Found 10 Records

CONTROL ID: 3276323
PRESENTER: MinJae Lee
PRESENTER (INSTITUTION ONLY): U of Texas McGovern Medical School
TITLE: A latent class based imputation under Bayesian quantile regression framework for longitudinal medication usage data with missing values
ABSTRACT BODY:
Abstract Body: Evaluating the association between diseases and the longitudinal pattern of pharmacological therapy has become increasingly important. However, in many longitudinal studies, self-reported medication usage data collected at patients' follow up visits could be missing for various reasons. These pieces of missing or inaccurate/untenable information complicate determining the trajectory of medication use and its complete effects for patients. Although longitudinal models can deal with specific types of missing data, inappropriate handling of this issue can lead to a biased estimation of regression parameters especially when missing data mechanisms are complex and depend upon multiple sources of variation. We propose a latent class based multiple imputation approach using a Bayesian quantile regression that incorporates cluster of unobserved heterogeneity for medication usage data with missing values. Findings from our simulation study indicate that the proposed method performs better than traditional imputation methods under certain scenarios of data distribution. We also demonstrate applications of the proposed method to real data obtained from the longitudinal cohort study that assesses a trajectory of medication usage and its association with disease progression, while self-reported medication usage data are incomplete during follow-up in the cohort.
AUTHORS/INSTITUTIONS: M. Lee, U of Texas McGovern Medical School, Houston, Texas, UNITED STATES|

CONTROL ID: 3338696
PRESENTER: Elizabeth Alison Thompson
PRESENTER (INSTITUTION ONLY): University of Washington
TITLE: A century of genotypic correlations and relatedness between individuals
ABSTRACT BODY:
Abstract Body: Almost 100 years ago, Sewell Wright gave the connection between allelic correlation and identity by descent (IBD) at a single locus in an idealized population. For many years the data did not allow for direct studies, and the only measures of IBD were expectations based on known pedigree structures, but it was already clear that the connection between genotypic correlation and IBD is weak when considering pairs of individuals beyond close family relationships.

The ease of computing pairwise genotypic correlations with modern SNP data, and the clarity of the basic formulae of Wright have led to wide use of the matrix of realized pairwise genotypic correlations (the GRM) being used as a measure of pairwise realized "relatedness" in large population samples from which apparent close relatives are often removed. Many forms of the GRM, involving weighting or other possible kernel-based adjustments, have been successfully used in studies of quantitative variation and in phenotypic prediction. As such these approaches are best viewed as regression approaches with random or fixed effects, and a very large number of predictors, rather than as deriving from ancestral identity by descent. The GRM can also provide measures of relatedness and structure among populations with divergent allele frequencies. We show that covariances of genotypes provide good measures of population-level relatedness if multiple samples are available from populations that have diverged due to random genetic drift.

However, the GRM is a poor estimate of realized IBD between pairs of individuals, and realized genome-wide IBD varies substantially about its pedigree expectation. We show that attempts to connect the GRM to individual relatedness and particularly to remote pedigree relationships are misguided. On the other hand SNP data provide location-specific information on segments of genome shared by descent. In remote relatives, segments of IBD are rare but not short. Although many relatives will share no IBD DNA, those segments of IBD that exist can be detected from SNP data. We show that location-specific inference of IBD segments jointly among relatives can provide information to resolve genetic traits, to detect regions of the genome subject to selection, and to model processes that go beyond a simple additive model.
AUTHORS/INSTITUTIONS: E.A. Thompson, Department of Statistics, University of Washington, Seattle, Washington, UNITED STATES|

CONTROL ID: 3351283
PRESENTER: Ana Gabriela Pereira Vasconcelos
PRESENTER (INSTITUTION ONLY): Universidade de São Paulo
TITLE: Integration of heterogeneous data: a multi-omics application
ABSTRACT BODY:
Abstract Body: Nowadays, a huge amount of data is being collected all the time, so high-dimension databases are becoming very common to encounter. More specifically, with the advance of technology many biological information are now available at low costs -- data from genome, miRNA, mRNA, gene expression, protein, methylation, lipids, metabolism, phenotypes and so on. Many different studies have been done separately with each type of data, but more recently there is an increasingly interest in combine different data to gather more information. However, many classical methodologies used to this end assume the data matrix to be completed and numerical. Therefore, the heterogeneity of dataset with different variable types is not considered.

Alternatively, the Generalized Low Rank Models (GLRM) is a tool capable of dealing with large datasets of heterogeneous data. It is used for a single database, but can handle abstract data by using different loss functions, adequate to each variable type. GLRM is a very powerful tool that can deal with many different problems, but it is very recent, so its potential to work with multi-omics is yet unknown. In spite of this, some initial simulation studies showed that GLRM can deal with omics data well.

