoverage by using sociodemographic characteristics based on census data was used to analyze the data. These methods provide unbiased estimates provided data are fully observed. However, nonresponse, inevitably, exists and can lead to biased estimates, when responders differ from non-responders. In EMENO, all participants filled in a questionnaire but some of them, upon their consent, also provided blood samples. We found that those who provided blood samples differ substantially from those who did not. Thus, generalizing results derived from the examination subsample to the whole population may lead to biased estimates. To adjust for nonresponse bias we considered two methods: a)the inverse probability weighting method (IPW) in which each observation of the subsample was weighted by the inverse of the probability of being included conditionally on covariates and b)multiple imputation by chained equations method (MICE) which is a widely-used method for filling missing values, iteratively, by using a sequence of univariate imputation methods with fully conditional specification (FCS) of prediction equations. MICE is a valid method providing the missigness mechanism is "missing at random" (MAR). We applied these methods to estimate the prevalence of elevated cholesterol levels in the Greek adult (≥18 years) population (total cholesterol≥240mg/ml). The prevalence, adjusting only for study design and non-coverage, was 15.54 (95% C.I: 14.45-16.70). Incorporating also the IPW method, the corresponding estimation was 14.90% (95% C.I: 13.81-16.06), similar to the one obtained after applying the MICE method: 14.91 (95% C.I: 13.89-15.93). Not taking into account nonresponse, led to overastimated prevalence of adults with elevated cholesterol levels. This is mainly attributed to the subasmple participants' older age and larger bmi, both associated with increased levels of cholesterol. In conclusion, when nonresponse exists, methods accounting for it should be applied to avoid biased estimates.

Keywords

inverse probability weighting, multiple imputation, survey nonresponse

Multiple Factor Analysis with Its Basic Properties and An Application

Siddik Keskin[1]

Uzuncu Yil University, Faculty of Medicine, Department of Biostatistics, Van-Turkey
skeskin973@hotmail.com

Abstract. In this study, Multiple factor analysis (MFA) is introduced with basic properties and geometrical viewpoint. In addition, an application is performed to help understanding of the subject. MFA is used to analyze relationships among the variables or characteristic in several tables. This analysis seeks the common structures among the variables and allows us to analyze both categorical and quantitative variables together. MFA is used in many areas such as sensory evaluation, medical research, economy, ecology, and chemistry. MFA is carried out in two steps. During the first step, Principal component analysis is performed on each set of data. Then the groups are normalized by dividing all the variables within a group by the first eigenvalue for that group. All the variables are then combined into a single data set and a global Principal component analysis is performed. In the analysis, the number of variables in each group may differ and the type of the variables (nominal or quantitative) can vary from one group to the other. The analysis generates an integrated configuration to present the relationships among the variables in two dimensional spaces.As compared to other alternative methods that can be used, this method is partially simple for interpretation of the resultand often preferable in the analysis of multiple tables.

Keywords

CONFIGURATION, CONSENSUS, EIGENVALUES NORMALIZATION, SENSORY

The Use of Joint Modeling Approach in Personalized Medicine

Naime Meric Konar[1], Eda Karaismailoglu[2], and Ergun Karaagaoglu[1]

[1] Hacettepe University, Faculty of Medicine, Department of Biostatistics, Ankara, Turkey
nmeric.konar@hacettepe.edu.tr

[2] Kastamonu University, Faculty of Medicine, Department of Biostatistics, Kastamonu, Turkey

Abstract. Personalized Medicine aims making decisions such as diagnosing, initiating treatment...Clinicians would like to have a prognostic tool to make decisions about patients, by examining the changes of measurements over time. Joint modeling approach is now being used in personalized medicine area for these aims as a prognostic tool. The aim of this study is to show the usage of joint modeling approach in this field. The main reason for utilizing this approach in the field of personalized medicine is dynamic predictions. As long as new measurements are taken, patients´ survival probabilities and longitudinal predictions are updated; and it gives predictions dynamic characteristic. Especially in the longitudinal part,given that $u>t$; it is possible to have predictions at time u by using measurements taken up to time t.This property makes joint models much more useful while making decisions. To show how joint modeling can be used to obtain these dynamic predictions, we used the data, which was collected retrospectively from Hacettepe University Emergency Department records. Data set includes repeated Troponin-I measurements. The follow-up time is planned as 240 hours. In conclusion, it is possible to use joint modeling approach in personalized medicine area as prognostic tool.By examining subject-specific longitudinal profile; a clinician can make decisions more accurately.Therefore, it could be possible to make an early intervention.

Keywords

Joint Modelling, Longitudinal Data, Diagnostic Tests

References

1. Rizopoulos, D. (2011). Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. Biometrics 67, 819:829.
2. Rizopoulos, D. (2010). JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. Journal of Statistical Software, 35:9.

A Modeling Approach for Predicting Disease Status Using Functional Data in the Absence of a Gold Standard

Amita Manatunga[1], Qi Long[2] and Andrew T. Taylor[3]

[1] Department of Biostatistics and Bioinformatics, Emory University, 1518 Clifton Road NE, Atlanta, Georgia 30322, U.S.A. amanatu@emory.edu
[2] Department of Biostatistics and Epidemiology, University of Pennsylvania, 3400 Civic Center Blvd, Philadelphia, Pennsylvania 19104, U.S.A.
[3] Department of Radiology and Imaging Sciences, Emory University, 1364 Clifton Road NE, Atlanta, Georgia 30322, U.S.A.

Abstract. Data from clinical studies involving risk predictions provide a wealth of opportunities for statistical research in particular to the development of prediction models and their evaluations. I will introduce a specific clinical decision making problem in nuclear medicine, present its statistical challenges and discuss some potential solutions. In our study, two consecutive curves over time are observed per subject and in some cases, the second curve for some subjects are not observed. There is no gold standard for determining disease status, instead, the ratings for disease status from multiple experts are available We consider a latent class modeling approach for predicting disease status of a subject based on observed functional data and its ratings from multiple experts. I will present our work including the modeling procedure, prediction models consisting of several prediction schemes, and their evaluation via simulation studies. I will demonstrate the practicality of our method and will show that proposed modeling procedure reasonably captures the patterns of observed curves and provide sensible clinical interpretations. I will conclude with a brief discussion of future work.

Keywords

GOLD STANDARD, LATENT CLASS MODELS, PREDICTION

References

ALBERT, P. S., MCSHANE, L. M., and SHIH, J. H. (2001): Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics, 57, 610–619.*

BAO, J., MANATUNGA, A., BINONGO, J. N. G., and TAYLOR, A. T. (2011): Key variables for interpreting 99mTc-mercaptoacetyltriglycine diuretic scans: development and validation of a predictive modeL. *AJR. American journal of roentgenology , 197, 325.*

Cox Regression for Doubly Truncated Data

Micha Mandel[1], Jacobo de Uña-Álvarez[2], David K. Simon[3], and Rebecca A. Betensky[4]

[1] Department of Statistics, The Hebrew University of Jerusalem, Jerusalem 91905, Israel msmic@huji.ac.il
[2] Department of Statistics and OR, University of Vigo, Vigo 36310, Spain jacobo@uvigo.es
[3] Department of Neurology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA 02115, USA dsimon1@bidmc.harvard.edu
[4] Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA betensky@hsph.harvard.edu

Abstract. Doubly truncated data arise when event times are observed only if they fall within subject-specific, possibly random, intervals. While non-parametric methods for survivor function estimation using doubly truncated data have been intensively studied, only a few methods for fitting regression models have been suggested, and only for a limited number of covariates. In this paper, we present a method to fit the Cox regression model to doubly truncated data with multiple discrete and continuous covariates, and describe how to implement it using existing software. The approach is used to study the association between candidate single nucleotide polymorphisms and age of onset of Parkinson's disease.

Keywords

BIASED DATA, INVERSE WEIGHTING, RIGHT TRUNCATION, U STATISTIC

Optimal Sampling Designs of Two-Compartment Nonlinear Regression Models

Noa Molshatski[1] and Sandrah P. Eckel[1]

Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA molshatz@usc.edu

Abstract. The fractional concentration of exhaled nitric oxide (FeNO) is a noninvasive biomarker of airway inflammation increasingly assessed in clinical, occupational and environmental epidemiology studies. At low flow rates FeNO originates primarily from the bronchial airway compartment and at higher flow rates from the alveolar compartment. Repeat FeNO maneuvers at multiple fixed exhalation flow rates (extended NO analysis) can be used to estimate parameters quantifying proximal and distal sources of NO in mathematical models of lower respiratory tract NO. Despite the growing number of multiple flow FeNO studies, there is no official standard flow rate sampling protocol. In this work, we provide information for study planning by deriving theoretically optimal flow rate sampling designs.

First, we reviewed previously published designs. Then, under a nonlinear regression framework for estimating NO parameters in the steady-state two compartment model of NO, we identified unbiased optimal four flow rate designs using theoretical derivations of the Fisher Information matrix and simulation studies. Optimality criteria included NO parameter standard errors (SEs). A simulation study was used to estimate sample sizes required to detect associations with NO parameters estimated from studies with different designs.

We found that most designs (77%) were unbiased. NO parameter SEs were smaller for designs with: more target flows, more replicate maneuvers per target flow, and a larger range of target flows. High flows were most important for estimating alveolar NO concentration, while low flows were most important for the proximal NO parameters.

In conclusion, there is a class of reasonable flow rate sampling designs with good theoretical performance. In practice, designs should be selected to balance the tradeoffs between optimality and feasibility of the flow range and total number of maneuvers.

Keywords

AIR POLLUTION, BIOMARKERS, NONLINEAR REGRESSION, FISHER INFORMATION MATRIX, STUDY DESIGN

Refining prognostication in CLL

Theodoros Moysiadis[1], Panagiotis Baliakas[2], Davide Rossi[3], Achilles Anagnostopoulos[4], Jonathan C. Strefford[5], Sarka Pospisilova[6], David Oscier[7], Gianluca Gaidano[5], Elias Campo[8], Paolo Ghia[9], Richard Rosenquist[2], and Kostas Stamatopoulos[1,2]

[1] Inst. of Applied Biosciences, Thes., Greece moysiadis.theodoros@certh.gr
[2] Department of IGP, Uppsala, Sweden panagiotis.baliakas@igp.uu.se
[3] Department of Translational Medicine, Univ. of Eastern Piedmont, Novara, Italy
[4] Hematology Dep. and HCT Unit, G. Papanicolaou Hospital, Thes., Greece
[5] Cancer Sciences, Faculty of Medicine, Univ. of Southampton, Southampton, UK
[6] CE Inst. of Technology, Masaryk Univ. and Univ. Hospital Brno, Czech Republic
[7] Department of Haematology, Royal Bournemouth Hospital, Bournemouth, UK
[8] Department of Pathology, University of Barcelona, Spain
[9] UniversitĂ Vita-Salute San Raffaele, Milan, Italy

Abstract. Chronic lymphocytic leukemia (CLL) is a malignancy of B lymphocytes and the most common adult leukemia in the West with remarkable clinical heterogeneity. To estimate the clinical outcome of CLL patients, the Rai and Binet clinical staging systems were developed, yet both have a limited ability at diagnosis to predict the clinical course for patients at an early clinical stage of the disease. Since most patients are diagnosed at early stages, this limitation highlights the need for alternative risk stratification approaches. The somatic hypermutation status of the IGHV genes [mutated (M-CLL), unmutated (U-CLL)], reflects fundamental differences in disease biology and clinical course. Thus, we followed a compartmentalized approach, addressing prognostication separately for M-CLL and U-CLL.

In a multi-institutional cohort of 2366 patients [M-CLL (58%); U-CLL (42%)], consolidated within ERIC, we assessed the clinical impact of various parameters regarding time-to-first-treatment (TTFT), focusing on early stage patients. Our statistical approach initially included the application of the Cox proportional hazards model. The stability of the results was validated using bootstrapping. A binary recursive partitioning algorithm, based on the development of conditional inference trees, further validated the results of Cox regression analysis.

Based on the statistical findings, we developed two prognostic indices for assessing TTFT, tailored specifically to M-CLL and U-CLL, respectively. In particular, within M-CLL and U-CLL, early stage patients were further stratified in two and three subgroups, respectively, with markedly different outcomes. We argue that such a compartmentalized approach may address the pronounced heterogeneity of CLL optimizing prognostication and, consequently, supersede previous attempts.

Nonparametric and parametric confidence intervals for the Youden index and its associated cutoff point

Christos T Nakas[1], Leonidas E Bantis[2], and Benjamin Reiser[3]

[1] Laboratory of Biometry, University of Thessaly, Volos, Greece. cnakas@uth.gr
[2] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, U.S.A. lebantis@mdanderson.org
[3] Department of Statistics, University of Haifa, Haifa, Israel. reiser@stat.haifa.ac.il

Abstract. The receiver operating characteristic (ROC) curve is commonly used to evaluate a continuous biomarker. The ROC curve is a plot of the sensitivity versus 1-specificity over all possible threshold values, c, of the marker. Although the area under the ROC curve is the most frequently used global index of diagnostic accuracy the maximum of the Youden Index, defined as $J = max_c\{sens(c) + spec(c) - 1\}$, is also used. J is equivalent to the Kolmogorov-Smirnov distance between the two populations. In practice, clinicians are often interested in determining a cutoff point for classification purposes. Frequently the "optimal" cutoff (c^*) is chosen as the value of c for which J is maximized. In the applied literature, confidence intervals for J and c^* are typically ignored. We provide new nonparametric kernel density and spline -based and parametric delta method -based approaches for constructing confidence intervals for both J and c^*. We compare our methods to currently available techniques through simulations and discuss some real examples.

