Dingyi Li

Email: dl922@cornell.edu Website: dingyili93.github.io Last Update: October, 2025

EDUCATION

- **Cornell University**
  Ph.D. in Applied Economics and Management 2019-2026 (Expected)
  M.S. in Applied Economics and Management 2017-2019
- **Renmin University of China**
  B.A. in Economics 2012-2016
  B.S. in Mathematics 2012-2016

RESEARCH INTERESTS

Econometrics, Applied Econometrics, Machine Learning, Empirical Industrial Organization

WORKING PAPERS

- "Doubly Robust Causal Inference with Confounders Missing Not at Random." **Job Market Paper**

  Abstract: Drawing causal inference from observational studies is challenging if confounders are missing not at random (MNAR). Existing nonparametric two-stage least squares estimation methods ensure consistency but often lack efficiency or require full parametric assumptions for the inference in high-dimensional settings. We derive the semiparametric efficiency bound for estimating the average treatment effect under the MNAR assumption with missingness conditionally independent of the outcome given the treatment and confounders. We propose a doubly robust estimator by solving two sets of Fredholm integral equations, which attains this bound. One set follows existing approaches that integrate over confounders, while the other, which integrates over outcomes with some unknown low-dimensional structure, is novel. Our estimator is doubly robust in two respects. First, the estimator remains consistent when either the confounder or the outcome has greater dimensionality, provided that other necessary assumptions are also satisfied. Second, in line with the double machine learning literature, the fourth-root rate convergence of nuisance parameters is critical to ensure $\sqrt{n}$-consistency, asymptotic linearity, and local efficiency of our estimator. We assess the estimator through simulations and three empirical applications: the impact of the Job Corps program on employment, the effect of smoking on blood lead levels, and the influence of education on general health satisfaction.

- "On the Testability of Anchor Words in Topic Models," with Simon Freyaldenhoven, Shikun Ke, and José Luis Montiel Olea. **Revise** & **Resubmit at Quantitative Economics**.

  Abstract: Topic models are a simple and popular tool for the statistical analysis of textual data. Their identification and estimation is typically enabled by assuming the existence of *anchor words*; that is, words that are exclusive to specific topics. In this paper we show that the existence of anchor words is statistically testable: There exists a hypothesis test with correct size that has nontrivial power. This means that the anchor-words assumption cannot be viewed simply as a convenient normalization. Central to our results is a simple characterization of when a column-stochastic matrix with known nonnegative rank admits a *separable* factorization. We test for the existence of anchor words in two different data sets derived from monetary policy discussions in the Federal Reserve and reject the null hypothesis that anchor words exist in one of them.

- "Identification and Estimation of Finite Mixtures of Multinomial Logit Models."

  Abstract: Finite mixtures of multinomial logit models can be used to capture consumer choice heterogeneity across multiple markets when only aggregate consumer choices per market are available. A motivating example is a nested logit where the composition of each mixture component (each nest of alternatives) is unknown a priori. We show that in order to identify these models, it suffices to require that each mixture component includes at least two component-exclusive alternatives. We refer to our assumption as the *pure-alternatives* condition, and we argue it is a natural extension of the *anchor-word* assumption used commonly in nonnegative matrix factorization problems in machine learning. Our identification result enables a consistent two-step estimator as the number of consumers, markets, and alternatives grow large. Applying this framework to the U.S. vehicle market, we find that consumer heterogeneity does not yield substitution patterns between electric and internal combustion engine vehicles, suggesting consumer segments are distinctly aligned with specific vehicle types without crossover substitution.

WORK IN PROGRESS

- "Systemic Risk, FOMC Statements, and Monetary Policy Shocks: A New Topic Model to Associate Text with Metadata," with Shawn Mankad.