Thus, in this work, considering multiomics applications, we explore the possibilities of the GLRM for dimensionality reduction and prediction. Also, the latent structure of the problem is compared with other consolidated techniques. To allow a latent structure selection between different techniques, a model stability metric is implemented. Furthermore, an expansion for studies based on family data is proposed.

Keywords: multi-omics, generalized low rank models, matrix factorization, multivariate analysis, latent structure.
Acknowledgment: To CAPES (Brazil, Finance code 001) and FAPESP (Brazil, Process 2017/051257) for partial financial support.
AUTHORS/INSTITUTIONS: A.P. Vasconcelos, J.M. Pavan Soler, Statistics, Universidade de São Paulo, São Paulo, SP, BRAZIL|

CONTROL ID: 3365371

PRESENTER: Gregory James Hunt

PRESENTER (INSTITUTION ONLY): William & Mary

TITLE: The Role of Scale in the Estimation of Cell-type proportions

ABSTRACT BODY:

Abstract Body: Complex tissues are composed of a large number of different types of cells, each involved in a multitude of biological processes. Consequently, an important component to understanding such processes is understanding the cell-type composition of the tissues. Estimating cell type composition using high-throughput gene expression data is known as cell-type deconvolution. In this talk, we first summarize the extensive deconvolution literature by identifying a common regression-like approach to deconvolution. We call this approach the Unified Deconvolution-as-Regression (UDAR) framework. While methods that fall under this framework all use a similar model, they fit using data on different scales. Two popular scales for gene expression data are logarithmic and linear. Unfortunately, each of these scales has problems in the UDAR framework. Using log-scale gene expressions proposes a biologically implausible model and using linear-scale gene expressions will lead to statistically inefficient estimators. To overcome these problems, we propose a new approach for cell-type deconvolution that works on a hybrid of the two scales. This new approach is biologically plausible and improves statistical efficiency. We compare the hybrid approach to other methods on simulations as well as a collection of eleven real benchmark datasets. Here, we find the hybrid approach to be accurate and robust.

AUTHORS/INSTITUTIONS: G.J. Hunt, Mathematics, William & Mary, Williamsburg, Virginia, UNITED STATES|J.A. Gagnon-Bartsch, Statistics, University of Michigan, Ann Arbor, Michigan, UNITED STATES|

CONTROL ID: 3366758
PRESENTER: Jeong Hoon Jang
PRESENTER (INSTITUTION ONLY): Indiana University School of Medicine
TITLE: Diagnostic Evaluation of Quantitative Features of Functional Markers
ABSTRACT BODY:
Abstract Body: With modern technology development, more and more diagnostic markers are being collected as functional data (i.e., functional markers). Each sample element of a functional marker is a smooth, continuous curve, whose dynamic structure over a time or space domain is a rich source of clinical information. In many clinical practices, it is standard to describe and diagnose a disease using a set of ``quantitative features'' that characterizes various dynamic, interpretable patterns of a functional marker, such as area under the curve, maximum value and time to reach maximum. Here, we present a novel statistical framework for evaluating the diagnostic accuracy of quantitative features using the area under the receiver operating characteristic curve (AUC). Based on a class of summary functionals that flexibly represents various quantitative features, we develop a two-stage non-parametric AUC estimator that addresses discreteness and noise in functional data and establish its asymptotic properties. To describe the heterogeneity of AUC in different subpopulations, we propose a sensible adaptation of a semi-parametric regression model, whose parameters can be estimated by the proposed estimating equations. We also propose an automated data-driven approach for trading off between bias and efficiency of the regression coefficient estimates when continuous covariates are considered. We demonstrate the application of our methods using a renal study.
AUTHORS/INSTITUTIONS: J. Jang, Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, UNITED STATES|A. Manatunga, Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, UNITED STATES|

CONTROL ID: 3367452

PRESENTER: Emily Getzen

PRESENTER (INSTITUTION ONLY): University of Pennsylvania

TITLE: Methods for Principal Component Analysis with Missing Data

ABSTRACT BODY:

Abstract Body:

Digital phenotyping studies involve collecting data from an individual's smartphone (such as GPS, call/text logs, surveys) to provide information relevant to psychiatric disorders and other illnesses. These data frequently contain too many explanatory variables, and principal component analysis (PCA) proves to be a useful dimension reduction tool. However, missingness in digital phenotyping data leads to problems with PCA and other downstream analyses. Existing methods for PCA in the presence of missingness rely on imputation, but can have trouble with reduced variance in the features as well as with interpretation of the principal components. We introduce a non-imputation based approach contingent on the pairwise empirical correlation matrix. We show in simulations how our method performs compared to imputation methods under varying correlation structures and proportions of missingness. We also compare competing methods using a digital phenotyping study of schizophrenia.