Keywords

BOX-COX TRANSFORMATION, DELTA METHOD, KERNELS, ROC CURVE, SPLINES, YOUDEN INDEX

References

FLUSS, R., FARAGGI, D. and REISER, B. (2005): Estimation of the Youden index and its associated cutoff point. *Biometrical Journal, 47, 458–472.*

BANTIS, L.E., NAKAS, C.T., and REISER, B (2014): Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point. *Biometrics 2014, 70, 212–223.*

The number of strata in propensity score subclassification

Markus Neuhäuser[1]

RheinAhrCampus, Koblenz University of Applied Sciences, Joseph-Rovan-Allee 2, 53424 Remagen, Germany neuhaeuser@rheinahrcampus.de

Abstract. For the statistical analysis of non-randomized studies propensity scores are increasingly being used. One of the possible methods is stratification, also called subclassification, based on propensity scores. We compare the standard of using five strata with alternative numbers of strata, both in a simulation study as well as based on real data. We use data from a study where patients with triple vessel disease undergoing coronary artery bypass surgery with and without previous percutaneous coronary intervention were compared (Thielmann et al. 2007). We present results given in our article Neuhäuser et al. (2017) and also additional more recent findings. According to our results more than five strata may be preferable, but more than ten strata hardly gives any further benefit. We conclude that establishing guidelines for choosing the number of strata is still an interesting avenue for future research as already mentioned by Lunceford and Davidian (2004).

Keywords

LOGISTIC REGRESSION, PROPENSITY SCORE, STRATIFICATION

References

LUNCEFORD, J.K. and DAVIDIAN, M. (2004): Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine, 23, 2937–2960.*

NEUHÄUSER, M., THIELMANN, M. and RUXTON, G.D. (2017): The number of strata in propensity score stratification for a binary outcome. *Archives of Medical Science, in press.*

THIELMANN, M., NEUHÄUSER, M., KNIPP, S., et al. (2007): Prognostic impact of previous percutaneous coronary intervention in patients with diabetes mellitus and triple-vessel disease undergoing coronary artery bypass surgery. *Journal of Thoracic and Cardiovascular Surgery, 134, 470–476.*

Dynamic borrowing through empirical power priors that control type I error

Stavros Nikolakopoulos[1] and Kit Roes[1]

Department of Biostatistics and Research Support
Julius Center for Health Sciences and Primary Care,
University Medical Center Utrecht, The Netherlands
S.N.Nikolakopoulos@umcutrecht.nl

Abstract. Incorporation of historical data in the design and analysis of a new clinical trial is of particular interest in the area of (very) rare diseases, where available data is scarce and heterogeneity is less well understood. Furthermore, prospectively planning such a task is paramount for control of operational characteristics. Particularly for borrowing evidence from a single historical study, the concept of power priors can be useful. Power priors employ a parameter $\gamma \in [0, 1]$ which in commonly encountered situations has a direct translation as the fraction of the sample size of the historical study that is included in the analysis of the new study. However, the possibility of borrowing data from a historical trial will usually be associated with an inflation of the type I error. We suggest a new, simple method of estimating the power parameter in the power prior formulation, suitable for the case when only one historical dataset is available. The method is based on predictive distributions and parameterized in such a way that the type I error can be controlled by calibrating the degree of similarity between the new and historical data. The method is developed for normal responses in a one or two group setting but the generalization to other models is straightforward.

Keywords

Power priors, Clinical trials, rare diseases, predictive p-values

The latent scale covariogram: a tool for exploring the spatial dependence structure of non-normal responses

Samuel D. Oman[1] and Jorge Mateu[2]

[1] Hebrew University, Jerusalem, Israel oman@mail.huji.ac.il
[2] Universitat Jaume I de Castellon, Castellon, Spain mateu@mat.uji.es

Abstract. Let Y_i be spatially dependent non-normally distributed responses (e. g., disease prevalence in different regions) which we wish to model in terms of vectors $\mathbf{x}_i$ of explanatory variables, using a hierarchical generalized linear model (GLIM) in which the dependence structure is expressed via a latent Gaussian field $\mathbf{Z} = \{Z_i\}$. At the exploratory stage, it is common practice to first fit a GLIM assuming independence, and then examine the variogram of the residuals $Y_i - \hat{Y}_i$ to determine a possible parametric model for the autocorrelation function of $\mathbf{Z}$. This is not appropriate, however, since (unless an identity link function is used) Y_i and Z_i are on different scales. We propose here an alternative, the latent scale covariogram (LSC), whose graph reflects the autocorrelation structure of the underlying Gaussian field. We illustrate its use on a large data set involving rat sightings in Madrid, and obtain results quite different from those obtained using the variogram. Moreover, fitting an exponential curve to the LSC, which is based on the residuals at the exploratory stage, gives virtually the same parameter estimates as those obtained after fitting a hierarchical GLIM.

Keywords

GENERALIZED LINEAR MODEL, SPATIAL CORRELATION,
VARIOGRAM

Fuzzy C Means Algorithm Applications on MiRNA Gene Expression data and evaluation of miRNA pathways

Suriye Ozgur[1], Bakiye Goker Bagca[2], Muhterem Duyu[3], Ozgur Cogulu[4], and Mehmet N. Orman[1]

[1] Ege University Department of Biostatistics and Medical Informatics
suozgur35@gmail.com mehmet.orman@ege.edu.tr
[2] Ege University Department of Medical Biology
[3] 3Medeniyet University, GÃ¶ztepe Training and Research Hospital, Department of Pediatric Health and Diseases
[4] Ege University Department of Medical Genetics

Abstract. In this study, unlike classical clustering algorithms fuzzy C-means clustering is applied to microarray data. Due to the classical approaches made according to the nature of the algorithms, exactly defined descriptions may cause some of the relationships to be overlooked. For this reason, hidden relations have been tried to be revealed via fuzzy method. This study was made to evaluate the effect of miRNAs on children with all by using miRNA expression data on healthy and ill children (study cases) with sets containing different numbers of elements of fuzzy c means algorithms. Structures of miRNAs in different sets and with different number of elements obtained from the algorithm were evaluated. MiRNAs of this structure were investigated for their common properties on mRNAs pathways. The significance was all set at p¡0.05, and fold change cut-off was used 2.0 for microarray. Only 108 of 1078 miRNAs were provided this condition. Differences in miRNA expression between the cases and controls were assessed by independent Students t-test. Only 46 of 108 miRNAs were significantly upregulated or down regulated. Fuzzy C means clustering was performed using R Project for Statistical Computing to 46 miRNAs. 2, 3, 4, 5 and 6 fuzzy clusters were obtained via fuzzy C means algorithm. In each cluster, miRNAs load to databases like KEGG, OMIM, miRWalk, TargetScan, miRANDA. Then characteristics common properties and effects of miRNAs on mRNAs were investigated. New pathways in which miRNAs are affected and their relationship with ALL have been investigated. We discuss, new pathways associated with ALL may be described and those pathways may provide guidance to open up new horizons in the field of miRNA studies.

Keywords

Fuzzy C Means Clustering, Cluster Validity, miRNA, Cancer, miRNA Target Prediction Tools

Classifying Resting State Functional Magnetic Resonance Imaging Data

Ebru Ozturk[1] and Ozlem Ilk[2]

[1] Department of Biostatistics,Faculty of Medicine,Hacettepe University, Ankara,Turkey ebru.ozturk3@hacettepe.edu.tr
[2] Department of Statistics, Faculty of Arts and Sciences, Middle East Technical University, Ankara,Turkey oilk@metu.edu.tr

Abstract. Most of the functional magnetic resonance imaging (fMRI) data are based on a particular task. The fMRI data are obtained while the subject performs a task. Yet, it is obvious that the brain is active although the subject is not performing a task. Resting state fMRI (R-fMRI) is a comparatively new and popular technique for assessing regional interactions when a subject is not performing a task. This study focuses on classifying subjects as healthy or patient with the diagnosis of schizophrenia by analyzing R-fMRI data. The resting state situation in the dataset of UCLA Consortium for Neuropsychiatric Phonemics LA5c Study is used to extract brain signals in the region of interest analysis. The default mode network (DMN) ROIs were selected since the DMN is a perception depend on an interconnected set of areas displaying higher activity during rest than task related activity (Raichle and Snyder, 2007). Pre-processing of fMRI images was achieved with toolbox of statistical parametric mapping version 8 (SPM8). ROI-based on brain signals were obtained from Functional Connectivity (CONN). After brain signals are obtained, the disease status is predicted by adjusting for the magnitude of brain signals, the time during resting state, the demographic informations of subjects such as gender and age. Generalized estimating equations(GEE) approaches are conducted to classify the subjects by using R-Studio(version 1.0.136).

Keywords

R-fMRI, GEE, DMN

References

RAICHLE, M. and SNYDER, A. (2007). A default mode of brain function: A brief history of an evolving idea. *Neuroimage, 37(4), 1083-1090.*

Three different approaches to diagnose of rheumatic diseases by using quality of life scores

Ozge Pasin[1], Handan Ankarali[2], and Safinaz Ataoglu[3]

[1] Istanbul University Faculty of Medicine, Biostatistics Department,Turkey
ozgepasin90@yahoo.com.tr
[2] Duzce University Faculty of Medicine, Biostatistics Department,Turkey
handanankarali@gmail.com
[3] Duzce University Faculty of Medicine, Physical Medicine and Rehabilitation Department,Turkey safinazataoglu@duzce.edu.tr

Abstract. The purpose of this study is to show similarities and dissimilarities in terms of Quality of Life(QoL) for fibromyalgia, osteoarthritis and rheumatoid arthritis.Thus, the effect of the QoL will be revealed in the separation of the three diseases.The data was obtained from 281 volunteers who were diagnosed with fibromyalgia (FMS, n=59), osteoarthritis (OA, n=169) and rheumatoid arthritis (RA, n=53). We used six different QoL scales as follow: SF-36, SF-12, SF-8, SF-6D, QuickDash, WHOQoL-Bref. Three new approaches were used for data analysis.In the first approach, SCT(Supervised Classification Tree) after K-means clustering and after Cascade K-means clustering algorithm.In second, UCT(Unsupervised Classification Tree) was used and in the last,only SCT algorithm was used for separating disease groups.The agreement of groups obtained by K-means clustering with real groups was statistically significant.According to these clusters, it was seen that QoL scores distinguish OA and RA better.By using SCT algorithm after K-means clustering, Physical component summary(PCS) SF-36,SF-6D, SF-12, Mental component Summary(MCS) SF-8, Quick-Dash and WHOQoL-Domain 2 scores had significant effects on the occurrence of these clusters. In addition,after Cascade K-means clustering and SCT, PCS SF-36, PCS SF-8, MCS SF-8 and WHOQoL-Domain 2 scores were found significant effects on clusters. In the second approach, 10 homogeneous groups were obtained by UCT and the majority of the patients in the first group of these clusters were RA, the majority in the 5th cluster was actually OA or RA. In the last approach, 66.1% of the 59 patients who were diagnosed with FMS, 58% of the 169 patients with OA, and 90.6% of the patients with RA were correctly classified by SCT. Total accuracy is 65.8%.According to our results, it can be said that the SCT algorithm distinguishes disease groups better than other algorithms.

Keywords

CART,UNSUPERVISED LEARNING,CLUSTERING,QUALITY of LIFE

Using Pseudo Individual Patient Data in Random-effects Meta-analysis

Katerina Papadimitropoulou[1], Saskia le Cessie[12], Olaf Dekkers[1], and Theo Stijnen[2]

[1] Clinical Epidemiology, Leiden University Medical Center - Albinusdreef 2 2333 ZA Leiden, The Netherlands a.papadimitropoulou@lumc.nl o.m.dekkers@lumc.nl

[2] Medical Statistics, Leiden University Medical Center - Einthovenweg 20 2333 ZC Leiden, The Netherlands s.le_cessie@lumc.nl t.stijnen@lumc.nl

Abstract. Meta-analysis, a method for synthesizing individual study findings in a quantitative manner, has gained a lot of attention over the years. When unexplained heterogeneity between study findings is present, it is advised to perform random-effects meta-analysis, assuming that the true effects measured in each study follow a normal distribution. Several methods have been proposed to deal with the between-study heterogeneity and the most common is the DerSimonian and Laird (DL) approach. The DL estimator has been challenged over the years due to known limitations i.e., the within-study variance being treated as fixed and known, the fact that small number of studies can lead to biased estimation, the assumption of equal variances in the control and treated group. When the outcome is continuous and individual patient data (IPD) are available, linear mixed modelling methods can be employed to address these limitations. However, IPD are seldom available. In this work, we develop an algorithm to generate pseudo individual patient data by using the aggregate mean and standard deviation within each study, i.e., the sufficient statistics. Three different modelling options are explored; assuming fixed study and treatment effects, fixed study but unexplained heterogeneity in treatment differences, i.e., treatment differences vary across studies and assuming both study and treatment effects to be random. Within each model, we investigate various variance-covariance modelling options for the within-study variance, arm-specific variances, trial-specific variances and simpler models assuming equal within-study variance between treatment arms. The methods are illustrated using SAS PROC MIXED. We explore the methods for the meta-analysis of continuous patient outcome data on two example datasets in Alzheimer's disease.

Keywords

META-ANALYSIS,RANDOM-EFFECTS MODEL, INDIVIDUAL PATIENT DATA, LINEAR MIXED MODELS

Individualized Dynamic Prediction of Survival under Time-Varying Treatment Strategies

Papageorgiou G.[1,2*], Rizopoulos D.[1], Mokhles M. M.[2], Takkenberg J. J. M.[2]

[1] Department of Biostatistics, Erasmus University Medical Center, PO Box 2040, 3000 CA, Rotterdam, The Netherlands
* correspondence author: g.papageorgiou@erasmusmc.nl

[2] Department of Cardio-Thoracic Surgery, Erasmus University Medical Center, PO Box 2040, 3000 CA, Rotterdam, The Netherlands

Abstract. Our work is motivated from a study conducted at the department of Cardio-Thoracic Surgery of the Erasmus University Medical Center in the Netherlands. This study concerns patients who received an allograft for Right Ventricular Outflow Tract (RVOT) reconstruction after previous Tetralogy of Fallot (ToF) correction and were thereafter monitored echocardiographically. Cardio-thoracic surgeons are interested in studying the change in the longitudinal profile of the echocardiography measurements after RVOT reconstruction, and utilizing this change in obtaining more accurate risk probabilities of survival for these patients.

