Program

Fourteenth Annual Winter Workshop
Causal Inference and Graphical Models

Department of Statistics
University of Florida
January 13-14, 2012

Contents

Sponsors

The College of Liberal Arts & Sciences, University of Florida Graduate School, Info Tech and Department of Statistics

Organizing Committee

Michael Daniels
Hani Doss
Kshitij Khare

Invited Speakers

Mathias Drton, University of Chicago
Joe Hogan, Brown University
Kshitij Khare, University of Florida
Bala Rajaratnam, Stanford University
Thomas Richardson, University of Washington
Don Rubin, Harvard University
Daniel Scharfstein, Johns Hopkins University
Martin Wainwright, UC Berkeley
Nanny Wermuth, Chalmers Technical University and University of Gothenburg (Sweden)

Acknowledgements

The organizers thank the Department of Statistics staff, Robyn Crawford, Tina Greenly, Summer Layton and Marilyn Saddler for their tremendous efforts in helping to set up this meeting and making it run smoothly.

Conference Schedule

All workshop sessions meet in the 209 Teaching Classroom, Emerson Alumni Hall

Thursday, January 12,2012

7:00-9:00 pm Reception Emerson Hall, Al and Judy Warrington Room

Friday, January 13,2012

8:00-8:45 am **Breakfast:** Continental

8:45-9:00 am Conference Welcome: Hani Doss, Organizing Committee
Alan Dorsey, Associate Dean of CLAS

9:00-10:40 am **Sessioni:** *Chair·. Hani Doss, University of Florida*

Sequences of Regressions and Their Independence Graphs
Nanny Wermuth, Chalmers Technical University & University of Gothenburg (Sweden)

Identifiability of Linear Structural Equation Models
Mathias Drton, University of Chicago

10:40-11:10 am **Break:** Refreshments

ll:10-Noon **Session 2:** **Chair:** *Xu Han, University of Florida*

Graphical Model Selection in High Dimensions: Practical Methods And Fundamental Limits
Martin Wainwright, University of California, Berkeley

Noon-12:30 pm **Conference photo on the Emerson Alumni Hall stairs to the second floor**

12:30-2:00 pm **Lunch:** (Gator Corner Dining Center)

2:00-3:40 pm **Session 3:** **Chair:** *Michael Daniels, University of Florida*

The Use of Randomized Incentives in Studies with Missing Outcome Data
Daniel Scharfstein, Johns Hopkins University

Nested Markov Properties for Acyclic Directed Mixed Graphs
Thomas Richardson, University of Washington

3:40-5:00 pm **Poster Session:** Emerson Alumni Hall, Presidents Room B (Refreshments)

All workshop sessions meet in the 209 Teaching Classroom, Emerson Alumni Hall

Saturday, January 14,2012

9:00-10:00 am Breakfast: Continental

10:00-11:40 am Session 4: *Chair·. Malay Ghosh,University of Florida*

For Objective Causal Inference, Design Trumps Analysis
Donald Rubin, Harvard University

Causal Inference About Mediation When There Are Several Mediators
Joseph Hogan, Brown University

11:40-1:30 pm Lunch: Free time

1:30-3:10 pm Session 5: Chair: *Nikolay Bliznyuk, University of Florida*

Generalized Hyper Markov Laws for Directed Markov Random Fields
Bala Rajaratnam, Stanford University

Cholesky Based Estimation In Graphical Models
Kshitij Khare, University of Florida

Invited Talks

Identifíability of Linear Structural Equation Models

Mathias Drton
University of Chicago

Structural equation models are multivariate statistical models that are defined by specifying noisy functional relationships among random variables. This talk treats the classical case of linear relationships and additive Gaussian noise terms. Each linear structural equation model can be represented by a mixed graph in which directed edges encode the linear equations, and bi-directed edges indicate possible correlations among noise terms. A basic problem in structural equation modeling is to determine which models are identifiable. In other words, each model corresponds to a parametrized set of positive definite covariance matrices and we wish to determine the graphs for which the edge coefficients and correlations appearing in the parametrization can be uniquely recovered from the covariance matrix. I will discuss recent results on this problem based on joint work with Jan Draisma, Rina Foygel and Seth Sullivant.

