Abstracts

Integration of anterior samples information and modeling in repeated survey with complex rotation pattern: an illustration with CPS
Daniel Bonnéry 2

Linking analysis of multiple datasets, and entity resolution, using probabilistic matrix factorization
Snigdhansu Chatterjee 3

Population synthesis for transportation microsimulation modeling
Cinzia Cirillo 4

Synthetic datasets for statistical disclosure control – a short introduction
Jörg Drechsler 5

Integration of multiple sources of data by the National Agricultural Statistics Service
Malay Ghosh 6

Big data in surveys: challenges, opportunities, and strategies
Jiming Jiang 7

Bayesian model based IntegrOmics
Bani K. Mallick 8

Finding hidden code of omics data to disentangle the genetic architecture of disease
Indranil Mukhopadhyay 9

Public health data integration
Takumi Saegusa 10

Entity resolution with societal impacts in machine learning
Rebecca C. Steorts 11

Variable selection in small area level models subject to sampling errors
Jiraphan Suntornchost 12

Integration of anterior samples information and modeling in repeated survey with complex rotation pattern: an illustration with CPS

Daniel Bonnéry

University of Maryland, USA

ABSTRACT

Longitudinal studies on a fixed panel allow to obtain estimation of changes of evolution of population characteristics of over time with great precision. Successive samples in longitudinal studies with a complex rotating do not necessarily overlap. For example, the Current Population Survey (CPS), which is sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS) and is the primary source of labor force statistics in the United States, uses a 4-8-4 rotating panel design: sampled household are in the sample four consecutive months two times at a one year interval. In this presentation, we show how to integrate all the information provided by the unit-level unemployment status time series to predict the population employment rate time series, and improve efficiency both for level and month to month change prediction. We will present simulation results on a design that mimics the CPS.

Linking analysis of multiple datasets, and entity resolution, using probabilistic matrix factorization

Snigdhansu Chatterjee

University of Minnesota, USA

ABSTRACT

We propose using probabilistic matrix factorization, which typically results in modeling a data matrix in terms of low-rank, sparse and noise terms, for entity resolution or record linkage problems. Often entity resolution is an intermediate step to a simultaneous analysis of multiple datasets, and our proposal facilitates such analysis.

Population synthesis for transportation microsimulation modeling

Cinzia Cirillo

University of Maryland, USA

ABSTRACT

The latest generation of large scale transportation models uses microsimulation to reproduce travel behavior of individuals and households. Given the impossibility to gather data from each single unit in the study area, a synthetic population approach is used to create agents.

In the field of transportation, the most prevalent methods have been based on numerical approximation techniques, specifically the Iterative Proportional Fitting (IPF). IPF essentially reweighs counts from a sample until the combined weight matches the constraints imposed by the aggregate statistics of the area of interest.

The talk presents an innovative approach using copulas to infer and capture the dependence structure between characteristics of people from sample survey data. The dependence structure can then be used to produce synthetic populations that conform to different marginal distributions with the same underlying dependence structure. The American Community Survey (ACS) of 2016 is used as the sample survey data, while the 2010 census data provides the marginals at the census tract level. Copulas are fit to the ACS at the Public Use Micro Area (PUMA) level. Cramér–von Mises distance is computed using the sample data to test the null hypothesis that the dependence structure belongs to the assumed parametric family of copulas. Parametric bootstrap is used to evaluate the Cramér–von Mises statistic and to compute p-values.

The concluding part will be about a real case study where ACS, IRS data, and data from the National Household Travel Survey are combined to estimate low income population at a small geographical level and to calculate the probability of not owning a car.

Synthetic datasets for statistical disclosure control – a short introduction

Jörg Drechsler

Institute for Employment Research (IAB) of the German Federal Employment Agency (BA), Germany

ABSTRACT

With the ever increasing amount of data that is collected and disseminated and the increased risk of re-identification that comes with this, traditional statistical disclosure control techniques such as swapping or top-coding might no longer be sufficient to adequately protect the data. Besides, applying these methods with great intensity can have severe deteriorating effects on the analytical validity of the protected data. In this context, the dissemination of synthetic data received increasing attention in recent years as an alternative that potentially preserves the analytical validity of the data while offering a high level of data protection. With this approach first proposed by Rubin in 1993 sensitive values in the dataset are replaced by synthetic values generated from a model fitted to the original data. These synthetic datasets are then released to the public. If the models for generating the synthetic data are carefully selected, many of the relationships between the variables are preserved and valid inferences can be obtained for a broad class of analyses based on the released data.

In this workshop I will provide a brief introduction to this topic. I will introduce the two main approaches for generating synthetic data, discuss how valid inferences can be obtained from the synthetic data, and illustrate the basic methodological steps required for generating synthetic data.

Integration of multiple sources of data by the National Agricultural Statistics Service

Malay Ghosh

University of Florida, USA

ABSTRACT

The National Agricultural Statistics Service (NASS), one of the wings of USDA, is responsible for collecting data on planted acres, harvested acres, production and yield of various crops such as corn, soybean, wheat, cotton and others The goal is to produce estimates of all the above at the national, states, agricultural statistics districts and finally counties. In addition, NASS is mandated to provide cash rent estimates for states, agricultural statistics districts and counties. At the lowest level of geography, such as counties, the data is very sparse to provide reliable direct estimates. Accordingly, one needs to integrate several sources of data to provide estimates of greater precision. In this talk, I will briefly review some of the existing methods of NASS with particular emphasis on the analysis of cash rent data.