To achieve this goal we propose here a flexible joint modeling framework for the longitudinal echocardiography measurements and the hazard of death that includes RVOT as a time-varying binary covariate in both the longitudinal and survival submodels. We consider a set of joint models that postulate different effects of RVOT in the longitudinal profile and the risk of death, and different formulations of the association structure. Based on these models we derive dynamic predictions of conditional survival probabilities, adaptive to time-varying RVOT reintervention strategies. The predictive accuracy of these predictions is evaluated with a repeated cross-validation procedure using a time-dependent ROC analysis. The results suggest that it is important to account for the change in the longitudinal profiles of RVOT.

Keywords

Joint models, Dynamic predictions, Time-varying treatment strategies, Longitudinal data analysis, survival analysis

Attenuated spline reconstruction technique for SPECT/CT

Nicholas E. Protonotarios[1,2], Athanassios S. Fokas[3] and George A. Kastis[1]

[1] Research Center of Mathematics, Academy of Athens, Athens 11527, Greece
protost@hotmail.com and gkastis@academyofathens.gr
[2] Department of Mathematics, National Technical University of Athens, Athens 15780, Greece
[3] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK t.Fokas@damtp.cam.ac.uk

Abstract. The analytical approach to single photon emission computed tomography (SPECT) requires the inversion of a certain generalization of the two-dimensional Radon transform, which is called *attenuated Radon transform*. Both Radon and attenuated Radon transforms are line integrals. Here we present a modification of the explicit formula for this inversion which was derived in 2006 by one of the authors, following the pioneering work of Novikov. We also present a numerical implementation of this *Inverse Attenuated Radon Transform* (IART), which we call the *attenuated Spline Reconstruction Technique* (aSRT). For this numerical implementation we utilize both the *attenuated sinogram* obtained from SPECT and the reconstructed attenuation coefficient obtained from a CT (computerized tomography) scan. These data can be provided by a SPECT/CT scanner. Our analytic formula of the IART involves the calculation of the Hilbert transform of the linear attenuation correction coefficient and the Hilbert transform of two sinusoidal functions of the attenuated sinogram. For the aSRT we have employed custom-made cubic splines, i.e. interpolation through piecewise-continuous third degree polynomials. The purpose of this work is to present the mathematical formulation of aSRT and to evaluate it via the reconstruction of various simulated phantoms, including an image-quality (IQ) phantom under Poisson noise.

Keywords

SINGLE PHOTON EMISSION COMPUTED TOMOGRAPHY (SPECT), ATTENUATED RADON TRANSFORM, MEDICAL IMAGING

References

FOKAS, A.S., ISERLES, A. and MARINAKIS, V. (2006): Reconstruction algorithm for single photon emission computed tomography and its numerical implementation. *J R Soc Interface, 3(6), 45–54.*

A new measure for prognostic index evaluation

Paola M.V. Rancoita[1]

University Centre of Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, Milano, Italy rancoita.paolamaria@unisr.it

Abstract. A challenge goal in many clinical studies is the definition of a prognostic index, i.e. of a classification scheme that divides the patients in groups with different event-free survival curve, on the basis of a subset of clinical variables. In the practice, this subdivision may be used by clinicians to decide for the most suitable treatment for each patient depending on the different grade of prognosis.

In the literature, once a new prognostic index is defined, the common assessment of its performance is usually done through a score (e.g. the c-index or the Brier score) that actually evaluates only one or two characteristics of an ideal prognostic index. In order to have a more comprehensive view, we defined a new measure of separation which consists of a weighted difference of the mean survival times of the resulting prognostic groups. The definition of the score allows it to account for three important features of the groups: 1) the ordering, 2) the reliability in terms of size, 3) the "spread" of the corresponding survival curves. Since this separation index (called ESEP) does not account for the goodness of survival prediction, it is intended to be used in practical applications together with an error measure of survival prediction (such as the Brier score) for a complete evaluation. In the present study, we show the theoretical properties of ESEP and its advantages with respect to other measures defined in the literature, using both simulated and real data.

Sequential Cox and marginal structural models to evaluate the role of anticoagulant therapy on mortality in haemodialysis patients

Paola Rebora[1], Emanuela Rossi[1], Simonetta Genovesi[1], and Maria Grazia Valsecchi[1]

School of Medicine and Surgery-University of Milano-Bicocca paola.rebora@unimib.it

Abstract. The evaluation of the risk/benefit ratio of oral anticoagulant therapy (OAT) in patients with atrial fibrillation and end-stage renal disease is complicated by the time dependent nature of this treatment and by the presence of time dependent confounders (such as bleeding events and the international normalized ratio), that can influence/cause the interruption of the treatment. We explored the ability of the sequential Cox and marginal structural models to deal with the complexity of this setting with the aim of obtaining an unbiased estimate of the effect of OAT. By using data from a prospective study, where detailed information on treatment intake and time varying covariates were collected beyond baseline data, we tackled these issues by a causal approach applying the marginal structural and the sequential Cox models. The main purpose of the present study is to compare the performance of these models in the evaluation of the relationship between OAT and mortality, accounting for time dependent confounders. The sequential Cox model has the limit than does not deal with intermittent treatments. In our application the main analysis only considered the first switch of therapy (stopping OAT). The two models gave similar results showing an advantage of OAT on mortality.

Keywords

marginal structural models, causal models, sequential Cox model

References

GRAN, J.M., ROYSLAN K., WOLBERS M. et al (2010): A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. *Stat Med, 29, 2757-68.*

HERNAN M.A., BRUMBACK B., ROBINS J.M. (2000): Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology, 11, 561-570.*

Choosing the number of classes in Bayesian finite mixture models using the posterior distribution of the mixing proportions

Kazem Nasserinejad[1], Joost van Rosmalen[1], Wim de Kort[2], and Emmanuel Lesaffre[1,3]

[1] Department of Biostatistics, Erasmus MC, Rotterdam, the Netherlands
[2] Sanquin Research, Department of Donor Studies, Amsterdam, the Netherlands
[3] L-Biostat, KU Leuven, Leuven, Belgium

Abstract. Identifying the number of classes in Bayesian finite mixture models is a challenging problem. Several criteria have been proposed, such as adaptations of the deviance information criterion, marginal likelihoods, Bayes factors, and reversible jump MCMC techniques. It was recently shown that in overfitted mixture models, the overfitted latent classes will asymptotically become empty under specific conditions for the prior of the class sizes. This result may be used to construct a criterion for finding the true number of latent classes, based on the removal of latent classes that have negligible mixing proportions. Unlike some alternative criteria, this approach is easily implemented in complex statistical models such as latent class mixed-effects models using standard Bayesian software.

We performed an extensive simulation study to develop practical guidelines to determine the appropriate number of latent classes based on the posterior distribution of the mixing proportions, and to compare this criterion with alternative criteria. We considered various scenarios with different degrees of separation between latent classes as well as scenarios with longitudinal data, to assess how this criterion performs in a realistic setting. The performance of the criterion is illustrated using a data set of repeatedly measured hemoglobin values of blood donors.

The simulation results show that the criterion based on the posterior distribution of the mixing proportions is more likely to find the true number of latent classes than alternative criteria, provided that the priors for the class-specific parameters as well as the hyperparameter of the Dirichlet distribution of the mixing proportions are chosen carefully. This criterion compares favorably to alternative model selection criteria for the number of latent classes in terms of both performance and ease of implementation.

Keywords

Bayesian statistics, Dirichlet prior, growth mixture models, mixture models

A new biosurveillance method based on convex hulls

Athanasios Sachlas[1,2], Polychronis Economou[3] and Sotiris Bersimis[2]

[1] Department of Statistics, Athens University of Business and Economics asachlas@unipi.gr
[2] Department of Statistics and Insurance Science, University of Piraeus sbersim@unipi.gr
[3] Department of Civil Engineering, University of Patras peconom@upatras.gr

Abstract. The assumption in biosurveilance is that events are uniformly distributed in the plane. In this work, we propose a new biosurveillance method, which is based on convex hulls. More specifically, the proposed test utilizes the area of a convex hull removing the furthest point, with criterion the largest reduction of the area by removing a point from the boundary of the convex hull. The rational of the test is that the area should be reduced uniformly at a rate of $1/n$ when we remove the most 'remote' point. The test statistic is corrected in order the original size (with all the points) to have size 1. A simulation study was conducted in order to determine the critical values of the test statistic. The numerical illustration showed an excellent performance of the new test.

Keywords

BIOSURVEILANCE, CONVEX HULLS

References

GRÜNBAUM, B. (2003): *Convex Polytopes, Graduate Texts in Mathematics (2nd ed.)*, Springer.

MATUSEK, J. (2002): *Lectures on discrete geometry*, Springer.

Deterministic Modeling and Inference of Biological Networks

Deniz Seçilmiş[1] and Vilda Purutçuoğlu[1,2]

[1] Informatics Institute, Middle East Technical University, Ankara, TURKEY
deniz.secilmis@metu.edu.tr
[2] Department of Statistics, Middle East Technical University, Ankara, TURKEY
vpurutcu@metu.edu.tr

Abstract. Developing technology renders the analysis of biological data highly effective by statistical and computational techniques. Inference of biological systems from the data is one of the promising outcomes of this situation since it is now crucial especially in personalized medicine. The mathematical description of biological networks can be performed mainly by stochastic and deterministic models. The former gives more information about the system, whereas, it needs very detailed measurements. On the other hand, the latter is relatively less informative, but, the collection of their data is easier than the stochastic ones, rendering it a more preferable modeling approach. In this study, we implement the deterministic modeling of biological systems due to the underlying advantage. Among many alternatives, we use the Gaussian graphical model (GGM) and evaluate its performance with respect to the random forest algorithm (RFA), which we suggest as an alternative approach to GGM. We estimate the model parameters, i.e., the structure of the networks, and assess their findings based on their accuracies. Finally, we extend the study by using copulas in the description of the data in order to reflect the non-normality, and apply the same modeling approaches to assess their effects. Under both normality and non-normality of the data, our suggested non-parametric approach, RFA, provides very promising outputs as well as GGM. These findings may enable us to unravel the true structure of the biological systems to detect the direct or indirect relationships among genes/proteins and diseases, which can be considered as a key point for the improvements in personalized and preventive medicine.

Keywords

SYSTEMS BIOLOGY, GAUSSIAN GRAPHICAL MODEL, RANDOM FOREST ALGORITHM, COPULAS.

References

BREIMAN, L. (2001): Random Forests. *Machine Learning, 45(1), 5–32.*

WHITTAKER, J. (2001): *Graphical Models in Applied Multivariate Statistics.* John Wiley and Sons.

NELSEN, R. (2013): An Introduction to Copulas. *Springer Series in Statistics, 53(9), 276.*

Comparing Logistic Regression and Decision Tree Analysis Results: A Simulation Study With An Application To Real Data

Yasar Sertdemir[1], Hacer Y. Yildizdas[2], Ferda G. Ozlu[3], Ilker Unal[4], Adnan Barutcu[5], Mehmet Satar[6], and Mustafa Akcali[7]

[1] Cukurova University School of Medicine Department of Biostatistics, Adana, Turkey
yasarser@cu.edu.tr
[2] Cukurova University School of Medicine Department of Pediatrics, Adana, Turkey
hyapicioglu@cu.edu.tr
[3] Cukurova University School of Medicine Department of Pediatrics, Adana, Turkey
ferdaozlu72@yahoo.com
[4] Cukurova University School of Medicine Department of Biostatistics, Adana, Turkey
ilkerun@cu.edu.tr
[5] Cukurova University School of Medicine Department of Pediatrics, Adana, Turkey
abarutcu@cu.edu.tr
[6] Cukurova University School of Medicine Department of Pediatrics, Adana, Turkey
msatar@cu.edu.tr
[7] Cukurova University School of Medicine Department of Pediatrics, Adana, Turkey
makcali@cu.edu.tr

Abstract. The main aim in a case control study is to define risk factors and obtaining a prediction model. The logistic regression (LR) and decision tree (DT) methods are suitable two methods for this purpose(s). LR method is probably preferred because it is better known and in use for many years and/or due to easy interpretation of the coefficient (increased risk). In recent years, the popularity of DT applications for classification in health research increased. The DT method is a useful tool because, it is highly reliable and easy to understand how the decisions are taken and convenience is taken to interpret the results. Booth got advantages but none is always superior. In a literature review we observed that in the last 10 years 14000 case control studies used LR and only 54 used DT for analysis whereas 40 of these 54 were in the last 5 years. To compare the performance of LR and DT, a data set with three categorical and three continuous variables will be simulated and 70% of simulated data will be used for training and the remaining 30% for verification. This procedure will be repeated 1000 times. To be able to compare the performance of LR and DT under different conditions, simulations will be done using: N=100, 250, 500 and 1000, case control ratio 0.3 and 0.5, inter-action between variables: Yes and No, the variance of the error term in model: high, medium and low resulting a sensitivity of (0.6,0.75 and 0.9). Two models will be applied for LR; Model1 where no interaction terms are defined in the model and Model2 where interaction terms are defined in the model and backward model selection will be applied to booth models. R 3.3 will be used for simulation and analysis. To compare the performance (sensitivity, specificity, AUC and Percentage of Correct Classification (PCC)) of logistic regression and decision tree analysis results and to find a good prediction model for the real data set. Preliminary simulation results showed that the difference in sensitivity for DT-LR is higher for data sets with case control ratio 0.3 compared to 0.5 and that the difference in specificity for DT-LR is higher for data sets with case control ratio 0.5

compared to 0.3. The mean Percentage of Correct Classification is slightly higher for LR but this difference is less than 10% for data sets without interaction and less 5% for data sets with interaction.

