JIAFENG (KEVIN) CHEN

jiafengchen@g.harvard.edu
Cell 818-658-4901
jiafengkevinchen.github.io

HARVARD UNIVERSITY

Littauer Center
1805 Cambridge St
Cambridge MA 02138

Placement Director: Claudia Goldin cgoldin@harvard.edu 617-495-3934
Placement Director: Lawrence F. Katz lkatz@harvard.edu 617-495-5079
Administrative Director: Brenda Piquet bpiquet@harvard.edu 617-495-8927

Education

Harvard University
Ph.D. Business Economics, 2019 to 2024 (expected)
S.M. Applied Mathematics, 2019
A.B. Applied Mathematics, *summa cum laude*, 2019

Fields

Econometrics
Causal Inference, Machine Learning, Labor Economics, Public Economics

References

Professor Isaiah Andrews
Massachusetts Institute of Technology
iandrews@mit.edu

Professor Elie Tamer
Harvard University
elietamer@fas.harvard.edu

Professor Jesse Shapiro
Harvard Business School
jesse_shapiro@fas.harvard.edu

Professor Edward Glaeser
Harvard University
eglaeser@harvard.edu

Fellowships & Awards

Opportunity Insights Fellowship, Harvard University, 2022
Lijun Lin and Ruilin Gong Graduate Student Fellowship, Harvard University, 2020-2021
Certificate of Distinction in Teaching, Harvard University, 2022
Thomas T. Hoopes Prize for excellent undergraduate thesis, Harvard University, 2019
Phi Beta Kappa (Junior year electee), 2018

Teaching

Data-Driven Leadership, Teaching fellow for Professor Michael Luca, Fall 2022
Ec 2140 Econometric Methods, Teaching fellow for Professor Wayne Gao, Spring 2022 (4.67/5)

Employment

Microsoft Research New England, Research Intern, 2020
QuantCo, Research Intern, 2019

Research

Research Assistant to Professor Isaiah Andrews, NBER, 2020-2023
Research Assistant to Professors Raj Chetty and Guido Imbens, Opportunity Insights, 2019-2020

Job Market Paper

Empirical Bayes When Estimation Precision Predicts Parameters
arXiv: 2212.14444

Empirical Bayes shrinkage methods usually maintain a *prior independence* assumption: The unknown parameters of interest are independent from the known precision of the estimates. This assumption is often theoretically questionable and empirically rejected, and imposing it inappropriately may harm the performance of empirical Bayes methods. We instead model the conditional distribution of the parameter given the standard errors as a location-scale family, leading to a family of methods that we call CLOSE. We establish that (i) CLOSE is rate-optimal for squared error Bayes regret, (ii) squared error regret control is sufficient for an important class of economic decision problems, and (iii) CLOSE is worst-case robust when our location-scale assumption is misspecified. We use our method to select high-mobility Census tracts targeting a variety of economic mobility measures in the Opportunity Atlas. Census tracts selected by CLOSE

are more mobile on average than those selected by the standard shrinkage method. The gain of CLOSE over the standard shrinkage method is substantial relative to two benchmarks.

Publications

Synthetic Control as Online Linear Regression
Econometrica, 2023

This paper notes a simple connection between synthetic control and online learning. Specifically, we recognize synthetic control as an instance of *Follow-The-Leader* (FTL). Standard results in online convex optimization then imply that, even when outcomes are chosen by an adversary, synthetic control predictions of counterfactual outcomes for the treated unit perform almost as well as an oracle weighted average of control units' outcomes. Synthetic control on differenced data performs almost as well as oracle weighted difference-in-differences, potentially making it an attractive choice in practice. We argue that this observation further supports the use of synthetic control estimators in comparative case studies.

Efficient estimation of average derivatives in NPIV models: Simulation comparisons of neural network estimators (with Xiaohong Chen and Elie Tamer)
Journal of Econometrics, 2023

Artificial Neural Networks (ANNs) can be viewed as nonlinear sieves that can approximate complex functions of high dimensional variables more effectively than linear sieves. We investigate the performance of various ANNs in nonparametric instrumental variables (NPIV) models of moderately high dimensional covariates that are relevant to empirical economics. We present two efficient procedures for estimation and inference on a weighted average derivative (WAD): an orthogonalized plug-in with optimally weighted sieve minimum distance (OP-OSMD) procedure and a sieve efficient score (ES) procedure. Both estimators for WAD use ANN sieves to approximate the unknown NPIV function and are root-n asymptotically normal and first-order equivalent. We provide a detailed practitioner's recipe for implementing both efficient procedures. We compare their finite-sample performances in various simulation designs that involve smooth NPIV function of up to 13 continuous covariates, different nonlinearities and covariate correlations. Some Monte Carlo findings include: (1) tuning and optimization are more delicate in ANN estimation; (2) given proper tuning, both ANN estimators with various architectures can perform well; (3) easier to tune ANN OP-OSMD estimators than ANN ES estimators; (4) stable inferences are more difficult to achieve with ANN (than spline) estimators; (5) there are gaps between current implementations and approximation theories. Finally, we apply ANN NPIV to estimate average partial derivatives in two empirical demand examples with multivariate covariates.

