Quantile Regression and Data Heterogeneity Workshop

Feb 12-13, 2022

UNIVERSITY OF MIAMI
MIAMI HERBERT
BUSINESS SCHOOL

Sponsored by the National Science Foundation; and co-sponsored by the Department of Management Science, University of Miami; Department of Statistics, University of Michigan; and the Deloitte Institute for Research andPractice in Analytics (DIRPA), University of Miami

Workshop Venue

Storer Auditorium

Miami Herbert Business School

University of Miami

5250 University Drive, Coral Gables, FL 33146

Organizing Committee

Xuming He (University of Michigan), Kengo Kato (Cornell University), Snigdha Panigrahi (University of Michigan), Lan Wang (University of Miami), Qi Zheng (University of Louisville)

Local Committee

Lan Wang (University of Miami), Doug Lehmann (University of Miami), Ganggang Xu (University of Miami), Emma Jingfei Zhang (University of Miami)

Saturday, February 12

8:15 am	Food & Drinks
8:50 am	Welcome

Session 1

9:00-10:20 am

Regina Liu (Rutgers University): Fusion learning: combine inferences from diverse data sources with heterogeneous data

Ivan Fernandz-Val (Boston University): Dynamic heterogeneous distribution regression panel models with an application to labor income processes

Chair/Discussant: **Roger Koenker**

10:20 am Coffee break

Session 2

10:40 am - 12:00 pm

Kean Ming Tan (University of Michigan): Convolution-type smoothing approach for quantile regression

Jingshen Wang (University of California, Berkeley): Debiased inference on heterogeneous quantile treatment effects with regression rank-scores

Chair/Discussant: **Lan Wang, Ganggang Xu**

Noon-1:30 pm lunch break

Session 3

1:30 - 2:50 pm

Jelena Bradic (University of California, San Diego): Causal Learning: excursions in double robustness

Ying Wei (Columbia University): Integrated quantile rank test (IQRAT) for gene-level associations

Chair/Discussants: **Emma Zhang, Hongyu Zhao**

2:50-3:10 pm   Coffee break

Session 4: Invited Presentations from Junior Scholars

3:10-4:30 pm

Shuangning Li (Stanford University): Random graph asymptotics for treatment effect estimation under network interference

Ben Sherwood (University of Kansas): Computationally efficient penalized quantile regression

Bo Wei (University of Michigan): Estimation of complier super-quantile causal effects with a binary instrumental variable

Harold Chiang (University of Wisconsin-Madison): Inference for high-dimensional exchangeable arrays

Chair/discussant: **Snigdha Panigrahi**

4:30-4:40 pm   break

4:40-5:10 pm FRG group report and discussion

6:00 - 8:00 pm Workshop dinner (Fontana at the Biltmore Hotel Restaurant)

Sunday, February 13

9:00 am Food & Drinks

Session 5

9:30-10:50 am

Matias Cattaneo (Princeton University): On Binscatter

Bodhi Sen (Columbia University): Measuring association on topological spaces using kernels and geometric graphs

Chair/Discussant: **Kengo Kato**

10:50-11:10 am coffee break

Session 6

11:10am -12:30pm

Sunil Rao (University of Miami): A Tour of Classified Mixed Model Predictions and Projections

Xiaohong Chen (Yale University): Adaptive Estimation and Uniform Confidence Bands for Nonparametric IV

Chair/Discussants: **Qi Zheng, Xuming He**

Afternoon free discussion and social time

Talk Titles and Abstracts

Session 1

1. **Regina Liu** (Rutgers University, rliu@stat.rutgers.edu)

Title: Fusion learning: combine inferences from diverse data sources with heterogeneous data

Abstract: Advanced data collection technology nowadays often makes inferences from diverse data sources easily accessible. Fusion learning refers to combining inferences from multiple sources or studies to make more effective inference than from any individual source or study alone. We focus on the tasks: 1) Whether/When to combine inferences? 2) How to combine inferences efficiently if you need to? We present a general framework for nonparametric and efficient fusion learning for inference on multi-parameters, which may be correlated. The main tool underlying this framework is the new notion of depth confidence distribution (depth-CD), which is developed by combining data depth, bootstrap and confidence distributions. We show that a depth-CD is an omnibus form of confidence regions, whose contours of level sets shrink toward the true parameter value, and thus an all-encompassing inferential tool. The approach is shown to be efficient, general and robust. Specifically, it achieves high-order accuracy and Bahadur efficiency under suitably chosen combining elements. It readily applies to heterogeneous studies with a broad range of complex and irregular settings. This property also enables the approach to utilize indirect evidence from incomplete studies to gain efficiency for the overall inference. This is joint work with Dungan Liu of University of Cincinnati and Minge Xie of Rutgers University.