Keywords

LOGISTIC REGRESSION, DECISION TREE, SIMULATION, INTERACTION, AUC

References

AGRESTI, A. (2002). Categorical Data Analysis. *New York: Wiley-Interscience.*

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., STONE, C.J. (1984). Classification and regression trees. . *Monterey, CA: Wadsworth and Brooks/Cole Advanced Books and Software.*

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J.H. (2001). The elements of statistical learning: Data mining, inference, and prediction. *New York: Springer Verlag.*

HOSMER, D. (2013). Applied logistic regression. Hoboken, . *New Jersey: Wiley.*

Development of high dimensional microbiome biomarkers

Ziv Shkedy, Nolen Joy Perualila and Rudradev Sengupta

Center for Statistics, Universiteit Hasselt, Martelarenlaan 42, B-3500 Hasselt, Belgium
ziv.shkedy@uhasselt.be

Abstract. In this study, high dimensional microbiome biomarkers were developed for a clinical outcome of interest., We discuss two settings in which longitudinal microbiome data are measured over time and a response variable of interest is available for each of the subjects. We consider two response types: continuous and time to event. Our goal is to link between the microbiome measurements and the response of interest. taking into account that the treatment may influence both microbiome and the response variable. We review different joint modeling approaches in which two aspects of the association between the response of interest and microbiome are modeled: (1) an association which is driven by the treatment effect and (2) an association reflecting the correlation between the microbiome variables and the response. We discuss both parametric and non-parametric approaches

Bayesian optimal cluster designs

Satya Prakash Singh[1] and Siuli Mukhopadhyay[2]

[1] Department of Statistics, University of Haifa, Mount Carmel, Haifa 31905, Israel
snghstyprksh@gmail.com

[2] Department of Mathematics, Indian Institute of Technology Bombay, Mumbai, India
siuli@math.iitb.ac.in

Abstract. Designing cluster trials depends on the knowledge of the intracluster correlation coefficient. To overcome the issue of parameter dependence, Bayesian designs are proposed for two level models with and without covariates. These designs minimize the variance of the treatment contrast under certain cost constraints. A pseudo Bayesian design approach is advocated that integrates and averages the objective function over a prior distribution of the intracluster correlation coefficient. Theoretical results on the Bayesian criterion are noted when the intracluster correlation follows a uniform distribution. Two data sets based on educational surveys conducted in schools are used to illustrate the proposed methodology.

Keywords

Bayesian designs, Cost function, Intracluster correlation, Multi-objective optimization, Pareto optimality

References

Atkinson, A. C., and Donev, A. N., and Tobias, R. D. (2007): *Optimum Experimental Designs, With SAS.* Oxford University Press.

Donner, A., and Klar, N. (2000): *Design and Analysis of Cluster Randomised Trials in Health Research.* Arnold, London.

Assessing Variable Selection Uncertainty in Linear Models

Aldo Solari[1], Ningning Xu[2], and Jelle J. Goeman[2]

[1] University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milan, Italy
aldo.solari@unimib.it

[2] Leiden University Medical Centre, Leiden, The Netherlands

Abstract. The problem of variable selection in regression is old but still very relevant, and some recent progress has been made in this area. Notably, selective inference has been used to design new variable selection methods. Both old and new variable selection methods, however, tend to come up with very different models, especially in the presence of collinearity. This suggests that the uncertainty in the results of variable selection should be taken into account.

In this talk we aim at quantifying the uncertainty in the variable selection process for linear models. Using the closed testing procedure, we construct a confidence set of models that covers (the best approximation of) the true model with $(1-\alpha)$ confidence, allowing for first-order model misspecification. We argue 1.) that such a confidence set represents the uncertainty in the variable selection process, and should always be taken into account when interpreting the results of a variable selection method; and 2.) that every admissible variable selection method should select a model from such a confidence set.

The confidence set is characterized by its minimal elements, the minimal adequate models (MAMs). Usually the confidence set is spanned by a small number of MAMs, so that it is relatively easy to work with. We show that the proposed simultaneous inference approach is considerably less conservative than Scheffé protection. We focus on the definition of the null hypothesis of model adequateness and provide relationships with both old (Mallows 1973, Spjøtvoll 1977, etc.) and new (Berk et al. 2013, G'Sell et al. 2016, etc.) literature. Finally, we illustrate with classical examples how to construct the confidence set by using the `cherry` R package.

Keywords

CONFIDENCE SET, VARIABLE SELECTION, SIMULTANEOUS INFERENCE

References

MALLOWS, C.L. (1973). Some comments on Cp. *Technometrics, 15, 661–675.*

Estimating the intervention effect and measurement error in calibration sub-studies

Michal Talitman[1,2], Malka Gorfine[1,3], and David M. Steinberg[1,4]

[1] Department of Statistics and Operation Research, Tel-Aviv University, Ramat-Aviv, 6997801 Tel-Aviv, Israel
[2] bennoac@post.tau.ac.il
[3] gorfinem@post.tau.ac.il
[4] dms@post.tau.ac.il

Abstract. The objective of our study is to investigate theory and methods that guide how to estimate the intervention effect in the analysis of intervention studies to reduce exposure to potential health hazards, that comprise a main study in which the outcome of the intervention is assessed only by self-report and a calibration sub-study in which the outcome is measured also by a biomarker. Keogh et al presented a novel measurement error model for such studies. Whereas Keogh et al. found MLE's for the parameters via numerical maximization, we show how to derive closed expressions for the MLE's. Our approach leads to simple formulas for the MLE's of both means and variance parameters. We investigated three ways of estimating the intervention effect, each of which we expected to be approximately unbiased: the biomarker data only method, Buonaccorsi's method and MLE method, which is a closed expression of the Keogh et al. MLE. We present results on the estimation accuracy of these methods.

Keywords

MEASUREMENT ERROR, INTERVENTION, SELF-REPORT, BIOMARKER

References

Ruth H. Keogh, Raymond J. Carroll, Janet A. Tooze, Sharon I. Kirkpatrick , Laurence S. Freedman. (2016): Statistical issues related to dietary intake as the response variable in intervention trials. *Statistics in Medicine, 2016, 35.25: 4493-4508.*

Buonaccorsi JP. Measurement error in the response in the general linear model. *J Am Stat Assoc. 1996; 91:633-642.*

Misspecifying the covariance structure in a linear mixed model under MAR dropout

Christos Thomadakis[1], Loukia Meligkotsidou[2], Nikos Pantazis[1], and Giota Touloumi[1]

[1] Department of Hygiene and Epidemiology, University of Athens, Greece
[2] Department of Mathematics, University of Athens, Greece cthomadak@med.uoa.gr

Abstract. When modeling CD4 cell counts during the HIV natural history, measurements taken after treatment initiation (cART) are by definition excluded, leading to a missing data problem. Likelihood-based methods ignoring the missingness mechanism are unbiased under random missingness (MAR). However, this only holds provided that the whole model is correctly specified, implying that both the mean evolution and the covariance structure in a linear mixed model (LMM) are modelled correctly. Pre-cART CD4 cell counts are usually analyzed using an LMM with a random intercept and a random slope assuming MAR dropout. When such a model does not provide an adequate fit, it is advisable to either add a stochastic process such as Brownian motion (BM) [1] or to use splines in the design matrix of the random effects [2]. In this work we analytically show that using a simple covariance structure when the true one is more complex leads to biased population parameters under MAR dropout, with the bias being clearly linked to the extend of dropout. It is also shown that the approach of adding a BM process performs better in terms of asymptotic bias compared to the approach of adding splines for the random effects. A simple random intercept and slope model fitted to CD4 data from the CASCADE study yielded a quite steeper CD4 decline than the LMMs with a more elaborate covariance structure, and it had the worst fit evaluated by the BIC criterion. Thus, our theoretical findings are further supported by the real data application's results.

Keywords

MAR, covariance structure, bias, linear mixed model

References

1. Oliver T. Stirrup et al. (2015): Fractional Brownian motion and multivariate-t models for longitudinal biomedical data, with application to CD4 counts in HIV-positive patients. *Statistics in Medicine, 35, 1514–1532.*
2. Rizopoulos, Dimitris: *Joint models for longitudinal and time-to-event data: With applications in R.* CRC Press.

Numerical approach to modelling mutation times in continuous time multi-type branching processes

Plamen Trayanov[1], Maroussia Slavtchova-Bojkova[2], and Stoyan Dimitrov[3]

[1] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", 5, J. Bourchier Blvd, 1164 Sofia, Bulgaria plament@fmi.uni-sofia.bg

[2] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", 5, J. Bourchier Blvd, 1164 Sofia, Bulgaria bojkova@fmi.uni-sofia.bg

[3] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", 5, J. Bourchier Blvd, 1164 Sofia, Bulgaria

[4] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", 5, J. Bourchier Blvd, 1164 Sofia, Bulgaria

Abstract. The multi-type Bellman-Harris branching process represents a model in which every cell has a random life-length and at the end of its existence it produces a random number of offspring, which can be either of the same type or mutated to another type. It is of interest to study the distribution of mutation times that lead to exponential growth in the mutant population as it could be used to model cancer growth development. The distribution depends on the model parameters and often it is impossible to derive a theoretical solution without making very strong restriction on the model. This paper represents an alternative approach to solve the required integral equations numerically, without imposing unrealistic model restrictions. At the same time this allows us to study the sensitivity of the model to its individual parameters and develop intuition of how it reacts to different parameter changes.

Keywords

Decomposable branching processes, Continuous time, Mutations, Waiting time to escape mutant, Hazard function

References

Slavtchova-Bojkova, M., Trayanov, P., Dimitrov, S. (2017): Branching processes in continuous time as models of mutations: Computational approaches and algorithms. Computational Statistics and Data Analysis. http://dx.doi.org/10.1016/j.csda.2016.12.013

Methodological challenges in the analysis of longitudinal RNAseq data

Roula Tsonaka[1]

Leiden University Medical Center s.tsonaka@lumc.nl

Abstract. Identification of genes with differentially expressed profiles in follow-up RNAseq experiments is crucial for understanding the transcriptional regulatory network. Experiments may involve samples repeatedly sequenced at a couple or even more occasions, and the number of samples can vary from a handful of patients assigned to two or more experimental conditions to hundreds of patients. Depending on the experimental design at hand, several complications may arise in the statistical analysis. Proper normalization, careful statistical modelling which addresses the research questions of interest and captures key features of longitudinal RNAseq data is crucial. Currently available statistical software for RNAseq experiments cannot be successfully used for the differential gene expression analysis in all cases. They may be limited to the analysis of single or at most paired measurements and testing can preserve good statistical properties only in small sample designs. For longer follow-up designs, time-dependent over-dispersion and within samples serial correlation may complicate the statistical analysis. Common mixed-effects models can be computationally intensive and fail to converge. In this talk we will discuss statistical challenges in studying the progression of RNAseq data from normalization to differential gene expression, provide an overview of state-of-the-art methods and present a recently developed approach which pairs methods for mixed-effects models with empirical Bayes methodology to stabilize estimation of differential gene expression over time.

Keywords

RNAseq, counts, longitudinal, Poisson log-normal.

Comparison of Estimation Methods for Area under the ROC Curve with Skewed Data

Ilker Unal[1] and Yasar Sertdemir[2]

[1] Cukurova University School of Medicine Department of Biostatistics, Adana, Turkey
ilkerun@cu.edu.tr

[2] Cukurova University School of Medicine Department of Biostatistics, Adana, Turkey
yasarser@cu.edu.tr

Abstract. The area under the ROC curve can be estimated by using parametric and nonparametric methods. The choice of estimation method depends on the distribution of measurements in subjects with/without event. In parametric methods, the main assumption is that the distribution of measurement in both groups should be Gaussian or it can be transformed to Gaussian by transformation methods. In nonparametric methods, the area under the ROC curve can be estimated using geometry, i.e. using trapezoids, or using probability density function derived by kernel smoothing. In this study, we will compare some of these methods (binormal model, binormal model with transformation, trapezoidal rule, kernel smoothing) using simulated skewed data with a variety of experimental conditions; changing the AUC values, the homogeneity of variances and sample size. The advantages and/or disadvantages of these methods will be discussed.

Keywords

ROC CURVE, THE AREA UNDER THE CURVE, SKEWED DATA

References

GODDARD, M.J., HINBERG I. (1990): Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine, 9, 325–337.*

ZOU, K.H., HALL, W.J., SHAPIRO, D.E. (1997): Smooth non-parametric receiver-operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine, 16, 2143–2156.*

A weighted composite likelihood estimators through an information criterion

Kostas Florios[1], Irini Moustaki[2], Dimitris Rizopoulos[3], and Vassilis Vasdekis[1]

[1] Athens University of Economics and Business cflorios@aueb.gr, vasdekis@aueb.gr
[2] London School of Economics, I.Moustaki@lse.ac.uk
[3] Erasmus University Medical Center, d.rizopoulos@erasmusmc.nl

Abstract. Composite likelihood estimation has been proposed in the literature for handling intractable likelihoods. In the context of multivariate latent variable models estimation, Vasdekis *et. al.* (2014) proposed a weighted estimator (WAVE) that is found to be more efficient than the unweighted pairwise estimator produced by separate maximizations of pairwise likelihoods. Florios *et. al.* (2015), proposed a modification to that weighted estimator (DWAVE) that lead to simpler computations and studied its performance through simulations and a real application. In this paper, we propose an even simpler weighted estimator (CWAVE), based on the concept of the Composite Likelihood Information Criterion (CLIC) which seems to combine the strengths of the unweighted estimator for the random effects parameters and the DWAVE/WAVE estimators for the fixed effects parameters. The new estimator CWAVE performs very well in both fixed and random effects parameters identification, with high coverage, especially when the number of time points at which measurements are obtained is large, regardless of the size of the cluster size.

Keywords

multivariate longitudinal data, composite likelihood, model averaging

References

FLORIOS, K., MOUSTAKI, I., RIZOPOULOS, D. and VASDEKIS, V.G.S. (2015): A modified weighted pairwise likelihood estimator for a class of random effects models *Metron, 73, 217–228.*

VASDEKIS, V.G.S., RIZOPOULOS, D. and MOUSTAKI, I. (2014): Weighted pairwise likelihood estimation for a general class of random effects models *Biostatistics, 15, 677–689.*

Simple bias formulas for mediation analysis with unmeasured confounding

Kai Wang

Department of Biostatistics, University of Iowa, Iowa CIty, IA 52242, USA
kai-wang@uiowa.edu

Abstract. It has been long recognized that ignoring unobserved confounders common to the mediator and the outcome can result in incorrect conclusions. Bias formulas are crucial in assessing the impact of the potential confounding. In this paper, I propose simple bias formulas in a model of continuous mediator and continuous outcome that contains mediator-treatment interaction. Compared to previous studies, these formulas involve only basic model model parameters, do not require normality assumption on the residual error, and do not assumed knowing the effect size of the confounder. They provide new insight into how unobserved confounding manifests its effect.