JUE Insight: The (non-)effect of opportunity zones on housing prices (with Edward Glaeser and David Wessel)
Journal of Urban Economics, 2023

Will the Opportunity Zones (OZ) program, America's largest new place-based policy in decades, generate neighborhood change? We compare single-family housing price growth in OZs with price growth in areas that were eligible but not included in the program. We also compare OZs to their nearest geographic neighbors. Our most credible estimates rule out price impacts greater than 0.5 percentage points with 95% confidence, suggesting that, so far, home buyers don't believe that this subsidy will generate major neighborhood change. OZ status reduces prices in areas with little employment, perhaps because buyers think that subsidizing new investment will increase housing supply. Mixed evidence suggests that OZs may have increased residential permitting.

Auctioneers sometimes prefer entry fees to extra bidders (with Scott Duke Kominers)
International Journal of Industrial Organization, 2021

We investigate a market thickness–market power tradeoff in an auction setting with endogenous entry. We find that charging admission fees can sometimes dominate the benefit of recruiting additional bidders, even though the fees themselves implicitly reduce competition at the auction

stage. We also highlight that admission fees and reserve prices are different instruments in a setting with uncertainty over entry costs, and that optimal mechanisms in such settings may be more complex than simply setting a reserve price. Our results provide a counterpoint to the broad intuition of Bulow and Klemperer (1996) that market thickness often takes precedence over market power in auction design.

A Semantic Approach to Financial Fundamentals (with Suproteem Sarkar)
ACL 2020 Workshop on Financial Technology and Natural Language Processing

The structure and evolution of firms' operations are essential components of modern financial analyses. Traditional text-based approaches have often used standard statistical learning methods to analyze news and other text relating to firm characteristics, which may shroud key semantic information about firm activity. In this paper, we present the Semantically-Informed Financial Index (SIFI), an approach to modeling firm characteristics and dynamics using embeddings from transformer models. As opposed to previous work that uses similar techniques on news sentiment, our methods directly study the business operations that firms report in filings, which are legally required to be accurate. We develop text-based firm classifications that are more informative about fundamentals per level of granularity than established metrics, and use them to study the interactions between firms and industries. We also characterize a basic model of business operation evolution. Our work aims to contribute to the broader study of how text can provide insight into economic behavior.

Working Papers

Logs with zeros? Some problems and solutions (with Jonathan Roth)
arXiv: 2212.06080; conditionally accepted at the *Quarterly Journal of Economics*

When the outcome of interest is nonnegative but can equal zero, economists frequently estimate average treatment effects (ATEs) for transformations of the outcome that behave like $\log(Y)$ when Y is large but are defined at zero (e.g. $\log(1+Y)$, $\text{arcsinh}(Y)$). This paper argues that ATEs for such log-like transformations should not be interpreted as approximating percentage effects, since they depend arbitrarily on the units of the outcome when the treatment has an extensive margin effect. Intuitively, this dependence arises because an individual-level percentage effect is not well-defined for individuals whose outcome changes from zero to non-zero when receiving treatment, and the units of the outcome implicitly determine how much weight is placed on the extensive margin effect of the treatment. We further establish that when the outcome can equal zero, there is no treatment effect parameter that is an average of individual-level treatment effects, unit-invariant, and point-identified. We discuss a variety of alternative approaches that may be sensible in settings with an intensive and extensive margin, including (i) expressing the ATE in levels as a percentage (e.g. using Poisson regression), (ii) explicitly calibrating the value placed on the intensive and extensive margins, and (iii) estimating separate effects for the two margins (e.g. using Lee bounds). We illustrate these approaches in three empirical applications.

Semiparametric Estimation of Long-Term Treatment Effects (with David Ritzwoller)
arXiv: 2107.14405; Minor revisions requested by the *Journal of Econometrics*

Long-term outcomes of experimental evaluations are necessarily observed after long delays. We develop semiparametric methods for combining the short-term outcomes of experiments with observational measurements of short-term and long-term outcomes, in order to estimate long-term treatment effects. We characterize semiparametric efficiency bounds for various instances of this problem. These calculations facilitate the construction of several estimators. We analyze the finite-sample performance of these estimators with a simulation calibrated to data from an evaluation of the long-term effects of a poverty alleviation program.

