Found 9 Records

CONTROL ID: 3340772
PRESENTER: Yiwen Wang
PRESENTER (INSTITUTION ONLY): School of Mathematics and Statistics, University of Melbourne
TITLE: A latent component based multivariate method to correct for batch effects in microbiome data
ABSTRACT BODY:
Abstract Body: In the past years, microbial research has made enormous progress with the advent of sequencing technologies to investigate the roles of all microorganisms in different ecological habitats. However, microbiome studies (based on 16S amplicon and shotgun sequencing) are difficult to replicate as they may suffer from different sources of batch effects. In this context, we define batch effect as any unwanted source of variation that is unrelated to, but obscures the biological factor of interest. Such batch effects range from biological, technical to computational factors. Batch effect correction is challenging in microbiome data because of the inherent characteristics of data, including sparsity, overdispersion, uneven library sizes, correlation between variables, compositional nature and small sample size. Thus, traditional statistical methods developed for microarray or RNA-seq data are not suitable in this context.

We propose two novel computational methods PLSDA-batch and sPLSDA-batch based on Partial Least Squares Discriminant Analysis regression to address some of these challenges. Data are transformed with centered log ratio to account for uneven library sizes and compositional constraints. We estimate latent components associated first with the outcome, then with batch effects. Using deflation, the batch effect is then subtracted from the original data while the effects of interest remain. The variant sPLSDA-batch includes lasso penalisation to select relevant microbial features. Both methods are non-parametric and can handle the skewed distribution caused by sparsity and overdispersion. Their multivariate property accounts for the data correlation structure.

On both simulated and real microbiome data, we showed that our approaches are superior in removing batch variation compared with existing methods such as ComBat, removeBatchEffect and Remove Unwanted Variation III (from the sva, limma, ruv packages). PLSDA-batch preserves more variation associated with the outcome, while sPLSDA-batch removes batch effects and selects relevant discriminative microbial variables.

Our proposed methods are useful alternatives to address batch effects in microbiome data and will be available through an R package.
AUTHORS/INSTITUTIONS: Y. Wang, K. Lê Cao, Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, AUSTRALIA|

CONTROL ID: 3343813

PRESENTER: Vicente Núñez-Antón

PRESENTER (INSTITUTION ONLY): University of the Basque Country UPV/EHU

TITLE: Alternative Structured Antedependence Bayesian Modelling Proposals for Longitudinal Data

ABSTRACT BODY:

Abstract Body: An important problem in Statistics is the study of longitudinal data taking into account the effect of other explanatory variables such as treatments and time and, at the same time, incorporate into the model the time dependence between observations on the same individual. The latter is especially relevant in the case of having nonstationary correlations, as well as nonconstant variances for the different time points at which measurements are taken. Structured Antedependence (SAD) models constitute a well known commonly used set of models that can accommodate this behavior. These covariance models can include too many parameters and estimation can be a complicated optimization problem requiring the use of complex algorithms and programming. In this paper, a new Bayesian approach for analyzing longitudinal data within the context of structured antedependence models is proposed. This innovative approach takes into account the possibility of having nonstationary correlations and variances, and proposes a robust and computationally efficient estimation method for this type of data. We consider the joint modelling of the mean and covariance structures for the general structured antedependence (SAD) model, estimating their parameters in a longitudinal data context. Our Bayesian approach is based on a generalization of the Gibbs sampling and Metropolis-Hastings by blocks algorithm, properly adapted to the SAD models longitudinal data settings. Finally, we illustrate the proposed methodology by analyzing several examples where SAD models have been shown to be useful: the small mice, the speech recognition and the race data sets.

AUTHORS/INSTITUTIONS: V. Núñez-Antón, Department of Econometrics and Statistics (A.E. III), University of the Basque Country UPV/EHU, Bilbao, Bizkaia, SPAIN|E. Cepeda-Cuervo, Department of Statistics, National University of Colombia, Bogotá, COLOMBIA|