Keywords

MEDIATION, BIAS FORMULAS, SEQUENTIAL IGNORABILITY, SENSITIVITY ANALYSIS

References

le Cessie, S. (2016): Bias Formulas for Estimating Direct and Indirect Effects When Unmeasured Confounding Is Present. *Epidemiology, 27, 125–132.*

Albert, J.M. and Wang, W. (2015): Sensitivity analyses for parametric causal mediation effect estimation. *Biostatistics, 16, 339-351.*

Optimal dose finding for confirmatory studies

Panagiota Zygoura[1,2] and Thomas Jaki[2]

[1] Frontier Science Foundation-Hellas, Athens, Greece pzygoura@frontier-science.gr
[2] Lancaster University, Bailrigg, Lancaster, United Kingdom
t.jaki@lancaster.ac.uk

Abstract. The aim of this dissertation was to use the optimal design theory to determine how dose-ranging studies can be designed. The optimal designs were found based on two different optimality criteria, the D-optimality and the c-optimality. Since the dose-response curve is usually nonlinear, we found optimal designs for both a sigmoid E_{max} model and an exponential model. The optimal designs were assessed in terms of efficiency using relative efficiencies of different arbitrary designs versus the optimal in each case. It was found that the more we deviate from the optimal design, the more efficiency is lost. Following that, simulated data were used in order to evaluate how challenging it is to fit a nonlinear model and derive the true parameters. In addition, data were simulated from a model different than the assumed one for the purpose of evaluating how easily we can estimate the true parameters when the best guesses that we have are far from the truth. Lastly, a two-stage design was implemented and two cases were explored: a) when the simulated data for the first stage had come from a sigmoid E_{max} model with our best guesses being the true parameters and b), when the simulated data for the first stage had come from a different sigmoid E_{max} model. The optimal design for each stage was reported, in the two cases along with the average optimality criterion.

Keywords

DOSE RANGING, OPTIMAL DESIGN, OPTIMALITY CRITERIA, SIGMOID E_{max} MODEL, DOSE-RESPONSE CURVE

References

FEDOROV, V.V. and LEONOV, S.L. (2014): *Optimal Design for Nonlinear Response Models.* Chapman & Hall/CRC Press, Boca Raton, Florida.

TING, N. (2006): *Dose Finding in Drug Development.* Springer, New York.

Poster session

A Non-parametric Bayes Approach for Instrumental Variable Analysis

Samrachana Adhikari[1] and Sharon-Lise Normand[2]

[1] Department of Health Care Policy, Harvard Medical School
adhikari@hcp.med.harvard.edu
[2] Department of Health Care Policy, Harvard Medical School
Sharon@hcp.med.harvard.edu

Abstract. For making a causal inference in the presence of unmeasured confounders, instrumental variable (IV) analysis plays a crucial role. Valid instruments that satisfy several assumptions about relationships between instruments, a treatment assignment and an outcome are necessary for the causal IV analysis. Most of the existing estimation methods utilize a method of moments approach within a structural models framework. While these models do not make any distributional assumptions, they are based on limiting assumptions, such as constant treatment effect or monotonicity, for identification of causal effects.

In this paper, we explore likelihood-based estimators in the IV analysis and the necessary assumptions to make a valid causal inference based on such estimators. We focus on a Bayesian approach to make a posterior inference on the estimates of causal effects of interest, with particular emphasis when there is treatment effect heterogeneity. In essence, we extend the approach outlined in Heckman et. al. (2014) by accounting for an unobserved heterogeneity via a flexible latent structure that leverages on Dirichlet process mixture priors. Our approach has several practical advantages. It provides a flexible framework to model complex latent structures and to account for correlation structures that standard approach does not. We utilize simulations to characterize operating characteristics of various estimators. A novel application to determine the causal effect of radial artery access on bleeding and vascular outcomes compared to femoral artery access for patients undergoing cardiovascular procedures demonstrates the utility of the Bayesian likelihood based IV analysis. Funded by R01-GM111339.

Keywords

latent variables analysis, treatment effect heterogeneity, health policy research

References

Heckman, J.J., Lopes, H.F., Piatek, R. (2014):Treatment Effects: A Bayesian Perspective. *Econometric reviews, 33(1-4):36-67.*

A New Hos Index for Academic Performance

Handan Ankarali[1], Ozge Pasin[2], and Seyit Ankarali[3]

[1] Duzce University Faculty of Medicine, Biostatistics Department,Turkey
handanankarali@gmail.com
[2] Istanbul University Faculty of Medicine, Biostatistics Department,Turkey
ozgepasin90@yahoo.com.tr
[3] Duzce University Faculty of Medicine, Physiology Department,Turkey
seyitankarali@duzce.edu.tr

Abstract. Scientific studies constitute an important part of academic activities. Various indexes are suggested in the literature compare different scholars. In this study, it was aimed to propose a new index called Hos index for better evaluation of academic performance. The proposed Hos index takes into account all publications of the scientist who have gone through the literature. Publications are calculated by giving a weighting according to increasing citation numbers. Number of citations received divided into intervals and number of publications per range is multiplied by a separate weighting coefficient. The allocation of citation numbers in the Web of Science database has been taken into account in determining the intervals. The sample was created by examined the h-index values of scientists from 5 different countries selected from 5 different continents or different regions of the world. A total of 210 scientists' 31375 publications are considered. In addition, the effect of the publication age on the Hos index is also eliminated. It has been determined that the distribution of citations of the each scientist's publications are similar to each other in our sample and it shows a right skewed distribution. Because of the similarity of the distribution, the various percentile values of the citations to 31375 were calculated and number of publications within the specified citation intervals. The effect of citation and publications is calculated by multiplied the number of publications entering the citation intervals with upper limit of the percentile value with gathering.This value is divided by the difference between the first and last publication year of the scientist and the adjusted performance value (Hos index) is calculated according to the publication age. Suggested Hos index is a more sensitive approach to assessing academic performance. This index takes into account all studies of researcher whether refering or not. In addition, as the number of citations increases, the performance value is higher.

Keywords

ACADEMIC PERFORMANCE,H-INDEX,GOOGLE SCHOLAR,WEB OF SCIENCE

Reliability and Validity of Short Forms of Some Scales for Used in Quality of Life in Rheumatic Diseases

Safinaz Ataoglu[1], Handan Ankarali[2], and Ozge Pasin[3]

[1] Duzce University Faculty of Medicine, Physical Medicine and Rehabilitation Department,Turkey safinazataoglu@duzce.edu.tr

[2] Duzce University Faculty of Medicine, Biostatistics Department,Turkey handanankarali@gmail.com

[3] Istanbul University Faculty of Medicine, Biostatistics Department,Turkey ozgepasin90@yahoo.com.tr

Abstract. The quality of life(QoL) is a multi-dimensional feature and the criteria used are affected by the disease and its severity. The most important QoL scale for evaluated the health of the patient better and to reveal the benefits and harms of the healthcare is SF-36. In this study our aim is to determine which scale measures the best quality of life for FMS(Fibromyalgia syndrome) patients and to discover another scale/scales (SF-12,SF-8, SF-6D) of SF-36. Thus the reliability and validity of the scales were calculated. The data in this study was obtained from face-to-face interviews with volunteers who were diagnosed with Fibromyalgia, Osteoarthritis, and Rheumatoid Arthritis who were referred to Duzce University Physical Therapy and Rehabilitation policlinic. The internal consistency of the scales and harmony between scores were examined by the Cronbach Alpha coefficient and ICC(Intra Class Correlation) and the validity was investigated by Spearman Rank correlation coefficient. WHOQOL-Bref and Quick-Dash scales were used to examining the validity of SF-12, SF-8 and SF-6D with SF-36. As a result of the study,in all of the sub-dimensions, the harmony of the scale scores was higher in the SF-12 and SF-6D than SF-8. When we evaluate both internal consistency and compliance coefficients, we have seen that SF-12, SF-8 and SF-6D are reliable scales in measuring the quality of life in FMS. We found the strongest association with the physical function sub-dimension of the SF-36 with the SF-6 scale. If we rank scales in terms of the physical and mental function, we can say that the best scales are SF-6 and SF-12. So we can say that, SF-12 can also be used to assess the quality of life for the patients in this study which contains all the sub-dimensions of SF-36. Sf-8 is not as effective as SF-12 and SF-6D in all dimensions and because of the removal of the mental health sub dimension its efficacy is reduced in FMS.

Keywords

QUALITY OF LIFE, RHEUMATIC DISEASE, SF-36

Developing an R-script for semi-automated statistical analysis

Avramopoulos A[1], Apostolidou Kiouti F[1], Goulis DG[1], Symeonidis A[2], and Haidich AB[1]

[1] Aristotle University of Thessaloniki, Health Sciences School, Faculty of Medicine
aasimakis@auth.gr

[2] Electrical and Computer Engineering Department, Aristotle University of Thessaloniki

Abstract. Data visualization, statistical analysis and accurate result reporting are important features of medical research. Various types of statistical software aid the process of data analysis. Among them, R has gained popularity due to its versatility. Despite its smooth learning curve, the statistical analysis performed by R takes time and might alienate the command-naive user from using it. The aim of this study was to construct an R script that takes medical databases as input and produces a report of the univariate statistical analysis along with quality plots in a semi-automated manner.

R version 3.3.1 was used for development of the script. Data visualization and analysis was performed by the use of few additional packages. The Markdown integration for R was facilitated for the final report. Effort was made to minimize user input for the sake of automation; the user was engaged only in the decision-making steps of the primary statistical analysis.

The script takes as input the three most common formats of databases (.csv, .xlsx, .txt) in wide format, prompts the user for information regarding those variable characteristics that are indispensable for the analysis, assesses normality of the continuous variables in an interactive way and performs the actual analysis employing the most suitable test. The final report is produced entailing a table of sample characteristics, the results of statistical tests run on outcome variables and the appropriate plots.

Keywords

R, MEDICAL DATA ANALYSIS, REPRODUCIBLE DOCUMENTS

MTHFR C677T and multiple health outcomes: An umbrella review.

Emmanouil Bouras[1], Christos Kotanidis[1], Anthoula Chatzikyriakidou[1], Sofia Kouidou[1], Anna-Bettina Haidich[1]

Aristotle University of Thessaloniki, Health Sciences School, Faculty of Medicine.
ebouras@auth.gr

Abstract. MTHFR C677T (rs1801133) is a common variant affecting a key enzyme in one carbon metabolism. As such it has been implicated in the pathogenesis of numerous different health outcomes, over the years. Our aim is to examine the strength of each unique association between polymorphism MTHFR C677T and different health outcomes and provide an overview of potential biases. We searched Pubmed and Scopus from January 1st 1990 to December 22nd 2016 to identify systematic reviews and meta-analyses of observational studies. For each meta-analysis odds ratios (OR), 95% confidence intervals (CI) and 95% prediction intervals were calculated using random and fixed effects models. Between-study heterogeneity was assessed with I^2. Overall, we examined 81 unique meta-analyses that synthesized data from 1444 studies on different outcomes. Almost half of the outcomes (37 out of 81 meta analyses with random effects model) showed that the T allele of rs1801133 was associated with increased risk of developing a disease ($p<0.05$). A suggestive evidence of class II (more than 1000 cases, $p<0.001$ by random-effects model, heterogeneity <50%, primary studies in Hardy Weinberg Equilibrium) was only found for gastric non-cardia cancer, ischemic stroke, epilepsy and Down Syndrome. There is substantial evidence linking MTHFR to several health outcomes, but a considerable number of them may reflect, residual confounding, information bias, gene-gene and gene-environment interactions.

Keywords

polymorphism MTHFR C677T, systematic review, umbrella review, health outcome

Big Data and Usage in Data Mining Applications on Medicine

BAKIRARAR Batuhan[1], KAR Irem[2], ELHAN Atilla Halil[3], and KOSE Serdal Kenan[4]

[1] Department of Biostatistics, Faculty of Medicine, Ankara University, Ankara, Turkey
bakirarar@ankara.edu.tr
[2] Department of Biostatistics, Faculty of Medicine, Ankara University, Ankara, Turkey
ikar@ankara.edu.tr
[3] Department of Biostatistics, Faculty of Medicine, Ankara University, Ankara, Turkey
elhan@medicine.ankara.edu.tr
[4] Department of Biostatistics, Faculty of Medicine, Ankara University, Ankara, Turkey
s.kenan.kose@medicine.ankara.edu.tr

Abstract. Big data is that businesses, states, hospitals and organizations integrate different digital data sets and use information that has remained hidden and surprise correlations through methods of statistics and data mining. With lower data storage costs and the emergence of NoSQL databases, Hadoop and similar database designs allowing the distribution of workload to multiple virtual servers when dealing with high-volume data, the flow of big data has been spreading even faster. Since newly developed technologies allow for analyzing different types of data for lower costs and in a faster way and emergence of new findings, big data has been accepted by all the world in a short notice and started to be used. Recently, especially the analysis of information such as videos, physician notes which cannot be stored and analyzed with classical methods through big data has facilitated obtaining several pieces of new and useful information. In this study data will be generated with the help of R which is proper to data mining applications on big data. With the help of Mahout and Scala(big data technologies), data mining methods of Random Forest and Multilayer Perceptron will be used. These data mining methods' will be compared via True Classification Rates and F-measure.