2. **Ivan Fernandz-Val** (Boston University, ivanf@bu.edu)

Title: "Dynamic Heterogenous Distribution Regression Panel Models with an Application to Labor Income Processes," joint with Wayne Gao, Yuan Liao and Francis Vella

Abstract: We consider estimation of a dynamic distribution regression panel data model with heterogeneous coefficients across units. The objects of interest are functionals of these coefficients including linear projections on unit level covariates. We also consider actual, stationary and counterfactual distributions of the outcome variable. We investigate how

changes in initial conditions or covariate values affect these objects. Coefficients and their functionals are estimated via fixed effect methods, which are debiased to deal with the incidental parameter problem. We propose a cross-sectional bootstrap scheme to perform uniform inference on function-valued objects. This avoids coefficient re-estimation and is shown to be consistent for a large class of data generating processes, including the reference point of coefficient homogeneity (conditional on covariates). We employ annual labor income data from the PSID to illustrate the variety of empirical issues we can address. First, we predict the impact of hypothetical tax policies and find substantially smaller effects than those from models based on homogeneous autoregressive and distributional regression processes. Second, we examine the impact on the distribution of labor income from increasing the education level of a chosen subsample of workers. Explicitly, we increase the education level of all workers with less than 12 years of schooling to that level of schooling and find short and long run increases in the distribution at the bottom tail with the upper tail relatively unaffected. Finally we uncover notable heterogeneity in income mobility implying substantial individual heterogeneity in the incidence to be trapped in poverty.

Session 2

1. Kean Ming Tan (University of Miami, keanming@umich.edu)

Title: Convolution-Type Smoothing Approach for Quantile Regression

Abstract: Quantile regression is a powerful tool for learning the relationship between a response variable and a multivariate predictor while exploring heterogeneous effects. However, the non-smooth piecewise linear loss function introduces challenges to the computational aspect when the number of covariates is large. To address the aforementioned challenge, we propose a convolution-type smoothing approach that turns the non-differentiable quantile piecewise linear loss function into a twice differentiable, globally convex, and locally strongly convex surrogate, which admits a fast and scalable gradient-based algorithm to perform optimization. In the low-dimensional setting, we establish nonasymptotic error bounds for the resulting smoothed estimator. In the high-dimensional setting, we propose the concave

regularized smoothed quantile regression estimator, which we solve using a multi-stage convex relaxation algorithm. Theoretically, we characterize both the algorithmic error due to non-convexity and statistical error for the resulting estimator simultaneously. We show that running the multi-stage algorithm for a few iterations will yield an estimator that achieves the oracle property. Our results suggest that the smoothing approach leads to a significant computational gain without a loss in statistical accuracy.

2. **Jingshen Wang** (University of California at Berkeley, jingshenwang@berkeley.edu)

Title: Debiased Inference on Heterogeneous Quantile Treatment Effects with Regression Rank-Scores

Abstract: Understanding treatment effect heterogeneity in observational studies is of great practical importance to many scientific fields because the same treatment may affect different individuals differently. Quantile regression provides a natural framework for modelling such heterogeneity. In this paper, we propose a new method for inference on heterogeneous quantile treatment effects that incorporates high-dimensional covariates. Our estimator combines a L1-penalized regression adjustment with a quantile-specific bias correction scheme based on quantile regression rank scores. We present a comprehensive study of the theoretical properties of this estimator, including weak convergence of the heterogeneous quantile treatment effect process to the sum of two independent, centered Gaussian processes. We illustrate the finite-sample performance of our approach through Monte Carlo experiments and an empirical example, dealing with the differential effect of statin usage for lowering LDL cholesterol levels for the Alzheimer's disease patients who participated in the UK Biobank study. This is joint work with Alexander Giessing at the University of Washington.