CONTROL ID: 3356278
PRESENTER: Andrew Yiu
PRESENTER (INSTITUTION ONLY): University of Cambridge
TITLE: A Bayesian framework for case-cohort Cox regression
ABSTRACT BODY:
Abstract Body: The case-cohort study design (Prentice, 1986) is an increasingly popular approach for collecting cohort data. The infeasibility of obtaining certain covariates on the full cohort is circumvented by restricting the measurements to a randomly sampled subcohort, along with all remaining incident cases. This may be necessitated by time and cost constraints, as well as concerns over the wastage of valuable biological material. Existing proposals for analyzing case-cohort data are largely based on the Cox proportional hazards model. Weighted Cox regression approaches are the current norm in practice, motivated by the intuition that the oversampling of cases can be balanced by an appropriate overweighting of the subcohort controls. However, the data are used inefficiently, and inverse probability weights can be unstable. Approaches based on multiple imputation and nonparametric maximum likelihood have been proposed to address this, but suffer from incompatibility and computational issues respectively. We introduce a novel Bayesian framework for case-cohort Cox regression which avoids the aforementioned problems. Posterior sampling is carried out in two stages, where samples from the first stage serve as inputs in a pseudo-marginal MCMC algorithm (Andrieu & Roberts, 2009). The model for the baseline cumulative hazard function is nonparametrically specified and integrated out, allowing for only the log hazard ratio to be sampled. We illustrate the methodology with simulations, and apply it to the EPIC-Norfolk study to investigate risk factors for type-2 diabetes.
AUTHORS/INSTITUTIONS: A. Yiu, R. Goudie, B.D. Tom, MRC Biostatistics Unit, University of Cambridge, Cambridge, UNITED KINGDOM|

CONTROL ID: 3364960

PRESENTER: Ali Mahmoudi

PRESENTER (INSTITUTION ONLY): The University of Melbourne, The University of Melbourne

TITLE: A probabilistic model to infer the ancestral recombination graph

ABSTRACT BODY:

Abstract Body: A challenging problem in population genetics is to infer the full genealogical history of a sample of DNA sequences—otherwise known as the Ancestral Recombination Graph (ARG)—under the coalescent with recombination. Inferring the ARG remains a problem since, for even a small number of DNA sequences, the state space of the ARG is large.

Many different methods have been proposed to perform the inference, however, most of them have been limited to small datasets. One reason that these methods are not efficient for large sample sizes is because of the way they store and represent the genealogies, i.e., the data structure. Previous methods use a data structure in which each marginal tree is stored separately. This leads to inefficiencies, as neighbouring trees in a genealogy share many parts. In order to gain efficiency and reduce processing time and storage capacity, taking these similarities into account is key.

In 2016, an efficient data structure known as Tree Sequence Recording (TS) was introduced by Kelleher, Etheridge, and McVean to store the genealogical trees at each site. In this method, identical parts of consecutive trees are stored only once. More recently, an inference method—tsinfer—was proposed to infer whole-genome genealogies. This method leverages the features of TS and is applicable to large data sets.

tsinfer infers the genealogical trees at each site, however, it is not a probabilistic inference model. Rather, it concentrates on compactly storing large datasets in a novel “evolutionary encoding” format that enables more efficient access and processing of the data.

In this work, we present a Markov chain Monte Carlo (MCMC) approach to perform probabilistic inference under the coalescent with recombination. Borrowing the idea of storing the genealogies with no repeated information from TS, we introduce a data structure to represent the full ARG.

Under the infinite sites mutation model, we infer the full ARG and, unlike tsinfer, our method infers both genealogical trees and event times. Hence, the time to the most common ancestor, the ancestral state at each time, and the total branch length are obtained.

We demonstrate the utility of our method by applying it to simulated datasets. Also, we compare our method with ARGweaver, the state-of-the-art probabilistic method.

AUTHORS/INSTITUTIONS: A. Mahmoudi, D. Balding, Y. Chan, School of Mathematics and Statistics, The University of Melbourne, Melbourne, Victoria, AUSTRALIA|A. Mahmoudi, D. Balding, Y. Chan, Melbourne Integrative Genomics, The University of Melbourne, Melbourne, Victoria, AUSTRALIA|D. Balding, School of BioSciences, The University of Melbourne, Melbourne, Victoria, AUSTRALIA|

CONTROL ID: 3367171

PRESENTER: Denis Rustand

PRESENTER (INSTITUTION ONLY): Bordeaux Population Health Center

TITLE: Joint modelling of a longitudinal semicontinuous biomarker and a terminal event.

ABSTRACT BODY:

Abstract Body: In cancer clinical trials, the sum of the longest diameter (SLD) of target lesions is a biomarker which reflects both the tumor burden and its evolution over time. Modelling this biomarker jointly with survival times improves the inference about treatment and prediction accuracy of the survival time. An excess of zero values and right skewness often characterize the SLD distribution. While a nonlinear transformation can easily handle the latter, the zero-inflation problem requires a more sophisticated approach. Left-censoring has been proposed (Król, A. et al. Biometrics 2016) as a way to handle the excess of zero values in a mixed-effects model (i.e. values below a detection limit are censored). Patients responding well to a treatment can reach the complete response state as defined by RECIST criteria, in which case the tumor size shrinks until reaching a 'true zero' value (i.e. not censored). We propose a two-part joint model that decomposes the distribution of the biomarker into a binary outcome (zero values vs. positive values) and a continuous outcome, both outcomes being modelled by a mixed effects regression model. Therefore, the binary part captures the effect of covariates on the probability of zero value of the biomarker. We propose two forms for the continuous part: the conditional form captures the effect of covariates on the expected value of the biomarker among positive values and the marginal form captures the effect of covariates on the marginal mean of the biomarker (Smith, V. A. et al. Stat. Med. 2014). The survival times are modelled using a Cox proportional hazards model with splines approximation for the baseline hazard. We propose several association structures to link the biomarker to the risk of event. We illustrate with simulation studies the performances of the models in terms of bias, accuracy and coverage probabilities when the model assumptions are misspecified, with different rates of zero excess. A real data application to a colorectal metastatic cancer trial comparing two treatment strategies is also proposed, showing how the biomarker evolution over time can bring additional information to evaluate the treatment effect on the risk of death. The zero inflation is a common problem in biomedical research, e.g. when quantifying exposure or measuring symptoms of a disease, and our proposed model is also relevant in this wide spectrum of applications.

AUTHORS/INSTITUTIONS: D. Rustand, V. Rondeau, Bordeaux Population Health Center, Bordeaux, FRANCE|L. Briollais, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, CANADA|L. Briollais, Biostatistics, Dalla Lana School of Public Health, Toronto, Ontario, CANADA|

CONTROL ID: 3367560

PRESENTER: WOOJOO LEE

PRESENTER (INSTITUTION ONLY): Inha University

TITLE: On the finite sample distribution of the likelihood ratio statistic for testing heterogeneity in meta-analysis

ABSTRACT BODY:

Abstract Body: In meta-analysis, hypothesis testing is one of the commonly used approaches for assessing whether heterogeneity exists in effects between studies.
The literature concluded that the Q-statistic is clearly the best choice and criticized the performance of the likelihood ratio test in terms of the type I error control and power.
However, all the criticism for the likelihood ratio test is based on the use of a mixture of two chi-square distributions with 0 and 1 degrees of freedom, which is justified only
asymptotically. In this study, we
develop a novel method to derive the finite sample distribution of the likelihood ratio test and restricted likelihood ratio test for testing the zero variance component
in the random effects model for meta-analysis.
We also extend this result to the heterogeneity test when meta-regression is applied.
A numerical study shows that the proposed statistics have superior performance to the Q-statistic,
especially when the number of studies collected for meta-analysis is small to moderate.

AUTHORS/INSTITUTIONS: W. LEE, S. Kuk, Inha University, Seoul, KOREA (THE REPUBLIC OF)|

CONTROL ID: 3367607
PRESENTER: Elisavet Syriopoulou
PRESENTER (INSTITUTION ONLY): University of Leicester
TITLE: Inverse probability weighting and doubly robust standardization in the relative survival framework
ABSTRACT BODY:
Abstract Body:
The event of interest in population-based cancer studies is usually death due to cancer. However, competing events that prevent the occurrence of the event of interest may be present. Relative survival is a commonly used measure that circumvents problems caused by inaccuracies in the cause of death information, incorporating other-cause mortality by matching cancer patients with individuals from the general population. Marginal estimates of relative survival summarise prognosis for a whole population, and contrasts of these (such as differences between subgroups) have a causal interpretation under certain assumptions. The causal inference literature in relative survival is scarce, with few application focussing on regression standardization that is the standard approach for obtaining marginal estimates in a modelling framework.

We propose two novel approaches to obtain marginal estimates of relative survival: i) inverse probability weighting (IPW) and ii) doubly robust standardization. In particular, we extend the IPW approach to use appropriate weights within the relative survival framework. With doubly robust standardization, a propensity score is first estimated as in the IPW approach. Then, the resulting weights are incorporated in a relative survival model alongside the exposure and all relevant covariates. With both methods, standard errors can be obtained by using either the delta method or M-estimation. The two methods outlined above as well as regression standardization are compared using a Monte Carlo simulation: we investigate the robustness of each approach to model misspecification, estimating both marginal estimates within exposed and unexposed as well as their difference.

Our results show that a higher degree of misspecification yields more biased estimates for regression standardization and IPW compared to doubly robust standardization, while the latter yields larger standard errors. All methods show less bias when the omitted variables are highly correlated with some of the variables included in the model. Finally, M-estimation tends to have better coverage compared to the delta method.