Keywords

Big Data, Data Mining, Classification, Prediction

References

CHANDRAMAN, T. (2015): Learning Apache Mahout. *Birmingham: Packt Publishing.*

LOHR, S. (2012): The age of big data. *New York Times, 11.*

NICK, P. (2015): Machine Learning with Spark. *Birmingham: Packt Publishing.*

PARTNERS, N. (2012): Big Data Executive Survey: Creating a Big Data Environment to Accelerate Business Value. *NewVantage Partners.*

Random Forest application on DNA methylation microarray data in CLL

Nestoras Chalkidis[1], Maria Tsagiopoulou[2], Nikos Papakonstantinou[2] and Theodoros Moysiadis[2]

[1] Master program on complex systems and networks, AUTH, Greece
chalkidisnestor@gmail.com

[2] Inst. of Applied Biosciences, Thes., Greece moysiadis.theodoros@certh.gr

Abstract. Chronic lymphocytic leukemia (CLL) is the most common adult leukemia in the western world. Most CLL patients will need treatment during the course of the disease and although initially responding to treatment, most patients will eventually relapse. The exact molecular mechanisms underlying disease relapse, are not yet fully understood. Moreover, the time to disease progression shows great variation among patients, consequently, there is a need for novel biomarkers predictive of the response to therapy. Previous studies have highlighted the important role of DNA methylation in CLL pathophysiology. In the present study we searched for DNA methylation signatures with potential predictive power regarding response to treatment. To this end, we evaluated DNA methylation array data from 38 CLL patients (pre-treatment/post-relapse).

We considered two biologically meaningful scenarios, where the patients were divided into two groups. The first grouping was based on the results of intra-individual differential methylation analysis (changes between pre-treatment and post-relapse). The second grouping classified patients as ultra-high risk or low risk cases according to the time to relapse. The Random Forest (RF) algorithm was selected to identify the most important methylation CpG sites towards optimizing the classification of CLL patients separately in the above scenarios. In particular, we applied a variable selection method, based on the RF algorithm, the reason being the need to focus on a small number of variables that could be further evaluated experimentally. Before applying this method, an appropriate filtering has been performed to reduce the data dimensionality (initially, approximately 450000 CpG sites/variables were available). Our results indicate that in both scenarios, the proposed methodology resulted in a very small number of CpG sites that were deemed important (approximately ten). Furthermore, the classification of patients in their original groups based on these CpG sites was very efficient in terms of error rate. Thus, we conclude that the methylation levels of specific CpG sites could be used as novel predictive biomarkers, especially regarding the time to relapse.

Algorithmic Construction of Optimal Block Designs for Two-Colour cDNA Microarray Experiments Using the Linear Mixed Effects Model

Legesse Kassa Debusho[1], Dibaba Bayisa Gemechu[1] and Linda M. Haines[2]

[1] Department of Statistics, University of South Africa, The Science Campus, GJ Gerwel (C-Block), Floor 6, Florida Park, Roodepoort, Private Bag X6, Florida 1710, South Africa (debuslk@unisa.ac.za; diboobayu@gmail.com)

[2] Department of Statistical Sciences, University of Cape Town, Rondebosch 7700, South Africa (linda.haines@uct.ac.za)

Abstract. In this paper, methods for efficient construction of A-, MV-, D- and E-optimal or near-optimal block designs for two-colour cDNA microarray experiments with array as the block effect are considered. Two algorithms, namely the array exchange and treatment exchange algorithms together with the complete enumeration technique are introduced. For large numbers of arrays or treatments or both, the complete enumeration method is highly computer intensive. The treatment and array exchange algorithms were compared on the basis of the computer time required to generate the designs and the efficiencies of the resultant criteria values. The treatment exchange algorithm computes the optimal or near-optimal designs faster than the array exchange algorithm. The three methods however produce optimal or near-optimal designs with the same efficiency under the four optimality criteria for the parameter combinations that were considered.

Keywords

Array exchange algorithm; A-, MV-, D- and E-optimal designs; complete enumeration; microarray experiment; treatment exchange algorithm

Computer-intensive methods for Model Selection in Animal Breeding

Nikos Demiris[1]

Athens University of Economics and Business nikos@aueb.gr

Abstract. This presentation is concerned with the use of certain classes of random effects models appropriate for individual selection in the presence of genetic relatedness. The first approach is based upon a standard linear mixed model where the fixed effects are correlated and their covariance matrix depends on the pedigree. The second model is using the SNPs from a genome wide association study in order to inform the mean and the covariance structure, thus incorporating Mendelian sampling directly. The effectiveness of the two models is illustrated using a dataset from animal science.

How to Overcome Class Imbalance Problem

Duygu Aydin Hakli, Dincer Goksuluk and Erdem Karabulut

Hacettepe University, Faculty of Medicine,Biostatistics Department
duygu.aydin@hacettepe.edu.tr

Abstract. In binary classification, when the distribution of numbers in the class is imbalanced, we are aimed to increase the accuracy of classification in classification methods. In our study, simulated data sets and actual data sets are used. When simulation work is planned, three different effects are considered which may affect the classification performance: sample size, correlation structure and class imbalance rates. Scenarios were created by considering these effects. 80 different scenarios including 4 different types of correlation structure, 5 different sample size and 4 different class imbalance ratios were prepared and each scenarios was repeated 1000 times. CART, SVM and RF methods have been used in the classification. SMOTE, SMOTEBOOST and RUSBOOST were used to decrease or completely remove the imbalance of the data before the classification methods were applied. Data generation, classification methods and performance were obtained using RStudio. The simulation results: the imbalance rate increases from 10 to 30, the effect of the 3 algorithms on the classification methods is similar accuracy. Because the class imbalance has become balanced. When sample size goes up to 2000, the classification accuracy has also increased in these algorithms.

Keywords

Classification methods, class imbalance, imbalance data

References

Nitesh V. Chawla and at all. (2002): SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research,16:321-357.*

D. Cieslak and N. Chawla (2008): *Learning decision trees for unbalanced data.* Springer, Germany.

X.Y. Liu, J. Wu, and Z.H. Zhou(2009): Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions,39(2):539-550.*

Estimation of blood pressure percentiles in healthy children using polynomial regression

Pembe Keskinoglu[1], Suriye Ozgur[2], Timur Kose[2], and Ahmet Keskinoglu[3]

[1] Dokuz Eylul University Department of Biostatistics and Medical Informatics
pembe.keskinoglu@gmail.com

[2] Ege University Department of Biostatistics and Medical Informatics
suozgur35@gmail.com

[3] Ege University Department of Pediatric Health and Diseases, Pediatric Nephrology

Abstract. Hypertension is one of the most important public health problem in worldwide. Hypertension in childhood and adolescence is a significant risk factor for cardiovascular disease in the adult life and it also causes early development of atherosclerosis and end-organ damage in childhood. The aim of this study is to determine the systolic and diastolic blood pressure (BP) percentiles using different models aged from 2 to 18 years in Turkish children and to evaluate the performances of the models(1).

A cross-sectional study was performed and the probability sampling method was applied. The total sample size is 3456 children. Of the total 5,417 children at 10 selected schools (nursery, pre-elementary school, primary school and high school), 4,984 (the reached rate 92%) were evaluated. Weight, height, systolic and diastolic blood pressure of these children were measured by suitable methods. The estimations of systolic and diastolic blood pressure percentiles in children using polynomial regression models were done. Fourth power of (age-10) and Z scores of the height and weight were used in polynomial regression model for the estimation of systolic and diastolic blood pressure (BP) percentiles (2).

Data were analyzed by using R Project. The mean values of systolic and diastolic blood pressure increased with age and height at both genders. Estimated values of pre-hypertension in child and adolescent via polynomial regression models were obtained lower grades from results of Europe and USA studies (3).

Keywords

Polynomial regression, blood pressure percentiles, hypertension

Evaluation of the performance of some data mining methods for hardly discriminated clinical diagnosis

Ahmet Keskinoglu[1], Suriye Ozgur[2], Caner Alparslan[3], Onder Yavascan[3], Burak Ordin[4], and Pembe Keskinoglu[5]

[1] Ege University, Department of Pediatric Health and Diseases, Pediatric Nephrology
pembe.keskinoglu@gmail.com

[2] Ege University Department of Biostatistics and Medical Informatics
suozgur35@gmail.com

[3] 2Ege University Department of Biostatistics and Medical Informatics

[4] Tepecik Training and Research Hospital, Department of Pediatric Health and Diseases

[5] Ege University Faculty of Science, Department of Mathematics

[6] Dokuz Eylul University Department of Biostatistics and Medical Informatics

Abstract. It is very difficult to distinguish between vesicoureteral reflux (VUR) and recurrent urinary tract infections (UTI) in children. Delayed diagnosis and treatment of VUR maybe the cause of some complications. Early diagnosis of VUR is important for preventing outcomes. Aim of this study is to evaluate the performance of some data mining methods for the clinical decision of VUR/UTI.

In this retrospective cross-sectional study, 611 pediatric patients who have applied to Medicine Faculty of Ege University Pediatric Nephrology Outpatient Clinic and Tepecik Training and Research Hospital were included. Informative data about the patients were obtained from hospital records and patient files. The conversion of records for data was carried out by pediatric nephrologists in the study team and the database was created. The analysis was done with Weka (Waikato Environment for Knowledge Analysis). 39 characteristics have been determined in the model by selecting the variables that are the best representatives of model for differential diagnosis of VUR/UTI with using Cfs Subset Eval(attribute evaluator) and Best fit methods. Results were obtained in this model via data mining algorithms which are cluster centroids, decision table (rule-based algorithm), multilayer perceptron, random forest, random tree and Bayesnet. In this research, 426 children (69.7%) with VUR and 185 (30.3%) children with UTI were evaluated. 41% of the children with VUR were boys and 42% of the children with UTI were boys. When the results were evaluated by data mining method, the performance of the method about the differential diagnosis of VUR/UTI was 56%-97.5%. We can conclude that this method is acceptable for differential diagnosis. Decision table (Rule-based approach) has the highest estimation rate (97.5%) for this data set. We suggest of clinical and laboratory variables in decision support system can achieves a good performance for the differential diagnosis of VUR/UTI. Prominent clinical and laboratory variables are restlessness, loss of appetite, urine PH and blood urea nitrogen. Presence of hydroureter and/or hydronephrosis in imaging methods which is important for the clinical diagnosis of VUR is also found out to be an important diagnosis method by data mining approaches. Finally, data mining approach is suitable for the differentiation of very closely related clinical diagnosis.

Keywords

Polynomial regression, blood pressure percentiles, hypertension

A meta-analysis and meta-regression approach to Obsessive-Compulsive Disorder

Laura Marciano[1] and Clelia Di Serio[2]

[1] Faculty of Psychology, Vita-Salute San Raffaele University, Milan, Italy
l.marciano@studenti.unisr.it
[2] University Center of Statistics for the Biomedical Sciences, Vita-Salute San Raffaele University, Milan, Italy diserio.clelia@hsr.it

Abstract. The aim of the present study is to systematically explore, by a meta- analytic and meta-regression approach, the non-verbal memory impairments in Obsessive-Compulsive Disorder (OCD), to describe how such a deficit can be linked to executive functioning. Neuropsychological findings of OCD patients have indi- cated deficits in a wide range of cognitive abilities, in particular in the executive functioning and in the non-verbal memory domains. Thus, this statistical approach may improve the understanding of memory underperformance.

Keywords

OBSESSIVE COMPULSIVE DISORDER, META-ANALYSIS, MEMORY DEFICIT

Perceived barriers of colorectal cancer screening in Northern Greece; a population survey of healthy adults.

Mastrokostas A[1], Gavana M[1], Touloumi G[2], Benos A[1], and Haidich AB[1]

[1] School of Health Sciences, Faculty of MedicineAristotle University of Thessaloniki, Greece, amastrokostas@auth.gr

[2] Medical School, National and Kapodistrian University of Athens, Athens, Greece

Abstract. Colorectal cancer is the third most frequent cause of cancer mortality in Greece and in most European countries. Although current guidelines highlight the beneficial effect of colorectal screening (with Fecal Occult Blood Test (FOBT) or colonoscopy) for the general population, uptake rates are very low in Greece.

We aimed to investigate factors influencing colorectal cancer screening tests participation rates in the Greek adult general population, and to explore the perceived barriers.

In this analysis, we used a sub-set of the EMENO study data (National Morbidity and Risk Factors Survey); those referring to Northern Greece (province of Macedonia and islands of Northern Aegean Sea). EMENO is a nationwide health examination survey. A representative sample of the Greek adult (≥18years) was selected using the multistage stratified random sampling method In addition to EMENO data, we interviewed all men and women EMENO participants who were over 50 years, lived in Northern Greece and had reported that they had never participated in a FOBT or colonoscopy screening test , or during the last 5 years. A semi-structured questionnaire, using a mixed method approach, was administered. All participants were asked to state the principal reason for not participating in colonoscopy and/or FOBT screening as well as to state their level of agreement, in a 4-level Likert scale, in a list of the most frequently cited barriers.Additional data on demographic and socioeconomic factors were drawn from the EMENO study database. Analysis was conducted with SPSS 23.0and STATA (version 13).

Five hundred seventeen adults were included in the analysis; 61.8% women, mean (SD) age of63.1 (12.7) years. The most frequently cited barrier for colonoscopy was lack of symptoms (40.7%), followed by lack of recommendation (19.5%) and negligence (15.3%). For FOBT, the principal perceived barrier was also lack of symptoms (42.1%),followed by lack of recommendation (22.3%) and negligence (11.6%). Interestingly, a substantial portion of the sample reported that they did not even know what the test is (8.5% for colonoscopy and 14.9% for FOBT). However, only 2.5% reported financial issues and 1.7% lack of insurance as the primary reason for not participating in colonoscopy (2.5% for FOBT). There were no significant differences between genders regarding the main reason for not participating in FOBT or colonoscopy.

Colorectal cancer screening participation of the general population in Greece is a multivariate issue. Behavioral as well as organizational factors contribute to diminished participation rates. The fact that most people value these tests suitable only in case of symptoms, and that many do not know the nature of the test, indicate a serious deficit in health literacy. Proper education of the public about the benefits of secondary prevention should be a public health priority.

