Found 9 Records

CONTROL ID: 3346908
PRESENTER: Kipoong Kim
PRESENTER (INSTITUTION ONLY): Pusan National University, Korea
TITLE: New statistical approach to identify pleiotropic variants associated with both quantitative and qualitative traits.
ABSTRACT BODY:
Abstract Body: In the last few decades, there have been numerous researches to identify genetic variants associated with a single trait. However, it has turned out that relatively many variants are associated with more than one trait. They are called pleiotropic variants. Identification of pleiotropic variants can play a critical role in understanding missing heritability because they have not been revealed by single phenotype association studies. However, most of existing statistical methods for pleiotropic variants are limited to only quantitative traits. In this work, we propose new statistical approach to identify pleiotropic variants which can be associated with quantitative traits only, qualitative traits only or both. The proposed approach is to unify multiple elastic-net regularization models where variants selected by individual elastic-net models are summarized as their selection probability. In our simulation studies, we demonstrated that the proposed approach can select more pleiotropic variants with a small or moderate effect size than single phenotype association. We also applied the proposed approach to cowpea genotype data with 18 different quantitative and qualitative traits. We could find potentially pleiotropic variants missed by a single association study.
AUTHORS/INSTITUTIONS: K. Kim, H. Sun, Statistics, Pusan National University, Korea, Busan, KOREA (THE REPUBLIC OF)|S. Wang, Biostatistics, Mailman School of Public Health, New York, New York, UNITED STATES|

CONTROL ID: 3367369
PRESENTER: Zhujie Gu
PRESENTER (INSTITUTION ONLY): University Medical Center Utrecht
TITLE: Statistical integration of methylation and IgG glycosylation data using Group Sparse O2PLS
ABSTRACT BODY:
Abstract Body: Summarizing multiple correlated omics datasets for dimension reduction and for interpretation is an open research topic. Various methods have been proposed for this purpose, such as PLS-related approaches, which decompose datasets into joint and residual parts. Omics data are heterogeneous (e.g. differences in source of variation, scale, dimensionality, etc.) and the joint parts estimated in PLS contain data-specific variations. O2PLS was proposed to capture the heterogeneity using data specific parts and better estimate the joint parts. However, the latent components spanning the joint subspace in O2PLS are linear combinations of all the variables, hampering interpretation. For better interpretation, variable selection is needed. To this end, we extend O2PLS to Group Sparse O2PLS (GSpO2PLS) which performs variable selection and incorporate group structures of variables.

Our motivating datasets are methylation (482,563 CpG sites) and IgG glycomics (22 glycan peaks) data from 646 samples in the TwinsUK study. IgG is an antibody whose functional diversity is mainly achieved by glycosylation. Methylation has an important role in the glycosylation pathways. We aim to identify groups of CpG sites that impact IgG glycosylation, and hence have influence on immune response.

To perform variable selection, L1 penalty is introduced on the loadings. Sparse solutions are obtained by retaining only variables with a large contribution to the covariance. GSpO2PLS imposes penalties on the sum of the group-wise L2 norms of loadings, which result in group-wise sparsity where variables of the same group are selected or dropped altogether based on the contribution to the covariance as a whole. If all the groups have size 1, the sum of group-wise L2 norms corresponds to L1 norm.

A simulation study shows that GSpO2PLS performs better than O2PLS in terms of variable selection and prediction, especially when group information is available. We apply GSpO2PLS to the motivating datasets, where we take CpG sites located around the same gene as a group. The genes corresponding to the selected regions of CpG sites are crucial for protein glycosylation and immune functions.

GSpO2PLS provides a framework to integrate two heterogeneous omics datasets and select relevant groups of variables, thereby facilitating interpretation.
AUTHORS/INSTITUTIONS: Z. Gu, S. El Bouhaddani, H. Uh, Biostatistics & Research Support, University Medical Center Utrecht, Utrecht, NETHERLANDS|J. Houwing-Duistermaat, Department of Statistics, University of Leeds, Leeds, UNITED KINGDOM|

CONTROL ID: 3373042
PRESENTER: Mirrelijn van Nee
PRESENTER (INSTITUTION ONLY): Amsterdam University Medical Centers
TITLE: Co-data learning in ridge models for high-dimensional data
ABSTRACT BODY:
Abstract Body: We consider generalised ridge regression in clinical prediction settings, in particular binary and survival, for high-dimensional data. We use complementary data ("co-data", e.g. related studies, genomic annotation or cell line data) to define possibly overlapping or hierarchical covariate groups (e.g. gene sets, known signatures, Gene Ontology trees) that may differ considerably in terms of predictive strength. If so, penalising these groups by different ridge penalties likely improves prediction.