Causal Inference about Mediation when there are Several Mediators

Joseph Hogan
Brown University

In the context of randomized intervention trials of behavior science, the typical objective is to understand how the effect of an intervention operates on a primary outcome through potential mediating variables. Often there is more than one mediation path, and the relations between potential mediating variables suggest that the multiple mediator model is more of interest than the single mediator model. We develop a model to infer separately the mediation effects for individual variables when there are several potential mediators. A causal model, parameterized in terms of natural direct and indirect effects (Pearl, 2001), is used to encode mediation. To identify the natural indirect effects, we require information about the joint potential outcomes distribution of each mediator. Our model identifies this joint distribution using information from baseline covariates and by placing targeted restrictions the correlation structure of the potential mediators. Unobserved potential mediators and associated potential outcomes can therefore be imputed under the model, and causal contrasts of interest can be computed in a straightforward manner. We illustrate our methods in both simulation studies and an analysis of a recent intervention trial designed to increase physical activity.

Co-authored with Jing Zhang

Cholesky Based Estimation In Graphical Models

Kshitij Khare
University of Florida

We consider the problem of sparse covariance estimation in high dimensional settings using graphical models. These models can be represented in terms of a graph, where the nodes represent random variables and edges represent their interactions. When the random variables are jointly Gaussian distributed, the lack of edges in such graphs can be interpreted as conditional and/or marginal independencies between these variables. We present a computationally efficient approach for high dimensional sparse covariance estimation in graphical models based on the Cholesky decomposition of the covariance matrix or its inverse. The proposed method is illustrated on both simulated and real data.

Generalized Hyper Markov Laws For Directed Markov Random Fields

Bala Rajaratnam
Stanford University

In the recent past, classes of flexible hyper Markov laws have been proposed (and subsequently analyzed) for the purposes of high dimensional Bayesian inference in important classes of Gaussian graphical models. A shortcoming of many of these theoretical endeavors is that they are often restricted to special class of graphs (such as decomposable graphs etc...), and are therefore not always readily applicable - though they have led to rich probability and statistical theory. In this talk we propose a novel approach that aims to move beyond this restriction for the class of directed Markov Random fields.

This is joint work with Emanuel Ben-David.

Nested Markov Properties for Acyclic Directed Mixed Graphs

Thomas Richardson
University of Washington

An acyclic directed mixed graph (ADMG) contains directed and bi-directed edges subject to the restriction that there are no directed cycles. ADMGs have been used to represent the conditional independence relations implied by a DAG with hidden variables on the distribution of the observed variables. However, it has long been understood that there are additional non-parametric constraints that arise directly from the factorization given by a partially observed DAG that do not correspond to conditional independence restrictions, such as the ”Verma constraint”. In this talk I will show that such constraints may be viewed as conditional independence restrictions in re-weighted distributions. I will introduce a 'nested' Markov property for ADMGs which implies these constraints. These re-weighted distributions have a natural causal interpretation in the context of the underlying DAG model.

This is a joint work Rubin Evans, James Robins and Ilya Shpitser.

For Objective Causal Inference, Design Trumps Analysis

Donald B. Rubin
Harvard University

For obtaining causal inferences that are objective, and therefore have the best chance of revealing scientific truths, carefully designed and executed randomized experiments are generally considered to be the gold standard. Observational studies, in contrast, are generally fraught with problems that compromise any claim for objectivity of the resulting causal inferences. The thesis here is that observational studies have to be carefully designed to approximate randomized experiments, in particular, without examining any final outcome data. Often a candidate data set will have to be rejected as inadequate because of lack of data on key covariates, or because of lack of overlap in the distributions of key covariates between treatment and control groups, often revealed by careful propensity score analyses. Sometimes the template for the approximating randomized experiment will have to be altered, and the use of principal stratification can be helpful in doing this. These issues are discussed and illustrated using the framework of potential outcomes to define causal effects, which greatly clarifies critical issues.