Session 3

1. **Jelena Bradic** (University of California, San Diego, jbradic@ucsd.edu)

Title: Causal Learning: excursions in double robustness

Abstract: Recent progress in machine learning provides many potentially effective tools to learn estimates or make predictions from datasets of ever-increasing sizes. Can we trust such tools in clinical and highly-sensitive systems? If a learning algorithm predicts an effect of a new policy to be positive, what guarantees do we have concerning the accuracy of this prediction? The talk introduces new statistical ideas to ensure that the learned estimates satisfy some fundamental properties: especially causality and robustness. The talk will discuss potential connections and departures between causality and robustness.

2**. Ying Wei** (Columbia University, yw2148@cumc.columbia.edu)

Title: Integrated Quantile Rank Test (IQRAT) for Gene-level associations

Abstract: Gene-based testing is a commonly employed strategy in genetic association studies. Gene-trait associations are complex due to underlying population heterogeneity, gene-environment interactions, and various other reasons. Existing gene-based tests, such as Burden and Sequence Kernel Association Tests (SKAT), focus on mean-level associations and may miss or underestimate higher-order associations that could be scientifically interesting. We introduce a new family of gene-level association tests that integrate the quantile-rank score process to accommodate complex associations better. The resulting test statistics enjoy multiple advantages. First, they are almost as efficient as the best existing tests when the associations are homogeneous across quantile levels and have improved efficiency for complex and heterogeneous associations. Second, they provide valuable insights into risk stratification. Third, the test statistics are distribution-free and could accommodate a wide range of underlying distributions; We established the asymptotic properties under the null and alternative hypotheses to validate the proposed tests theoretically. We also conducted extensive simulations to assess its empirical performance compared to the existing approaches. Finally, we illustrate its real-world applications to identify genes associated with lipid traits using a Metabochip dataset and identifying eGenes from the multi-tissue gene-expression data in GTEx.

Session 4

1. Shuangning Li (Stanford University, lsn@stanford.edu)

Title: Random Graph Asymptotics for Treatment Effect Estimation under Network Interference

Abstract: The network interference model for causal inference places all experimental units at the vertices of an undirected exposure graph, such that treatment assigned to one unit may affect the outcome of another unit if and only if these two units are connected by an edge. This model has recently gained popularity as means of incorporating interference effects into the Neyman--Rubin potential outcomes framework; and several authors have considered estimation of various causal targets, including the direct and indirect effects of treatment. In this paper, we consider large-sample asymptotics for treatment effect estimation under network interference in a setting where the exposure graph is a random draw from a graphon. When targeting the direct effect, we show that---in our setting---popular estimators are considerably more accurate than existing results suggest, and provide a central limit theorem in terms of moments of the graphon. Meanwhile, when targeting the indirect effect, we leverage our generative assumptions to propose a consistent estimator in a setting where no other consistent estimators are currently available. We also show how our results can be used to conduct a practical assessment of the sensitivity of randomized study inference to potential interference effects. Overall, our results highlight the promise of random graph asymptotics in understanding the practicality and limits of causal inference under network interference. This is joint work with Stefan Wager.

2. Ben Sherwood (University of Kansas, ben.sherwood@ku.edu)

Title: Computationally efficient penalized quantile regression

Abstract: Quantile regression with a lasso penalty can be framed as a linear programming problem. If a group lasso penalty is used, then it becomes a second order cone programming problem. These approaches become computationally burdensome for large values of n or p.

Using a Huber approximation to the quantile function allows for the use of computationally efficient algorithms that require a differentiable loss function that can be implemented for both penalties. These algorithms then can be used as the backbones for implanting penalized quantile regression with other penalties such as Adaptive Lasso, SCAD, MCP and group versions of these penalties.

3. Bo Wei (University of Michigan, boweinju@umich.edu)

Title: Estimation of Complier Super-quantile Causal Effects with a Binary Instrumental Variable

Abstract: Estimating causal effect of a treatment or exposure for an interested subpopulation is a fundamental interest in many biomedical and economical studies. Super-quantile, also known as expected shortfall, is an attractive measure of risk for the subpopulation because it can capture heterogeneity and aggregated local information of effect over a range of distribution of outcomes simultaneously. In this work, we propose a complier super-quantile causal effect (CSQCE) model under instrumental variable (IV) framework to quantity the CSQCE for the data with unmeasured confounders. By utilizing the special characteristic of binary IV, we propose a simple and easily-implemented two-step estimation procedure, which can simply be solved by weighted linear regression and weighted quantile regression. We rigorously justify the asymptotic properties for the proposed estimator. Extensive simulations have been conducted to confirm its validity and satisfactory finite-sample performance. An application to a dataset from National Job Training Partnership Act (JTPA) study demonstrates the practical utility of the proposed method.