We present an Empirical Bayes approach to estimate the group penalties. Here, we provide an extra level of shrinkage to obtain stable group parameter estimates and to account for structure of the co-data. Any type of shrinkage can be used at this level, rendering a new, flexible framework to improve predictions. Moreover, the framework allows for integration and weighting of multiple co-data sets, plus posterior variable selection.

We demonstrate the method on an application to cancer genomics, in which we combine various sources of co-data and shrinkage types of the group parameters. Besides, we compare predictive performance with other commonly used methods, such as group lasso, which account for one single grouping structure, but which are not able to both shrink and estimate multiple group penalties from multiple sources. We show that the multi-group penalties stabilise variable selection, and improve the performance of parsimonious prognostic models.
AUTHORS/INSTITUTIONS: M. van Nee, M. van de Wiel, Epidemiology & Biostatistics, Amsterdam University Medical Centers, Amsterdam, NETHERLANDS|M. van de Wiel, MRC Biostatistics Unit, Cambridge University, Cambridge, UNITED KINGDOM|

CONTROL ID: 3377253

PRESENTER: Youngjoo Cho

PRESENTER (INSTITUTION ONLY): The University of Texas at El Paso

TITLE: Causal effect estimation for competing risk data in randomized trial: adjusting high-dimensional covariates to gain efficiency

ABSTRACT BODY:

Abstract Body: The double blinded randomization trial is considered as the gold standard to estimate the average causal effect (ACE). It is known that the crude estimator without adjusting any covariate is consistent. However, in most cases, incorporating the information of covariates that are strong predictors of the outcome by adjusting them could reduce the issue of unbalance covariate distribution between treated and controlled group and can improve efficiency. Although the adjusted estimator for ACE and its properties have been well studied with low-dimensional setting with pre-specified parametric form of the covariate effect, it is unknown whether the same results hold when the dimension of covariates increase with the sample size and the form of the covariate effect is unknown. Recent work has shown that thanks to the randomization, for linear regression, an estimator under risk consistency (e.g., Random Forest) for the regression coefficients could maintain the $n^{1/2}$ convergence rate even when nonparametric model are assumed for the effect of high dimensional covariates. Also, such adjusted estimator will always lead to efficiency gain comparing to the crude unadjusted estimator. In this paper, we extend this result to the competing risk data setting and showed that under similar assumptions, the augmented inverse probability censoring weighting (AIPCW) based adjusted estimator has the same $n^{1/2}$ convergence rate and efficiency gain. Extensive simulation was performed to show the efficiency gain in the finite sample setting and a mimic clinical trial was presented to illustrate our method with adjusting for high-dimensional cytogenetic abnormality.

AUTHORS/INSTITUTIONS: Y. Cho, Mathematical Sciences, The University of Texas at El Paso, El Paso, Texas, UNITED STATES|C. Zheng, Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, UNITED STATES|M. Zhang, Biostatistics, Medical College of Wisconsin, Milwaukee, Wisconsin, UNITED STATES|

CONTROL ID: 3386088

PRESENTER: David Balding

PRESENTER (INSTITUTION ONLY): University of Melbourne, UCL Genetics Institute

TITLE: Where In The Human Genome Does Complex Trait Heritability Lie

ABSTRACT BODY:

Abstract Body: The advent of very large, richly-phenotyped and high-quality human genomics datasets, together with the development of models that allow joint analyses of all GWAS test statistics, have led to big advances in understanding the genomic architecture of complex traits. Analysis of summary statistics is effectively unlimited in sample size and number of genetic variants, and avoids confidentiality issues associated with individual-level data. However, the models rest on assumptions that were initially unchallenged but have since become the subject of controversy. A key focus has been to answer detailed questions about how much causal variation lies in specific genomic regions, particularly those with functional annotations. Different approaches have led to very different estimates, giving discordant pictures of the genomic architecture of complex traits. The problem has been different assumptions about how the causal effect of a SNP varies according to genomic properties known a priori, for example minor allele fraction and linkage disequilibrium. I will review recent progress in using genome-wide SNPs to assess the heritability of complex human traits and its distribution across the genome, as well as the effect of confounding on GWAS test statistics. We have also derived improved estimates of selection parameters, leading to new insights into the effects of purifying selection on various traits. I will describe a consensus position that we have recently reached, and our reasons for believing that we are close to robust answers about genomic architecture, even though there remain many avenues for model refinements.