AUTHORS/INSTITUTIONS: E. Getzen, I. Barnett, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES|

CONTROL ID: 3367972

PRESENTER: Stanley B Pounds

PRESENTER (INSTITUTION ONLY): St. Jude Children's Research Hospital

TITLE: Bootstrap of Association Matrices (BAM): A Robust Method for Integrated Analysis of Multiple Omic Data Sets with Multiple Clinical Outcomes

ABSTRACT BODY:

Abstract Body: Modern biomedical research studies routinely collect multiple matrices of data on thousands that each contain thousands to millions of variables in order to explore the associations of these data matrices with multiple pharmacologic and/or clinical outcomes. For example, in oncology research, there may be a data matrix for each of several types of molecular data such as DNA genotype, DNA abnormalities, or RNA expression levels. The objective is to identify genes for which one or more of these molecular data forms are robustly associated with one ore more clinical characteristics or outcomes such as risk of disease development, disease histology, response to chemotherapy, time to relapse, and time to death.

Here, bootstrap of association matrices (BAM) is introduced as a robust method that can be applied to a broad class of these types of research studies. First, BAM computes estimates of a matrix of association parameters for each clinical outcome with each molecular variable for the observed data set. Next, BAM computes a series of bootstrap estimates of this association matrix for each of many bootstrap data sets obtained by resampling subjects with replacement. This bootstrap procedure produces a cloud of points in multivariate space; each point representing the association vector for one bootstrap. Finally, for each gene with a predefined association submatrix of scientific interest, BAM uses a recursive peeling algorithm to quantify statistical significance by characterizing the position of the origin (null) relative to the point cloud of bootstrap association estimates.

BAM possesses several practical advantages over other widely used methods for integration of multi-omic data with multi-endpoint data. BAM provides a more elegant approach to adjust for clinically relevant covariates than do widely used permutation methods. BAM does not require specification and evaluation of priors like Bayesian methods or specification of data integration weights like projection onto the most interesting statistical evidence.

The performance of BAM will be evaluated and compared to that of other methods in simulation studies and in the analysis of example data sets from pediatric cancer research. BAM will be made freely available as an R package on GitHub and/or CRAN.

AUTHORS/INSTITUTIONS: S.B. Pounds, L. Shi, Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, UNITED STATES|X. Cao, Preventative Medicine, University of Tennessee Health Science Center, Memphis, Tennessee, UNITED STATES|

CONTROL ID: 3386029

PRESENTER: Qiwei Li

PRESENTER (INSTITUTION ONLY): The University of Texas at Dallas

TITLE: Bayesian Modeling of Metagenomics Sequencing Data for Discovering Microbial Biomarkers in Colorectal Cancer

ABSTRACT BODY:

Abstract Body: Colorectal cancer (CRC) is a major cause of morbidity and mortality globally. It is known that CRC survival is highly dependent upon the stage of disease at diagnosis. Reductions in mortality can be achieved through the detection and treatment of early-stage CRC patients. Optical colonoscopy, which allows for direct visualization of colonic polyps, is currently the most effective CRC screening test in nowadays. However, it is costly, invasive, and requires anesthesia, which discourages many people from pursuing routine tests. A simple noninvasive test with high accuracy for CRC is urgently needed, as it might increase adherence rates, resulting in better clinical outcomes. Several recent CRC studies have demonstrated a significant association between tumorigenesis and abnormalities in the microbial community. Those findings shed light on utilizing microbial taxa as noninvasive CRC biomarkers. In this paper, we propose a Bayesian hierarchical framework to identify a set of differentially abundant taxa, which could potentially serve as microbial biomarkers. The bottom level is a multivariate count generative model that links the observed counts in each sample to their latent normalized abundances. For the choice of a zero-inflated negative binomial model as the bottom level, we use the Dirichlet process as a flexible nonparametric mixing distribution to model all unknown factors that could influence sequencing depth. The top level is a Gaussian mixture model with a feature selection scheme for identifying those taxa whose normalized abundances are discriminatory between different phenotypes. The model further employs Markov random field priors to incorporate phylogenetic tree information to identify microbial biomarkers at different taxonomic ranks. A simulation study on both simulated and synthetic data is conducted. A CRC case study demonstrates that a resulting diagnostic model trained by the microbial signatures identified by our model in a CRC cohort can significantly improve the current predictive performance in another independent CRC cohort.