Big data in surveys: challenges, opportunities, and strategies

Jiming Jiang

University of California, Davis, USA

ABSTRACT

We discuss the notion of Big Data in the context of surveys. Examples of different types of survey Big Data, and the computational challenges, especially associated with complex statistical analysis, such as fitting a mixed effects model, are discussed. Some strategies of learning from the Big Data are proposed. In particular, we discuss a trade-off between statistical efficiency and computation complexity. Some empirical results, including simulation studies and analysis of Big GWAS (genome-wide association study) data are presented.

Bayesian model based IntegrOmics

Bani K. Mallick

Texas A&M University, USA

ABSTRACT

Accurate prognostic prediction using molecular information is a challenging area of research which is essential to develop precision medicine. We discuss translational models to identify major actionable proteins that are associated with clinical outcomes like the survival time of the patients. There are considerable statistical and computational challenges due to the large dimension of the problems. Furthermore, the data are available from different sources hence data integration is desirable. We consider integration of data from different sources like: (1) multiple cancers (pan cancer), (2) multiple responses (like survival time, type of tumors), or (3) multiple platforms (like protein expression and copy number). We develop sparse Bayesian models and computation techniques for each of this data integration problem. These proposed methods have been used to analyze data from the recently curated *The Cancer Proteome Atlas* (TCPA) which contains RPPA based high quality protein expression data, other genomics data as well as detailed clinical annotation including survival times.

Finding hidden code of omics data to disentangle the genetic architecture of disease

Indranil Mukhopadhyay

Duke University, USA

ABSTRACT

To understand the molecular architecture of disease, an unprecedented amount of next generation heterogeneous genomic data are generated. But to comprehend the knowledge of molecular mechanism of complex diseases, it is important to analyse the heterogeneous genomic data together, rather than individually. Aberrant proteins that most drugs target for therapeutic interventions quite often reveal dysfunction in the intermediate stages of protein formation such as gene expression or methylation assays, mutation or genotyping assays etc. Integration of information from such genomic data could provide crucial insight/knowledge in therapeutic target identification. But diversity in the characteristics of omics data and the variation in the sample sizes of omics data across all platforms are two of the major challenges in the data integration paradigm. We propose generic data integration methods to tackle these challenges.

Public health data integration

Takumi Saegusa

University of Maryland, USA

ABSTRACT

Various data sets collected from numerous sources have a great potential to enhance the quality of inference and accelerate scientific discovery. Inference for merged data is, however, quite challenging because such data may contain unidentified duplication from overlapping inhomogeneous sources and each data set often opportunistically collected induces complex dependence. In public health research, for example, epidemiological studies have different inclusion and exclusion criteria in contrast to hospital records without a well-defined target population, and when combined with a disease registry, patients appear in multiple data sets. In this talk, we present several examples in public health research which potentially enjoy merits of data integration. We overview existing research such as random effects model approach and multiple frame surveys and discuss their limitations in view of inferential goals, privacy protection, and large sample theory. We then propose our estimation method in the context of the Cox proportional hazards model. We illustrate our theory in simulation and read data example. If time permitted we discuss extensions of our proposed method in several directions.

Entity resolution with societal impacts in machine learning

Rebecca C. Steorts

Duke University, USA

ABSTRACT

Very often information about social entities is scattered across multiple databases. Combining that information into one database can result in enormous benefits for analysis, resulting in richer and more reliable conclusions. Among the types of questions that have been, and can be, addressed by combining information include: How accurate are census enumerations for minority groups? How many of the elderly are at high risk for sepsis in different parts of the country? How many people were victims of war crimes in recent conflicts in El Salvador? In most practical applications, however, analysts cannot simply link records across databases based on unique identifiers, such as social security numbers, either because they are not a part of some databases or are not available due to privacy concerns. In such cases, analysts need to use methods from statistical and computational science known as entity resolution (record linkage or de-duplication) to proceed with analysis. Entity resolution is not only a crucial task for social science and industrial applications, but is a challenging statistical and computational problem itself. In this talk, we describe the past and present challenges with entity resolution, with an application to the El Salvadorian conflict. More specifically, I will discuss unsupervised Bayesian entity resolution models, which are able to identify duplicate records in the data, while quantifying uncertainty. I will highlight the importance of choosing flexible priors and in implementing scalable inference algorithms. I will present preliminary results from the El Salvadorian conflict.

This is joint work with Bihan Zhuang (undergraduate student, Duke University), Neil Marchant (PhD candidate, University of Melbourne), and Ben Rubinstein (Associate Professor, University of Melbourne).

Variable selection in small area level models subject to sampling errors

Jiraphan Suntornchost

Chulalongkorn University, Thailand

ABSTRACT

Variable Selection plays crucial role in regression analysis. Several variable selection criteria have been proposed in literature. However, in the context of small area level models, dependent variables are usually collected subject to sampling errors. This leads to biases in the classical variable selection criteria. Therefore, Lahiri and Suntornchost (2015) proposed an adjustment to classical variable selection criteria. The study showed that their adjustment reduces biases of the variable selection criteria caused from sampling errors. The method has been extended to other regression settings such as linear panel data models and multivariate linear regression models.