Keywords

colorectal cancer, screening, FOBT, colonoscopy, secondary prevention, EMENO study

Statistical evaluation of HPV-E7 oncoprotein detection as a triage method to colposcopy

Theodoros Moysiadis[1], Theodoros Agorastos[2], Kimon Chatzistamatiou[2], Andreas M. Kaufmann[3], Alkmini Skenderi[4], Irini Lekka[5], Isabel Koch[6], Erwin Soutschek[6], Oliver Boecher[6], Vasilis Kilintzis[5], Stamatia Angelidou[7], Evangelia Katsiki[7], Ingke Hagemann[8], Eleonora Boschetti-Gruetzmacher[3], Athena Tsertanidou[2], Eleftherios Angelis[9], Nikolaos Maglaveras[5], Pidder Jansen-Duerr[10]

[1] Inst. of Applied Biosciences, Thes., Greece moysiadis.theodoros@certh.gr
[2] 4th Clinic of Obstetrics and Gynecology, AUTH, Greece agorast@auth.gr
[3] Clinic for Gynecology, CharitĂ©-Universitaetsmedizin Berlin, Germany
[4] Laboratory of Cytology, Hippokratio General Hospital, Thes., Greece
[5] Laboratory of Bioinformatics, Dept of Medicine, AUTH, Thes., Greece
[6] Mikrogen GmbH, Neuried, Germany
[7] Laboratory of Histopathology, Hippokratio General Hospital, Thes., Greece
[8] MVZ Im Mare, Kiel, Germany
[9] School of Informatics, Faculty of Sciences, AUTH, Thes., Greece
[10] Research Institute for Biomedical Aging Research, Univ. of Innsbruck, Austria

Abstract. High-risk (hr) HPV detection with HPV16/18 genotyping tends to be considered a better method of primary cervical screening than cytology. The purpose of this study is to assess the performance of the detection of E7HPV protein as a triage method for women either hrHPV (non16/18) or HPV16/18 positive, and, consequently, evaluate alternative triage processes that could possibly result in a new algorithm towards better selection of at risk women.

The study included 1473 potentially eligible participants. E7HPV protein, being a continuous variable, was dichotomized based on ROC curve analysis and the Youden's J statistic. Different indices, such as sensitivity, specificity, PPV, NPV were computed to evaluate the accuracy of the different screening methods considered. The accuracy differences between the methods were statistically compared. In addition, the positive likelihood ratio and the odds ratio of CIN2+ (the endpoint in this study) were computed for all screening methods.

We conclude that triage of either HPV 16/18 positive women or hr (non16/18) HPV positive women, to colposcopy with the E7 test results in very interesting findings and a better performance.

Assessment of diagnostic test results detected in independent samples, for correspondence with the Suzuki and standards methods

Mustafa S. Senocak[1] Hayriye Ertem Vehid[1] Nurgul Bulut[2] and Gokalp Eral[1]

[1] Istanbul University, Cerrahpasa Faculty of Medicine, Department of Biostatistics mssenocak@gmail.com

[2] Istanbul Medeniyet University, Faculty of Medicine, Department of Biostatistics and Medical IT nrgl.bulut@hotmail.com

Abstract. In the assessment of diagnosis tests, using methods other than the standard criteria yields beneficial information. It is crucial to compare the beneficial aspects of different diagnostic tests based on the same golden standard. Suzuki has suggested a method for comparing the benefits of two different diagnostic tests, based on the proportion of OR values yielded by the 2x2 contingency table used in diagnostic test practices. The Suzuki allows not only the general comparison of diagnostic tests, but also assesses them in terms of sensitivity and specificity. We attempted to assess the comparison of the features of two different diagnostic tests, by focusing on the similarities and contradictions between the Suzuki and the standard methods that compares AUC levels. The results of two different diagnostic tests for a total golden standard inventory of 120 ill and 120 healthy individuals, were collected by randomly appointing ill or healthy individuals to these tests. The AUC, SE-AUC, OR values for the two diagnosis tests were calculated by the software we designed. Based on the resulting ROR value confidence intervals and standard comparisons, z values were calculated. This was repeated 5000 times. It has been found that the method's results are largely consistent in terms of significance, and that discrepancies arise only for the very critical "z" values, in confidence intervals very close to 1. The Suzuki appears to be a rather valid option in such comparisons, since its diagnostic tests provide OR's, and additional information which includes the OR ratios.

Keywords

ODDS RATIO, ROR, AUC, SENSITIVITY AND SPECIFICITY.

References

Suzuki, S. (2006): Conditional relative odds ratio and comparison of accuracy of diagnostic tests based on 2 x 2 tables.*Journal of Epidemiology, 16,145-153.*

A review of Bayesian methods in meta-analysis of diagnostic accuracy test studies

Eirini Pagkalidou[1], Eleni Verykouki[1], and Anna-Bettina Haidich[1]

Laboratory of Hygiene, Medical School, Aristotle University of Thessaloniki, Thessaloniki, Greece. pagalidou@auth.gr

Abstract. Meta-analysis of diagnostic test accuracy studies differs from the usual meta-analysis of therapeutic/interventional studies in that, it is required to simultaneously analyze a pair of two outcome measures such as sensitivity and specificity, instead of a single outcome. Several different methods have been proposed for meta-analysis of diagnostic test accuracy studies, but there is still considerable uncertainty regarding the best method to synthesize those studies.

There are currently two analytical models available for hierarchical modeling: the bivariate model (1) and the hierarchical summary receiver operating characteristic (HSROC) model (2). Both models utilize a hierarchical structure of the distributions of data in terms of two levels, and provide equivalent summary estimates for sensitivity and specificity under the special condition. Using Bayesian methods, improved estimates of sensitivity and specificity are possible, especially when prior information is available on the diagnostic accuracy of the reference test.

We will review Bayesian methods that are used in meta-analysis of diagnostic test accuracy studies and point their advantages and weaknesses.

Keywords

Meta-analysis, Network Meta-analysis, Diagnostic test accuracy, Hierarchical models, Bayesian Statistics

References

REITSMA, J.B., GLAS, A.S. ET AL (2005): Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol 2005, 58, 982-990.*

RUTTER, C.M. AND GATSONIS, C.A. (1995): Regression methods for metaanalysis of diagnostic test data. *Acad Radiol 1995, 2 Suppl 1, S48-S56.*

Analyzing Data from BCI Control of Robotic Arms

N. Pandria, A. Athanasiou, G. Arfaras, P.D. Bamidis,

Lab of Medical Physics, Faculty of Medicine, School of Health Sciences, Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece npandria@gmail.com

Abstract. Brain-Computer Interfaces (BCIs) have used motor imagery (MI) to interpret brain activity into control of robotics. While kinesthetic (KMI) and visual motor imagery (VMI) both induce brain activation patterns similar to an actual motor task, they correspond to distinct neural systems and their suitability for BCI applications has not been investigated. We analyzed BCI skill training, performance and “Godspeed” questionnaire answers of 30 healthy individuals (18 male, 12 female) during a comparative experiment incorporating VMI and KMI paradigms as BCI (Emotiv) control modalities to operate two robotic arms (Mercury 2.0) in simple motions. Demographic data with regards to age across sexes and different hand dominance were explored using the Mann-Whitney (U) and Kruskal-Wallis (H) tests respectively with no significant differences. Godspeed’s total and sub-categories scores were analyzed with no significant findings (sex and superior performance in KMI or VMI as grouping factors; normality tested using Shapiro-Wilk Test). KMI against VMI skill training percentages were compared separately for each hand, across training blocks using Paired t-test (t) comparisons and differences for each block were further explored grouping by hand dominance and sex. Significant difference was found only in KMI against VMI skill training percentages for the Right Hand, gathered during training block 2 for female participants. Descriptive data highlighted that even though KMI could result in increased skill rates, there is a “fatigue” effect. Finally, success rates of BCI control between KMI and VMI were treated using Wilcoxon Signed Ranks Test. While VMI outperforms KMI, differences were not significant.

Keywords

ROBOTIC ARM, BRAIN COMPUTER INTERFACE, VISUAL MOTOR IMAGERY, KINESTHETIC MOTOR IMAGERY

Accounting for heterogeneity in a sparse-event meta-analysis of a few small trials

Konstantinos Pateras[1], Stavros Nikolakopoulos[2], and Kit Roes[3]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands, k.pateras@umcutrecht.nl
[2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands, S.N.Nikolakopoulos@umcutrecht.nl
[3] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands, K.C.B.Roes@umcutrecht.nl

Abstract. Clinically important outcomes in randomized clinical trials (RCTs) are often expressed as a dichotomous variable. In rare diseases, only a small number of patients is available per RCT. Such small sample sizes increase the chance of observing zero events. When the number of zero events increases, a meta-analysis (MA) via a simple frequentist normal-hierarchical model, induces bias and improper (interval) estimation of the overall treatment effect.

A full Bayesian model, which is a viable solution for MA in rare diseases, can be used instead. In such sparse settings, Bayesian models are sensitive to the choice of variance (heterogeneity) prior distributions. This choice heavily influences posterior inference of the heterogeneity parameter and results in improper (interval) posterior estimation of the overall treatment effect.

We performed a simulation study to evaluate the impact of alternative methods for estimating heterogeneity on the overall treatment effect in a sparse-events MA of a few small RCTs under a frequentist and a Bayesian hierarchical model. We evaluated each method by reporting the frequentist operational characteristics of bias and coverage.

The Bayesian approach performed more robustly than the frequentist one. When a Bayesian MA is performed under such sparse conditions we recommend using priors whose density is concentrated but not restricted to plausible τ values such as a $Uniform(-10, 10)$ prior on the $log(\tau^2)$ scale.

The impact of the proposed methods is illustrated via two meta-analyses of rare diseases.

Keywords

Bayesian, meta-analysis, heterogeneity, rare events, rare diseases

Evaluation of Different Statistical Methods Used to Test for Agreement between the Results of Continuous Measurements

Ozge Pasin[1], Ahmet Dirican[1], Basak Gurtekin[1], Sevda Ozel Yildiz[1], and Rian Disci[1]

Istanbul University Faculty of Medicine, Biostatistics Department,Turkey
ozgepasin90@yahoo.com.tr, diricana@yahoo.com, bgurtek@istanbul.edu.tr,
sevda@istanbul.edu.tr, rian@istanbul.edu.tr

Abstract. New approaches and instruments are developing for the purpose of measuring various variables, with the aim of providing cheaper, more convenient and safe methods. When a new method of measurement or instrument is invented, the quality of it has to be assessed. In such studies, linear regression analysis are often used for investigating degree of agreement of two quantitative measurements. But linear regression allows measurement error only in Y variable. So in this case, researchers can use Deming regression analysis and Bland-Altman methods for accordance. This methods allows measurement error in both X and Y variables. In this study, we discussed Deming regression, Bland-Altman and ICC (intraclass correlation coefficient) methods, including the application of inappropriate statistical methods, multiple statistical methods, and the strengths and weaknesses of each methods. As an application, we compare digital thermometry of incubator and axillary fever measurements for newborn. We have found a good accordance between measurements with methods of Deming, Bland Altman and ICC. For example, correlation values were obtained whereupon 0.70 in the comparison of the variables between the two measurements and their repeated values. As a result, the methods of Deming regression, Bland-Altman and ICC are useful and give similar results for measuring accordance between two numerical measurements.

Keywords

DEMING REGRESSION, BLAND ALTMAN, ICC,METHOD COMPARISON

Methods for estimating the cumulative effect of correlated exposures: an application on the health effects of air pollution in Athens, Greece

Sophia Rodopoulou[1], Klea Katsouyanni[1], Pagona Lagiou[1], and Evangelia Samoli[1]

Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Greece srodopoyl@med.uoa.gr

Abstract. Assessment of the cumulative effect of correlated exposures is an open methodological issue with several applications. In environmental epidemiology it is encountered in the evaluation of the cumulative health effects of mixtures of pollutants. Previous studies have applied regression models with interaction terms [1] or identified the most health relevant components of the pollutant mixture by dimension reduction methods [2]. In cancer epidemiology, the combined effect has been evaluated through the use of a score of exposures of interest that incorporates weights based on the strength of the component-specific associations with health outcomes [3]. We estimated the joint effect of six air pollutants (particulate matter with diameter less than 2.5 μm and between 2.5-10 μm, nitrogen dioxide, sulfur dioxide, carbon oxide and ozone) on daily natural and respiratory mortality in Athens, Greece, in 2007-2012. We assessed three different methods under the context of a Poisson regression allowing for overdispersion for the investigation of the effects of short term exposure to the pollutants: a) including all pollutants and their first-order interactions, b) including only pollutants selected via adaptive LASSO and c) using a weighted exposure score with pollutant-specific weight the corresponding effect retrieved from published reviews. The pollutants correlations ranged from -0.48 to 0.73. The estimation of their joint effect on natural mortality was similar between methods (% increase and 95% confidence interval (CI) per 1 interquartile range (IQR) increase per pollutant or score; all pollutants model: 1.12% (-1.22%, 3.52%), adaptive LASSO model: 1.08% (-1.24%, 3.46%) and score model: 1.00% (0.14%, 1.87%)). When we assessed the joint effect on respiratory mortality, which presented less number of events and less variance compared to natural mortality, the score model resulted in a different cumulative estimate compared to the other methods (% change and 95% CI per 1 IQR increase per pollutant or score; all pollutants model: -0.56% (-14.56%, 15.73%), adaptive LASSO model: -0.61% (-14.55%, 15.61%) and score model: 1.21% (-3.73%, 6.40%)). Our findings suggest that, when an endpoint with sufficient number of cases is investigated, the use of an exposure score estimates a similar joint effect to a multi-pollutant or a dimension reduction method, while the effects are differentiated when analyzing more sparse health indices.