4. Harold Chiang (University of Wisconsin-Madison, hdchiang@wisc.edu)

Title: Inference for high-dimensional exchangeable arrays

Abstract: We consider inference for high-dimensional separately and jointly exchangeable arrays where the dimensions may be much larger than the sample sizes. For both exchangeable arrays, we first derive high-dimensional central limit theorems over the rectangles and subsequently develop novel multiplier bootstraps with theoretical guarantees. These theoretical results rely on new technical tools such as Hoeffding-type decomposition and maximal inequalities for the degenerate components in the Hoeffiding-type decomposition for the exchangeable arrays. We exhibit applications of our methods to uniform confidence bands for density estimation under joint exchangeability and penalty choice for l1-penalized regression under separate exchangeability. Extensive simulations demonstrate precise uniform coverage rates. We illustrate by constructing uniform confidence bands for international trade network densities.

Session 5

1. **Matias D. Cattaneo** (Princeton University, cattaneo@princeton.edu)

Title: On Binscatter

Abstract: Binscatter, or a binned scatter plot, is a very popular tool in applied microeconomics. It provides a flexible, yet parsimonious way of visualizing and summarizing mean, quantile, and other nonparametric regression functions in large data sets. It is also often used for informal evaluation of substantive hypotheses such as linearity or monotonicity of the unknown function. This paper presents a foundational econometric analysis of binscatter, offering an array of theoretical and practical results that aid both understanding current practices (i.e., their validity or lack thereof) as well as guiding future applications. In particular, we highlight important methodological problems related to covariate adjustment methods used in current practice, and provide a simple, valid approach. Our results include a principled choice for the number of bins, confidence intervals and bands, hypothesis tests for parametric and shape restrictions for mean, quantile, and other functions of interest, among other new methods, all applicable to canonical binscatter as well as to nonlinear, higher-order polynomial, smoothness-restricted and covariate adjusted extensions thereof. Companion general-purpose software packages for Python, R, and Stata are provided. From a technical perspective, we present novel theoretical results for possibly nonlinear semi-parametric partitioning-based series estimation with random partitions that are of independent interest.

2. **Bodhi Sen** (Columbia University, bodhi@stat.columbia.edu)

Title: Measuring association on topological spaces using kernels and geometric graphs

Abstract: We propose and study a class of simple, nonparametric, yet interpretable measures of association between two random variables X and Y taking values in general topological spaces. These nonparametric measures -- defined using the theory of reproducing kernel Hilbert spaces -- capture the strength of dependence between X and Y and have the property that they are 0 if and only if the variables are independent and 1 if and only if one variable is a measurable function of the other. Further, these population measures can be consistently estimated using the general framework of graph functionals which include k-nearest neighbor graphs and minimum spanning trees. Moreover, a sub-class of these estimators are also shown to adapt to the intrinsic dimensionality of the underlying distribution. Some of these empirical measures can also be computed in near-linear time. Under the hypothesis of independence between X and Y, these empirical measures (properly normalized) have a standard normal limiting distribution. Thus, these measures can also be readily used to test the hypothesis of mutual independence between X and Y. In fact, as far as we are aware, these are the only procedures that possess all the above mentioned desirable properties.

Session 6

1. J. Sunil Rao (University of Miami, JRao@med.miami.edu)

Title: A Tour of Classified Mixed Model Predictions and Projections

Abstract: Many practical problems are related to prediction where the main interest is at the subject or small sub-population level. In such cases, it's possible to make substantial gains in prediction accuracy by identifying a class that a new subject belongs to and associating the new subject with a random effect corresponding to the same class in the training data so that mixed model prediction can be used. In this talk, we first introduce the original classified mixed model prediction idea and then discuss some newer developments in multivariate classified mixed model prediction and classified mixed model projections for new data outside the range of the training data. This is joint work over the years with Jiming Jiang (UC-Davis), Thuan Nguyen (OHSU) and former UM students Menying Li (Moderna), Hang Zhang (Biogen) and Jie Fan (Novartis).