The methods we described can be extended to obtain several estimates of interest such as marginal all-cause survival and marginal crude probabilities, and can also be incorporated into mediation analysis approaches within the relative survival framework.
AUTHORS/INSTITUTIONS: E. Syriopoulou, M.J. Rutherford, P. Lambert, Health Sciences, University of Leicester, Leicester, UNITED KINGDOM|P. Lambert, Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, SWEDEN|

CONTROL ID: 3385362

PRESENTER: Khue-Dung Dang

PRESENTER (INSTITUTION ONLY): University of Technology Sydney, ARC Centre of Excellence for Mathematical & Statistical Frontiers

TITLE: Bayesian Structural Equations Modelling to characterize dose and pattern of adverse effects in the diagnosis of Fetal Alcohol Spectrum disorder

ABSTRACT BODY:

Abstract Body: While it is well known that high levels of prenatal alcohol exposure (PAE) result in serious cognitive and behavioural deficits in children, the exact nature of the dose response is less well understood. In particular, there is a pressing need to identify the levels of PAE associated with an increased risk of clinically significant adverse effects. To address this issue, data have been combined from six birth cohort studies in the United States which assessed PAE along with developmental outcomes measured throughout childhood. We argue that structural equation models (SEMs) provide the most appropriate way to capture the association between multiple observed outcomes, as well as to characterise the underlying variable of interest, cognition, and then relate this to PAE. Unfortunately, classic SEM software does not work well in our context, due to the variation in the outcomes being measured among the six studies as well as the complex nature of the variables. We show how a Bayesian approach can be used to fit a multilevel structural model that maps cognition to a broad range of observed variables that in turn map to several different subdomains and which also may be measured at multiple ages. The model adjusts for confounding through propensity scores. In comparison to more commonly used frequentist approaches, the Bayesian model allows us to incorporate expert knowledge into the model via weakly informative prior distributions. This approach is shown to have more stable performance and produce more reliable estimates than frequentist approaches.

AUTHORS/INSTITUTIONS: K. Dang, L. Ryan, School of Mathematical and Physical Sciences, University of Technology Sydney, Sydney, New South Wales, AUSTRALIA|K. Dang, L. Ryan, ARC Centre of Excellence for Mathematical & Statistical Frontiers, Sydney, New South Wales, AUSTRALIA|T. Akkaya-Hocagil, Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario, CANADA|J. Jacobson, S. Jacobson, Department of Psychiatry and Behavioral Neurosciences, Wayne State University, Detroit, Michigan, UNITED STATES|

CONTROL ID: 3392774
PRESENTER: Fiona Evans
PRESENTER (INSTITUTION ONLY): Murdoch University, Curtin University
TITLE: Analysis of data from large of on-farm experiments using Bayesian latent Gaussian models
ABSTRACT BODY:
Abstract Body: On-farm experimentation (OFE) has been proposed as a process to enhance adoption of digital technologies in agriculture, develop farmer competence and improve farm profitability. OFE is a farmer-centric process where farmers work with consultants or researchers to design and implement experiments to test management practices. Farmers use their own machinery to conduct large field-scale experiments. Data are usually measured by yield monitors that are standard in most modern harvesting machinery. However, data recorded by other types of sensors mounted on machinery, satellites or drones may also be used.

Agricultural experiments have historically been performed by researchers using small plot sizes and principles of experimental design to so that effects of treatments can be analysed while minimising or accounting for environmental effects. In contrast, analysis of data from large on-farm experiments is challenging because of large spatial variation in yield that: (1) is not due to the treatment being tested; (2) is influenced by spatial auto-correlation; and (3) is often much larger than the variation due to treatment effects. In addition, the relationship between treatment and yield may also vary spatially.

We investigate the use of Bayesian latent Gaussian modelling (BLGM) for analysis of data from large on-farm experiments. Latent Gaussian models combine random Gaussian effects (that may be fixed, structured or unstructured) into a linear predictor. We consider the case where the structured random effects include spatial dependency modelled as a Gaussian field with Matérn covariance. Bayesian inference is performed using integrated nested Laplace approximation. Examination of the posterior densities for the Matérn parameters shows they can be accurately estimated from the data using this approach.

We also consider the issue of model selection for BLGM; in particular determining whether treatment effects can be modelled globally or whether a spatially varying coefficient (SVC) model is required. We compare three approaches to model selection: (1) ad-hoc bandwidth based on experimental design; (2) k-fold cross validation; and (3) spatial k-fold cross-validation.
AUTHORS/INSTITUTIONS: F. Evans, S. Cook, Big Data in Agriculture, Murdoch University, Murdoch, Western Australia, AUSTRALIA|F. Evans, S. Cook, Centre for Digital Agriculture, Curtin University, Bentley, Western Australia, AUSTRALIA|Z. Cao, Statistics in the Australian Grains Industry (SAGI-West), Curtin University, Bentley, Western Australia, AUSTRALIA|