AUTHORS/INSTITUTIONS: D. Balding, Melbourne Integrative Genomics, University of Melbourne, University of Melbourne, Victoria, AUSTRALIA|D. Balding, D. Speed, UCL Genetics Institute, London, UNITED KINGDOM|D. Speed, Aarhus Institute of Advanced Studies, Aarhus, DENMARK|

CONTROL ID: 3386128

PRESENTER: Boram Kim

PRESENTER (INSTITUTION ONLY): Seoul National University

TITLE: Pathway-based Integration of Metabolome and Microbiome Data

ABSTRACT BODY:

Abstract Body: It is clear that the human's microbiome is associated with various disease. Much work has been done to analyze microbiome data to identify microbiota related with specific disease. However, in the case of study using only microbiome data, even if we identify microbiota related to a particular disease, it is difficult to know the functional potential of the microbiome. The metabolomes produced from the microbial community are known to play a role in connecting host phenotype and microbiome function. Using both metabolomics and metagenomics has an advantage of understanding functional potentials of the microbiome and interactions with the host. However, integration of these two omics data remains a challenge, usually requiring a more advanced method. In this study, we proposed the hierarchical structural component model (HisCoM-MnM) that integrates microbiome and metabolome data. In particular, we used pathway information for integrate these two omics datasets to provide insight into biological interactions between different biological layers in relation to host phenotype. We applied our model to analyze real datasets generated from specific diseases. These real datasets were used to demonstrate whether our model is able to identify the pathways known to be related with disease. This analysis shows our HisCoM-MnM can identify disease related pathways from KEGG database and provide significant metabolomic and metagenomic components of pathway.

AUTHORS/INSTITUTIONS: B. Kim, Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|T. Park, Department of Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|

CONTROL ID: 3386218

PRESENTER: Johan Verbeeck

PRESENTER (INSTITUTION ONLY): UHasselt

TITLE:

Critical appraisal of Generalized Pairwise Comparison Methods to evaluate Composite Endpoints

ABSTRACT BODY:

Abstract Body: In many clinical areas, it is common to combine several clinically meaningful endpoints in a composite endpoints to evaluate the full clinical benefit of a treatment. The standard method of analysis for a composite endpoint is a time to first event analysis, which has several shortcomings. The results are often driven by lesser important components, it ignores information on subsequent events and only survival data can be combined. Hence, a time to first event analysis may still not evaluate the full treatment benefit.

Generalized pairwise comparison (GPC) analysis methods, such as the win ratio and net benefit, have been described that alleviate the shortcomings of the time to first event analysis.

Although the GPC methods are increasingly being used, their properties are not well understood. We aim to evaluate the properties of the GPC methods. Early criticism mentioned the dependence of the net benefit and win ratio on the censoring distribution. We found additionally that the break-down of the treatment effect on the individual components of the composite endpoint does not represent well the true treatment effect on the component. Hence, the prioritized GPC analysis should not be used to evaluate the treatment effect on individual components. When investigating unbiasedness, sufficiency and completeness of the GPC statistics, we conclude that the win ratio statistic is not an unbiased estimator, while the net benefit statistic is. In a univariate setting with complete data, sufficiency and completeness of the net benefit statistic depends on the distribution of the data. When the distribution for the observations belongs to an unrestricted family (Bernoulli), the net benefit statistic is the uniformly minimum-variance unbiased estimator (UMVUE), while it is not when observations come from a more restrictive family of distributions (normal, exponential or Poisson). Theory and simulations show that the loss in efficiency for the net benefit statistic is limited in realistic scenarios. Since, the underlying distribution of the observations is in practice often unknown, the net benefit statistic is the safer option.

In conclusion, the limitations of the GPC statistics in the presence of censoring and on the interpretation of individual component treatment effects should be recognized. Moreover, the net benefit has better theoretical properties than the win ratio.