Keywords

multi-pollutant models, air pollution cumulative effect, adaptive LASSO, weighted exposure score

References

[1] Sun Z, Tao Y, Li S, et al. (2013) Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environ Health 12(1): 85*

[2] Winquist A, Kirrane E, Klein M, et al. (2014) Joint effects of ambient air pollutants on pediatric asthma emergency department visits in Atlanta, 1998-2004. *Epidemiology 25(5): 666-73*

[3] Tworoger SS, Rosner BA, Willett WC, et al. (2011) The combined influence of multiple sex and growth hormones on risk of postmenopausal breast cancer: a nested case-control study. *Breast Cancer Res 13(5): R99*

Non-orthogonal multi-stratum design of experiments

Luzia A. Trinca[1]

Institute of Biosciences, Unesp, Botucatu, São Paulo, Brazil ltrinca@ibb.unesp.br

Abstract. Multi-stratum experiments arise when practical restrictions on the randomisation process lead to information appearing at different levels of experimental unit from different treatment factors. Analysis of data is performed by fitting mixed effects models. When resources are limited the classic orthogonal layout is not viable and we need methods to construct efficient designs. Usually optimum designs for point estimation of fixed effects allow few degrees of freedom for estimating pure error variance components. However such estimates are required in case inferences are to be performed using the experimental data. Here we use the modified optimality criteria for inference of Gilmour and Trinca (2012) in the stratum-by-stratum design approach of Trinca and Gilmour (2015) in order to construct multi-stratum designs that allow for pure error degrees of freedom in the several strata. We explore a few combinations of properties in the criteria, including compound criteria, and study several properties of the obtained designs. Besides being better for inferences on the model parameters than standard designs our designs are shown to be quite competitive in terms of response and difference response predictions as well. As illustration we use a biotechnology laboratory experiment.

Keywords

OPTIMUM DESIGN, PURE ERROR, MIXED MODEL

References

TRINCA, L. A. and GILMOUR, S. G. (2015): Improved Split-Plot and Multistratum Designs. *Technometrics, 57, 145–154.*

GILMOUR, S. G. ; TRINCA, L. A. (2012): Optimum design of experiments for statistical inference. *Applied Statistics, 61, 345–401.*

Differences in correlation structure of gene expression data, used for variable selection in Support Vector Machines algorithm

Athina Tsanousa[1] and Lefteris Angelis[2]

[1] Aristotle University of Thessaloniki, Greece atsanous@csd.auth.gr
[2] Aristotle University of Thessaloniki, Greece lef@csd.auth.gr

Abstract. This paper concerns Support Vector Machines (SVM), a supervised learning technique which is applied to data that are already labeled as members of categories. SVM are useful for classification and prediction and the typical application involves a learning phase on a training dataset in order to predict the classes of new cases of a test dataset. As with most machine learning techniques, variable selection methods can increase the SVM accuracy of prediction by reducing the dimensions of the dataset, especially in cases of big data.The basic idea of the present work is that in certain datasets the correlation structure of variables is complicated and different in specific classes; therefore the recognition of underlying patterns is of greater importance than studying the behavior of individual variables. Here, by comparing correlation coefficients of two distinct groups, we choose the variables that participate in most pairs that are significantly different and we use this procedure as a variable selection method in order to eventually determine the variables that will next be included in a SVM classifier. The combination of the aforementioned variable selection with SVM was applied to gene expression data comprised of various subcategories and improved the accuracy of the predicted categories of the test set, compared to the results received by including all the variables in the algorithm.

Keywords

Comparing correlations, Support Vector Machines, Gene expression data

References

Jennrich, R. I. (1970): An asymptotic chi-square test for the equality of two correlation matrices. *Journal of the American Statistical Association, 65(330), 904-912.*

Burges, C. J. (1998): A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery, 2(2), 121-167.*

Joint Modeling of Longitudinal Measurements and Survival Time Data- a Simulation Study of Endogenous Time Dependent Covariates in JM versus in Extended Cox Model

Ebru Turgal[1] and Beyza Doğanay Erdoğan[2]

[1] Hitit University, School of Medicine, Department of Biostatistics, Çorum, Turkey
ebruturgal@gmail.com

[2] Ankara University, School of Medicine, Department of Biostatistics, Ankara, Turkey
beyzadoganay@gmail.com

Abstract. Beside lots of advantageous of joint modeling of longitudinal and time-to-event data, one of the usage area is to handle with endogenous (internal) time-dependent covariates. The extended Cox model assumes that the covariates are external, so for internals joint modeling should be used.

Time-dependent covariates may be internal or external (exogenous) covariates. An internal time dependent covariate relates directly to the patient and can only be measured when an individual is alive. Examples of internal variables in medical applications are biomarkers and clinical parameters, such as the serum bilirubin levels for patients with primary biliary cirrhosis, CD4 cell counts for HIV-infected patients, the prothrombin index for patient with liver cirrhosis, and aortic gradient level for patients with aortic stenosis. On the other hand, external time-dependent covariates are not related to the body conditions or status. For example; age, levels of air pollution, air temperature, etc.

We compare the results from extended cox regression and the new approach (JM) on AIDS dataset. The approaches are demonstrated via both simulation studies and the data set analysis. We carry out this simulation-based investigation under various different scenarios. Results from the simulation study and from the analysis will be presented. We will discuss scenarios in which case they are advantageous. All statistical analyses were conducted using SAS 9.0 (SAS Institute, Cary, NC) and R version 3.3.2 (R Foundation for Statistical Computing, Vienna, Austria).

Keywords

LONGITUDINAL DATA ANALYSIS, SIMULATION, SURVIVAL ANALYSIS.

References

ABDEL HAMID, H. (2012): *Flexible Parametric Survival Models with Time-Dependent Covariates for Right Censored Data (Doctoral dissertation, University of Southampton).*

ANDRINOPOULOU, E. R. (2014): *Joint Modelling of Longitudinal and Survival Data with Applications in Heart Valve Data. (Doctoral dissertation, Erasmus University).*

RIZOPOULOS D. (2012): *Joint Models for Longitudinal and Time-to-Event Data: with Applications in R.* Boca Raton: Chapman and Hall/CRC.

Discovery Of Ensemble Methods On A Real Delirium Data From Hospital Database

Merve Gulsah Ulusoy[1], Cigdem Dinckal[2] Pinar Tosun Tasar[3], Suriye Ozgur[4], Nur Ozge Akcam[5], Ozan Fatih Sarikaya[2], Aysin Noyan[5], and Soner Duman[2]

[1] Novagenix Bioanalytical Research and Development Centre, mgulsahulusoy@gmail.com
[2] Ege University Hospital, Department of Internal Medicine, sonerduman@hotmail.com
[3] Erzurum Training and Research Hospital,
[4] Ege University Hospital, Department of Biostatistics,
[5] Ege University Hospital, Department of Psychiatry, suozgur35@gmail.com

Abstract. Data mining is obtaining the valuable information that can be from the big data. In this way, it is possible to reveale the relationship between the data and when required the predictions for the future. By using data mining methods in health database it is possible to obtain information that may help hospital or health care facility to make predictions. It is necessary to apply data mining in health area due to the magnitude and vital importance of health data.

In this study decision tree algorithms of classification methods have been used. The aim of the study was comparing individual and ensemble decision tree algorithms. In accordance with this purpose individual algorithms (CART and C5) and ensemble algorithms (bagging, boosting and random forest) have been compared according to performance measures (accuracy, specificity etc.) obtained from confusion matrix. All algorithms have been performed by using R software. Hold-out validation method was used to validate the model. Therefore, dataset was splitted 70% and 30% for training set and test set, respectively.

The algorithms have been applied on a real dataset obtained from Ege University Faculty of Medicine. The dataset consists information of detailed clinical examination and evaluation of delirium. The outcome variable is the presence or absence of delirium. All of the patient cards (n=12962), have been evaluated by the psychiatrist and assessed for psychiatric disorders, that were reached between 2005-2013. Files were scanned and patients over 65 years old were chosen for the study.

Results showed that ensemble algorithms predicted better in all algorithms. Among ensemble algorithms bagging algorithm predicted the best with an accuracy of 85%.

Keywords

data mining, individual and ensemble trees, delirium, confusion matrix

Impact of outliers in Meta-Analysis

Mutlu Umaroglu[1] and Pinar Ozdemir[1]

Hacettepe University Department of Biostatistics, Turkey
mutlu.umaroglu@hacettepe.edu.tr

Abstract. Meta-analysis is a statistical method that combines the results from multiple independent studies. In meta-analysis, effect sizes are usually combined. Firstly, the effect size of studies must be calculated. After calculating effect size, homogenity of effect sizes must be examined. Cochrans Q test, I^2 or graphical methods can be used to asses heterogenity. Fixed effect model is used when studies are homogeneous otherwise random effect model is used. Assessing heterogenity is a critical issue in meta-analysis because chosen model may change the overall effect size. It should also be examined whether there is any outlier in studies using in meta-analysis. When outliers are included in meta-analysis, it is recommended that alternative models should be used instead of random effect model. If an outlier is included in a meta-analysis study, the weights should be reduced. Aim of this methods reduce the weight of outlier(s). One solution about overcoming outliers problem is using t-distribution instead of using normal distribution random effect model. If only one study is assumed to be an outlier, outlier can be deleted. But instead of deleting an outlier, mixture model is recommended as an alternative to random effect model.

Keywords

Meta-analysis, Outliers, Heterogenity

References

1. Ken J. Beath, A finite mixture method for outlier detection and robustness in meta-analysis *Research Synthesis Methods, 2014.*
2. Wolfgang Viechtbauer, Mike W.- L. Cheung, Outlier and influence diagnostics for meta-analysis *Research Synthesis Methods, 2010.*
3. Rose Baker, Dan Jackson, A new approach to outliers in meta-analysis *Health Care Management Science, 2009.*
4. Lifeng Lin, Haitao Chu, James S. Hodges, Alternative Measures of Between-Study Heterogeneity in Meta-Analysis: Reducing the Impact of Outlying Studies *Biometrics, 2016.*

Index

Ağraz M., 49
Abrahamowicz M., 3
Adesi G.B., 48
Adhikari S., 125
Alfò M., 4
Alioum A., 5
Andreis F., 52
Andrinopoulou E. R., 51
Ankarali H., 126
Antolini L., 6, 53
Ataoglu S., 127
Avramopoulos A., 128

Baart S., 54
Bakirarar B., 130
Bakoyannis G., 7
Balan T. A., 55
Bantis L.E., 56
Bartzis G., 57
Basol M., 58
Berghold A., 59
Bernasconi D. P., 8
Betensky R.A., 9
Birrell P., 10
Bofill M., 60
Bonetti M., 11
Bougioukas K., 63
Boulieri A., 62
Bouras E., 64, 129
Brewer M., 65
Bulut N., 142

Camli O., 66
Carstensen B., 12
Chalkidis N., 131
Chrysanthopoulou S. A., 67
Cribben I., 13
Cugnata F., 69

Dattner I., 70
Debusho L.K., 132
Demétrio C., 14
Demiris N., 15, 133

Edefonti V., 71, 72
Eleftheriou D., 73
Erar B., 74
Erler N. S., 75
Evangelopoulos D., 76

Farcomeni A., 78
Fernández-Fontelo A., 16
Finos L., 79
Fischer K., 80

Geroldinger A., 81
Giordani P., 17
Giussani A., 82
Goetghebeur E., 18
Goldberg Y., 86
Grigorova D., 83
Gutman R., 19

Hakli D. A., 134
Haneuse S., 21
Harezlak J., 20
Hayek S., 77
Heinze G., 22
Hinde J., 23
Hogan J., 24
Hong L., 84

Inan G., 85

Kakourou A., 87
Kalish M. M., 88
Kalogeropoulos K., 25
Kalpourtzi N., 89
Karlis D., 26

Keskin S., 90
Keskinoglu A., 136
Keskinoglu P., 135
Kipnis V., 27
Konar N. M., 91

Langohr K., 28
Li S., 29, 30

Manatunga A., 92
Mandel M., 93
Marciano L., 138
Mastrokostas A., 139
Mira A., 31
Molshatski N., 94
Moysiadis T., 95, 141
Murad H., 32

Nakas C. T., 96
Neuhäuser M., 97
Nikolakopoulos S., 98
Normand S.L.T, 1
Ntzoufras I., 33

Oman S. D., 99
Ozgur S., 100
Ozturk E., 101

Pagkalidou E., 143
Pandria N., 144
Papadimitropoulou K., 103
Papageorgiou G., 104
Paparoditis E., 34
Papaspiliopoulos O., 35
Pasin O., 102, 146
Pateras K., 145
Perez-Alvarez N., 36
Presanis A., 38
Protonotarios N., 105
Puig, P., 37

Rancoita P. M. V., 106
Rebora P., 107
Reiss P. T., 39
Rodopoulou S., 147
Rossell D., 40

Sachlas A., 109
Sangalli L. M., 41
Scheipl F., 42
Secilmis D., 110
Sertdemir Y., 111
Shaw P.A., 43
Shkedy Z., 113
Singh S. P., 114
Slavtchova-Bojkova M., 61
Solari A., 115

Talitman M., 116
Thomadakis C., 117
Trayanov P., 118
Trinca L. A., 149
Tsanousa A., 150
Tsonaka R., 119
Tudorascu D., 44
Turgal E., 151

Ulusoy M. G., 152
Umaroglu M., 153
Unal I., 120

van Rosmalen J., 108
Vasdekis V., 121

Wang K., 122

Yiannoutsos C. T., 45

Zhang Y., 46
Zygoura P., 123

About this book

This book of abstracts was based on the abstracts submitted. For all abstracts not in the LaTeX format provided, we tried to reproduce the abstracts as close as possible by retyping them, taking care of equations but not changing any grammar issue. We also tried to homogenize them as possible. We apologize to the authors if the transfer from other editors to LaTeX led to any change in their abstract.

NINETH CONFERENCE OF THE EASTERN MEDITERRANEAN REGION AND THE ITALIAN REGION OF THE INTERNATIONAL BIOMETRIC SOCIETY

9th EMR-IBS and Italian Region Conference

8-12 May 2017
Thessaloniki, Greece

The meeting is devoted to the memory of Prof Marvin Zelen

http://stat-athens.aueb.gr/~emribs/page/emr2017.html

Thessaloniki, 8-12 May 2017