2. Xiaohong Chen (Yale University, xiaohong.chen@yale.edu)

Title: Adaptive Estimation and Uniform Confidence Bands for Nonparametric IV

Abstract: We introduce computationally simple, data-driven procedures for estimation and inference on a structural function h_0 and its derivatives in nonparametric models using instrumental variables. Our first procedure is a bootstrap-based, data-driven choice of sieve dimension for sieve nonparametric instrumental variables (NPIV) estimators. When implemented with this data-driven choice, sieve NPIV estimators of h_0 and its derivatives are adaptive: they converge at the best possible (i.e., minimax) sup-norm rate, without having to know the smoothness of h_0, degree of endogeneity of the regressors, or instrument strength. Our second procedure is a data-driven approach for constructing honest and adaptive uniform confidence bands (UCBs) for h_0 and its derivatives. Our data-driven UCBs guarantee coverage for h_0 and its derivatives uniformly over a generic class of data-generating processes (honesty) and contract at, or within a logarithmic factor of, the minimax sup-norm rate (adaptivity). As such, our data-driven UCBs deliver asymptotic efficiency gains relative to UCBs constructed via the usual approach of undersmoothing. In addition, both our procedures apply to nonparametric regression as a special case. We use our procedures to estimate and perform inference on a nonparametric gravity equation for the intensive margin of firm exports and find evidence against common parameterizations of the distribution of unobserved firm productivity. (Authors: Xiaohong Chen, Tim Christensen and Sid Kankanala.)

Workshop Participants

(in alphabetical order)

Hyoin An (Ohio State University, an.355@osu.edu)

Jelena Bradic (University of California at San Diego, jbradic@ucsd.edu)

Matias Cattaneo (Princeton University, cattaneo@princeton.edu)

Xiaohong Chen (Yale University, xiaohong.chen@yale.edu)

Harold Chiang (University of Wisconsin-Madison, hdchiang@wisc.edu)

Ying Cui (Emory University, ying.cui@emory.edu)

Ivan Fernandz-Val (Boston University, ivanf@bu.edu)

Xuming He (University of Michigan, xmhe@umich.edu)

Feifang Hu (George Washington University, feifang@gwu.edu)

Yingsi Huang (University of Miami, yxh953@miami.edu)

Kengo Kato (Cornell University, kk976@cornell.edu)

Roger Koenker (University College London, UK, rkoenker@illinois.edu)

Ji Hyung Lee (University of Illinois, jihyung@illinois.edu)

Shuangning Li (Stanford University, lsn@stanford.edu)

Yuanzhi Li (University of Michigan, yzli@umich.edu)

Regina Liu (Rutgers University, rliu@stat.rutgers.edu)

Yanyuan Ma (Pennsylvania State University, yanyuanma@gmail.com)

Aaron J. Molstad (University of Florida, amolstad@ufl.edu)

Snigdha Panigrahi (University of Michigan, psnigdha@umich.edu)

Lekha Patel (Sandia National Laboratories, lpatel@sandia.gov)

Sunil Rao (University of Miami, JRao@med.miami.edu)

Bodhi Sen (Columbia University, bodhi@stat.columbia.edu)

Ritwik Sadhu (Cornell University, rs2526@cornell.edu)

Ben Sherwood (University of Kansas, ben.sherwood@ku.edu)

Kean Ming Tan (University of Michigan, keanming@umich.edu)

Sudaraka Tholkage (University of Louisville, donramesh.tholkage@louisville.edu)

Ganggang Xu (University of Miami, gangxu@bus.miami.edu)

Qian Xu (University of Louisville, qian.xu@louisville.edu)

Dewei Wang (University of South Carolina, deweiwang@stat.sc.edu)

Jingshen Wang (University of California at Berkeley, jingshenwang@berkeley.edu)

Lan Wang (University of Miami, lanwang@mbs.miami.edu)

Shuoyang Wang (Auburn University, szw0100@auburn.edu)

Bo Wei (University of Michigan, boweinju@umich.edu)

Ying Wei (Columbia University, yw2148@cumc.columbia.edu)

Emma Zhang (University of Miami, ezhang@bus.miami.edu)

Shushu Zhang (University of Michigan, shushuz@umich.edu)

Hongyu Zhao (Yale University, hongyu.zhao@yale.edu)

Tuoyi Zhao (University of Miami, txz311@miami.edu)

Qi Zheng (University of Louisville, qi.zheng@louisville.edu)