AUTHORS/INSTITUTIONS: J. Verbeeck, UHasselt, Diepenbeek, BELGIUM|G. Molenberghs, I-BioStat, Hasselt University & KU Leuven, Hasselt, BELGIUM|

CONTROL ID: 3386320

PRESENTER: Jos Hageman

PRESENTER (INSTITUTION ONLY): Wageningen University

TITLE: Assessing the Usefulness of Data Fusion on Untargeted Metabolomics Data for Prediction Purposes

ABSTRACT BODY:

Abstract Body: Interest in the prediction of all kinds of properties, like sensory traits of foodstuff, from untargeted metabolomics data is high. To achieve the highest accuracy possible, there is a tendency to fuse data from multiple analytical platforms. However, there is no consensus on how to combine said data or what modelling techniques to use. Here, we evaluate the performance of random forests, stochastic gradient boosted machines, regularized regressors, support vector machines, partial least squares, as well as a more traditional method in the form of stepwise regression. In literature, three approaches to combine data sets are proposed, namely low, intermediate and high data fusion. An exploratory analysis of the various methods' performance across all datasets and data fusion levels is conducted. In addition, the effect of the data fusion strategies is statistically evaluated for each method and dataset. This evaluation is performed on 4 real, independent datasets. Using the Friedman, and Nemenyi post-hoc tests, we observe that regardless of method or data set, no fusion level significantly outperforms the single most informative data block. Furthermore, we note that no individual method consistently outperforms all others.

AUTHORS/INSTITUTIONS: J. Hageman, K. Mildau, H. Ehlers, F. van Eeuwijk, Applied Statistics, Wageningen University, Wageningen, NETHERLANDS|

CONTROL ID: 3393551

PRESENTER: Sarah Flora Jonas

PRESENTER (INSTITUTION ONLY): Institut Gustave Roussy, Inserm, CESP, Oncostat

TITLE: Estimating the difference in restricted mean survival time accounting for trial effect in individual-patient-data meta-analyses

ABSTRACT BODY:

Abstract Body: Summary

The difference in restricted mean survival times (RMSTs) between two treatment groups is a useful tool to provide information on the average causal treatment effect in a randomized clinical trial. This method is particularly appealing since it does not require a proportional hazards assumption, unlike the hazard ratio based on the Cox model. The RMST can be obtained by integrating under the survival curve up to a predetermined horizon. A particular concern in individual patient data meta-analyses (IPD-MA) is to properly account for the trial effect in the estimation of the difference in RMSTs.

Our objective is to estimate the difference in RMSTs in an IPD-MA using various propositions where the RMST is adjusted on covariates and includes a random trial effect.

Method(s): In our 1st proposition, we use the Breslow estimator as proposed by Zucker (JASA 1998) to have an estimation of the baseline hazard adjusted on the trials and on other covariates, and calculate the area under the curve. A variance estimator is obtained with an added term in the expression (Chen and Tsiatis Biometrics 2001) to remove the conditioning on the trial. The 2nd proposition is based on a general linear mixed model which includes a random trial effect. A Poisson regression allows the estimation of the survival baseline hazard. Finally, the 3rd proposition is based on a Breslow estimator which takes into account the random trial effect (Gorfine et al. Biometrika 2006). An IPD-MA from the GASTRIC (JAMA 2010) is used for illustration.

Results: A simulation study has been set up generating Weibull data with random trial and treatment-by-trial effects. A true “RMST” is calculated through a double integration of the parametric survival function; and the bias and variance of our estimation methods are evaluated with respect to this value. We also plan to make simulations under a non-proportional hazard hypothesis. The IPD-MA included 3288 patients with resectable gastric cancer from 14 randomized trials of adjuvant chemotherapy versus surgery alone. At the 10y [20y] horizon, the estimated difference in RMSTs is 141[354] days (SE 45 [104] days; p=0.002 [p<0.001]) between the chemotherapy and surgery alone groups.

Conclusions: We proposed estimation methods for the difference of RMSTs between two treatment groups in the IPD meta-analyses context.

AUTHORS/INSTITUTIONS: S.F. Jonas, S. Michiels, Institut Gustave Roussy, Villejuif, FRANCE|S.F. Jonas, S. Michiels, Inserm, CESP, Oncostat, Villejuif, FRANCE|D.M. Zucker, Dept. of Statistics and Data Science, The Hebrew University of Jerusalem, Jerusalem, ISRAEL|