The Use of Randomized Incentives in Studies with Missing Outcome Data

Daniel Scharfstein
Johns Hopkins University

Non-response is a common problem in experimental and observational studies. Point identification of population-level parameters relies on untestable assumptions about the relationship between non-response and outcomes. Rather than point identification, Manski (2001) recommends reporting identification regions, where regions narrow as more assumptions are imposed. In this talk, we focus on the development of confidence regions for two types of identification regions. The first depends on no assumptions and the second depends on the (testable) existence of a randomized incentive designed to influence the response rate. We develop our methods in the context of the Three-City Study, which surveyed low to middle income families in three major U.S. cities.

This is a joint work with Hui-Chun Hsu, Thomas Richardson and Robert Moffltt.

Graphical Model Selection In High Dimensions: Practical Methods And Fundamental Limits

Martin Wainwright
University of California, Berkeley

Given samples from a Markov random field, how to determine the unknown graph? This graph selection problem is important for many applications of graphical models, and has received a great deal of attention over the past years. In this talk, we present a simple polynomial-complexity method, based on pseudolikelihood and ℓ_l-regularization, for estimating graph structure. We show that it can recover the correct graph structure high probability as long as $n = \Omega(d^2 \log p)$, where n is the sample size, p is the number of vertices, and d is the maximum degree. Using information-theoretic methods, we show that no method can recover the graph correctly when the sample size is much smaller than this critical amount. We then discuss extensions of these methods to problems with noisy, missing and/or dependent data.

Based on joint works with John Lafferty, Po-Ling Loh, and Pradeep Ravikumar.

Sequences of Regressions and Their Independence Graphs

Nanny Wermuth
Chalmers Technical University and
University of Gothenburg

The study of pathways of dependences has a long history in the special case of sequences of linear regressions with single and joint responses. However, our knowledge on sequences of regressions in which joint responses may be quantitative or categorical variables or of both types, have only in the last decade developed into a general framework for studying development over time, in both observational and interventional studies. In this lecture, I summarize some of the recent results and of applications.

Poster Abstracts

Positive Definite Completion Problems For Bayesian Networks

Emanuel Ben-David
Stanford University

A positive definite completion problem pertains to determining whether the unspecified positions of a partial (or incomplete) matrix can be completed in a desired subclass of positive definite matrices. In this poster we present an important and new class of positive definite completion problems where the desired subclasses are the spaces of covariance and inverse-covariance matrices of probabilistic models corresponding to Bayesian networks (also known as directed acyclic graph models). We provide fast procedures that determine whether a partial matrix can be completed in either of these spaces and thereafter proceed to construct the completion matrices. We prove an analog of the positive definite completion in the context of directed acyclic graphs. We also proceed to give closed form expressions for the inverse and the determinant of a completion matrix as a function of only the elements of the corresponding partial matrix.

Spatial Graphical Models with Discrete and Continuous Components

Xuan Che
Oregon State University

Graphical models use the Markov property to establish associations among dependent variables. To estimate spatial correlation and other parameters in a graphical model, the conditional independences and the joint density of the graph need to be specified. We can rely on the Gaussian multivariate model to derive the joint density when all the nodes of the graph are assumed to be marginally normally distributed. However, when some of the nodes are discrete random variables, the Gaussian model no longer affords an appropriate joint density for the graph. We develop a method for specifying the joint distribution of a chain graph with both discrete and continuous nodes and spatial correlation among the discrete nodes. We structure the graph as a generalized tree network and then use a multivariate Gaussian copula to model the dependence structures of the graph while accommodating the discrete marginal distributions of the appropriate nodes. Generalized tree networks partition a graph's joint density according to its subgraphs, while copula models help separate variables' correlations from their marginal distributions. We analyze a dataset with disease incidence and demographic information and compare our results to those from a spatial random effects model for spatially correlated Bernoulli responses with fixed explanatory information.