  Abstract: In this research paper, we investigate the regulations guiding monetary policy communications through the development of a novel machine learning method called the Cluster Sentence Structural Topic Model (CSSTM). Our approach incorporates covariates in the data generation process and accounts for the correlation of sentences within each document by utilizing the equilibrium of sentences' topics. In the estimation process, we sort the equilibrium in the M step. Our method outperforms the Latent Dirichlet Allocation (LDA) and the Structural Topic Model (STM) by increasing the held-out likelihood by 20 percent and 10 percent. Using our method, we analyze FOMC statements and observe that the Fed places more emphasis on inflation expectations as opposed to current rates. According to our results, FOMC statements rely more on production instead of consumption. More importantly, we find that monetary policy communication started to consider systemic risk shortly after the 2007 financial crisis. By our method, we are able to decompose monetary policy shocks. The new measure has large and significant effects on systemic risk.

- "Pollution Avoidance and Willingness-to-Pay: Evidence from Travel Mode Choice in Beijing," with Shanjun Li and C.-Y. Cynthia Lin Lawell.

  Abstract: We estimate the short-term willingness-to-pay (WTP) to avoid air pollution by modeling the trade-off between avoidance behavior and its associated costs. Using fine-scale travel survey data from Beijing, we examine mode choices between indoor and outdoor commuting for mandatory work trips during heavily polluted hours. We estimate a short-term WTP of \$0.00223 per hour to avoid a 1 $\mu g/m^3$ increase in $PM_{2.5}$, which serves as a lower bound for the long-term WTP of approximately \$11.54 per year. Our estimation employs a machine learning instrumental variable approach in a high-dimensional econometric setting. We find that greater potential exposure discourages walking and cycling. Older individuals (55+) exhibit a 28% higher WTP than younger people, and wealthier individuals are willing to pay 36% more to avoid pollution. Finally, we find that behavioral adjustments occur only after substantial media coverage, highlighting the role of information in shaping responses to pollution.

- "Weak Sparse Models and Methods for Instrumental Variables."

- "Casual Inference with Missing Not at Random and Unobserved Confounders under Multiple Outcomes," with Peng Ding.

TEACHING EXPERIENCE

- Teaching Assistant for AEM 2300 International Trade and Finance, Cornell University 2022-2024 Spring
- Teaching Assistant for AEM 3310 Introduction to Business Regulation, Cornell University 2023 Fall
- Teaching Assistant for AEM 4110 Introduction to Econometrics (4.9/5.0), Cornell University 2021 Fall
- Teaching Assistant for AEM 6120 Applied Econometrics (4.4/5.0), Cornell University 2020 Fall

AWARDS

- Graduate Research Fellowship, Cornell University 2025
- Flaim and Neenan Family Fellowship, Cornell University 2025
- Graduate Research Fellowship, Cornell University 2024
- Ashley Graduate Fellowship, Cornell University 2023
- Transportation Networks and Smart Mobility Scholarship, Massachusetts Institute of Technology 2022
- Edward and Janet Heslop Fellowship, Cornell University 2021-2022
- Academic Excellence Scholarship (awarded annually to the Top 3 students in the honors program), Renmin University of China 2012-2015

PRESENTATIONS

- CES North American Conference (University of Michigan) 2025
- New York Camp Econometrics XVIII (Syracuse University), North American Summer Meeting of the Econometric Society (Vanderbilt University), Hong Kong University of Science and Technology, International Association for Applied Econometrics (Xiamen University; University of Macedonia), ESIF Economics and AI+ML Meeting (Cornell University) 2024
- Asia Meeting of the Econometric Society, EREE Student Workshop (Department of Agricultural and Resource Economics, UC Berkeley), Econometrics Student Conference (Department of Economics, UC Berkeley) 2023
- North American Summer Meeting of the Econometric Society (Tsinghua University) 2022
- World Conference of Spatial Econometrics Association (Tokyo, Japan) 2021

REFERENCES

José Luis Montiel Olea (Chair)
Associate Professor
Department of Economics
Cornell University
montiel.olea@gmail.com

Peng Ding
Associate Professor
Department of Statistics
University of California, Berkeley
pengdingpku@berkeley.edu

Francesca Molinari
Professor
Department of Economics
Cornell University
fm72@cornell.edu

C.-Y. Cynthia Lin Lawell
Associate Professor
Department of Applied Economics and Management
Cornell University
clinlawell@cornell.edu