AUTHORS/INSTITUTIONS: Q. Li, Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, Texas, UNITED STATES|S. Jiang, Department of Statistical Science, Southern Methodist University, Dallas, Texas, UNITED STATES|A.Y. Koh, G. Xiao, Y. Xie, X. Zhan, The University of Texas Southwestern Medical Center, Dallas, Texas, UNITED STATES|

CONTROL ID: 3386521
PRESENTER: Nabihah Tayob
PRESENTER (INSTITUTION ONLY): Dana-Farber Cancer Institute, Harvard Medical School
TITLE: Longitudinal biomarker screening algorithms for early detection of hepatocellular carcinoma
ABSTRACT BODY:
Abstract Body: Advanced hepatocellular carcinoma (HCC) has limited treatment options and poor survival. Early detection of HCC is critical to improve the prognosis of these patients. Current guidelines for high-risk patients include six-month ultrasound screenings but these are not sensitive for early HCC. Alpha-fetoprotein (AFP) is a widely used diagnostic biomarker but has shown limited use in HCC screening with a fixed threshold.

Approaches that incorporate longitudinal AFP have shown potentially increased earlier detection of HCC. A parametric empirical Bayes algorithm, first proposed by McIntosh and Urban (2003), defines a patient-specific threshold that is a weighted average of the population mean and the sample mean of the patient screening history. The PEB algorithm has been applied to HCC screening with AFP in data from the Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) trial, a randomized clinical trial, as well as a cohort constructed from the electronic medical records from the Department of Veteran's Affairs Hepatitis C Clinical Case Registry. In both studies, we observed gains in the discriminatory performance of AFP via the PEB algorithm.

However, in most cancer settings, including HCC, a single biomarker will not cover the heterogeneous subtypes present in the target surveillance population. Des-gamma-carboxy prothrombin (DCP) is a serum biomarker that has been evaluated in Phase-2 biomarker studies and approved by the FDA for HCC risk. We have developed a multivariate parametric empirical Bayes (mPEB) algorithm that identifies optimal patient-specific thresholds for a panel of biomarkers. A minimal model for the biomarker levels in both cases and controls is specified. The optimal set of patient-specific thresholds ensures the false positive rate is at most f_0 while maximizing the likelihood of a positive screen if the patient has HCC, conditional on the patient screening history. The mPEB algorithm was compared to alternatives in the HALT-C Trial (using cross-validation) and in simulations studies under a variety of possible scenarios for biomarker trajectories.
AUTHORS/INSTITUTIONS: N. Tayob, Data Science, Dana-Farber Cancer Institute, Boston, Massachusetts, UNITED STATES|A.S. Lok, Internal Medicine, University of Michigan, Ann Arbor, Michigan, UNITED STATES|Z. Feng, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, UNITED STATES|N. Tayob, Harvard Medical School, Boston, Massachusetts, UNITED STATES|

CONTROL ID: 3386672

PRESENTER: Shelley B Bull

PRESENTER (INSTITUTION ONLY): Lunenfeld-Tanenbaum Research Institute, Sinai Health System, University of Toronto

TITLE: Single-region and multiple-region testing of rare-variant association in affected sibling pairs

ABSTRACT BODY:

Abstract Body: Next generation sequencing technologies have made it possible to investigate the role of rare variants (RVs) in disease etiology. Because RVs associated with disease susceptibility tend to be enriched in families with affected individuals, study designs based on ascertainment of affected sib pairs (ASP) can be more powerful than conventional case-control studies of unrelated individuals. We construct tests of RV-set association in ASPs for single genomic regions as well as for multiple regions. Single-region tests can efficiently detect a genomic region harboring susceptibility variants, while multiple-region extensions are meant to capture signals dispersed across a biological pathway, potentially as a result of locus heterogeneity. Within ascertained ASPs, the test statistics contrast the frequencies of duplicate rare alleles (usually appearing on a shared haplotype) against frequencies of a single rare allele copy (appearing on non-shared haplotypes); we call these allelic parity tests. Incorporation of minor allele frequency estimates from reference populations can markedly improve test efficiency. Under various genetic penetrance models, application of these tests in simulated ASP datasets demonstrates good type I error properties as well as power gains over approaches that regress ASP rare allele counts on sharing state. The improved performance of the allelic parity tests can be explained by the fact that allele parity counting (ie. whether alleles appear as singles or duplicates) is a better discriminator between susceptibility and null regions at the sib-pair level. As proof of principle, we apply gene-based and pathway-based tests to an established DNA pathway in a dataset derived from whole exome sequencing of sisters ascertained with early onset breast cancer. Considering practical issues, we address robustness of the allelic parity methods to sequencing error, the presence of genetic linkage, population stratification, and misspecification of reference population allele frequencies.

AUTHORS/INSTITUTIONS: R. Romanescu, S.B. Bull, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, CANADA|S.B. Bull, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, CANADA|