Using Propensity Score Matching On Click-Stream Data

Douglas Galagate
University of Maryland

My research focuses on applying different statistical methods to click-stream data. This summer, I had the chance to present at Joint Statistical Meetings in Miami, Florida and also at the useR! conference at the University of Warwick in the UK. The work was on using click-stream data to understand how potential customers behave online. Most of the work that I did was descriptive and mostly exploratory data analysis. Recently, I have been pinning down more specific measures of the effects that advertising has on consumer behavior and have applied causal inference techniques.

It is usually easier to understand a process by doing experiments, but a lot of times, all we have is observational data. Learning as much as we can from observational data is one of the reasons for using matching techniques such as propensity score matching to create similar treated and non-treated groups. From there, we try to measure the effect of how a treatment such as a different advertisement can lead to results. We look at causal inference and its application in evaluating advertising effects.

An Outcome-Free Procedure for Interval Estimation of Causal Effects

Roee Gutman
Brown University

The estimation of population/subpopulation average treatment effects has been a subject of extensive research. In a randomized experiment, common practice entails estimating the average treatment effect by calculating the difference between the average outcome in the treatment group and the average outcome in the control group. In cases where additional covariates that affect the outcome exist, estimation of the treatment effect is typically controlled (adjusted) by using a model that combines the treatment effect and a function of the covariates in an additive manner. This type of model relies on the assumption that the response surfaces from the outcome given the covariates are parallel in the control and treatment groups. When this assumption is incorrect, the estimation of the treatment effect may be unreliable. In observational studies, the effect of this assumption is even more substantial, because the distributions of the covariates in the two groups are usually different.

This talk will focus on the proposal of an outcome-free three-stage procedure based on Rubin's framework for causal inference. First, we create subclasses that include observations from each group based on the covariates. Next, we independently estimate the response surface in each group using flexible spline model. Lastly, multiple imputations of the missing potential outcomes are performed. A simulation analysis which resembles real life situations and compares this procedure to other common methods is carried out. In relation to other methods and in many of the experimental conditions examined, our proposed method is the only one that produced a valid statistical procedure while providing a relatively precise point estimate and a relatively short interval estimate.

Bayesian Inference For The Causal Effect Of Mediation With Baseline Covariates

Chanmin Kim
University of Florida

Mechanistic models of behavior change inherently describe causal processes; however, most analyses in the behavioral literature rely on the Baron and Kenny (1986) regressi on-based approach, which generally cannot be used to infer causal effects. Their research is based on randomization of a post-randomization variable which cannot be controlled in practice. Researchers may be reluctant to use causal models due to lack of full understanding of the models. This paper is intended to bridge the gap between theory and practice of causal modeling for assessment of mediation in behavioral intervention studies through Bayesian methodologies.
We propose nonparametric Bayesian approaches to estimate the natural direct and indirect effects through a mediator in the setting of a continuous mediator and a binary or continuous response. In addition to this, we incorporate baseline covariates to increase the efficiency in estimating NDE and NIE. Several conditional independence assumptions are introduced (with corresponding sensitivity parameters) to make these effects identifiable from the observed data. This approach is used to assess mediation in a recent weight management clinical trial.

This is a joint work with Michael Daniels.

Assessing the Surrogate Value of a Biomarker: An Approach Combining Matching and Sensitivity Analysis

Julian Wolf son
University of Minnesota

Statisticians have developed a number of frameworks which can be used to assess the surrogate value of a biomarker, i.e. establish whether it can be used in place of a clinical endpoint for quantifying the effect of a treatment. The most commonly applied of these frameworks is due to Prentice (1989), who proposed a set of criteria which a surrogate marker should satisfy. However, verifying these criteria using observed data can be challenging due to the presence of unmeasured simultaneous predictors (i.e. confounders) which influence both the potential surrogate and the outcome. In this work, we adapt a technique proposed by Rosenbaum (2002) for observational studies, in which observations are matched and the odds of treatment within each matched pair is bounded. This yields a straightforward and interpretable sensitivity analysis which can be performed particularly efficiently for certain types of test statistics. We discuss the details of this technique, and illustrate with an example drawn from the infectious disease literature.