Nonparametric Treatment Effect Identification in School Choice
arXiv: 2112.03872

We study nonparametric identification and estimation of causal effects in centralized school assignment. We characterize the full set of identified treatment effects in common school choice

settings, under unrestricted heterogeneity in individual potential outcomes. This exercise highlights two points of caution for practitioners: We find that lack of overlap poses a challenge to regression-based estimators; we also find that, asymptotically, regression-based estimators that aggregate across many treatment contrasts put zero weight on treatment effects identified from regression-discontinuity (RD) variation, when the mechanism allows for both RD and lottery-based variation. Due to the complex interplay between heterogeneous causal effects and school choice algorithms, we recommend empirical researchers clearly decompose aggregate causal effect estimates by sources of variation in these settings. Lastly, we provide estimators and accompanying asymptotic results for causal contrasts identified by RD variation in school choice.

Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models
(with Daniel L. Chen and Greg Lewis)
arXiv: 2011.06158; Appeared in *NeurIPS Workshop on Machine Learning and Economic Policy, 2020*

We show how to use machine learning in the first-stage of the standard linear IV model to construct optimal instruments (Chamberlain, 1987, 1992). Doing so extracts non-linear covariation between the treatments and instruments, boosting statistical precision. The estimator is "mostly harmless" because it constrains the optimal instruments to be linear in the included covariates, enforcing only the plausible moment conditions. It also preserves standard intuitions and interpretations of linear instrumental variable methods, including under weak identification, and provides a simple, user-friendly upgrade to the applied economics toolbox. We illustrate our method with an example in law and criminal justice.

Mean-variance constrained priors have finite maximum Bayes risk in the normal location model
arXiv: 2303.08653; Submitted

Consider a normal location model $X \mid \theta \sim N(\theta, \sigma^2)$ with known σ^2. Suppose $\theta \sim G_0$, where the prior G_0 has zero mean and unit variance. Let G_1 be a possibly misspecified prior with zero mean and unit variance. We show that the squared error Bayes risk of the posterior mean under G_1 is bounded, uniformly $G_0, G_1, \sigma^2 > 0$.

Papers in Progress

Optimal Conditional Inference in Adaptive Experiments (first author; with Isaiah Andrews)

We study statistical inference using data from batched bandit experiments. A simple inference procedure is to use only the last batch of the data. This procedure adapts to the bandit algorithm, data-dependent choice of stopping time, and data-dependent choice of inference target. We show that, for many bandit algorithms, even without knowing the precise algorithm, there is a simple procedure that dominates the last-batch-only procedure while preserving the adaptivity properties. Moreover, if we have knowledge of the bandit algorithm, then we can design optimal conditional procedures. These procedures are computationally tractable for a large class of discrete assignment algorithms that we call polyhedral algorithms.

Robust Inference for Imperfectly Linked Data (second author; with Ross Mattheis; Ross Mattheis's job market paper)

Estimating intergenerational mobility often requires linking data across multiple sources. However, mistakes in record linkage can introduce biases in the subsequent estimates. This paper studies inference for linear models with imperfectly linked data. We show that any ambiguity in the linked data result in a loss of point identification without additional assumptions. To recover identification, it is sufficient to assume independence of the information used for linkage and the information used for model estimation. Building on this result, we propose a new method which we call Robust Inference for Imperfectly Linked data (RIILink). We apply the RIILink estimator to re-examine intergenerational mobility in the United States between 1850 and 1940.

Seminars & Conferences	Southern Economic Association Meeting (scheduled), 2023 Pennsylvania State University, 2023 International Seminar on Selective Inference (discussant), 2023 NBER Summer Institute Labor Studies, 2022 NBER Summer Institute Urban Economics (discussant), 2022 Princeton DataX Workshop on Synthetic Control Methods (poster), 2022 NABE TEC, 2022 Virtual Quant Marketing Seminar (panelist), 2022 AEA/ASSA, 2021 Society of Labor Economists Meeting, 2021 Conference on Digital Experimentation (CODE@MIT), 2021 NeurIPS Workshop on Machine Learning for Economic Policy, 2020 ACL FinNLP Workshop, Brookings Institution, 2020 North American Winter Meeting of the Econometric Society, 2019
Academic Service	Referee service for *American Economic Review*, *Econometrica*, *Quarterly Journal of Economics*, *Review of Economics and Statistics*, *Journal of Political Economy: Microeconomics*, *Journal of Econometrics*, *Management Science*, *Journal of Human Resources*, *Journal of Business and Economic Statistics*, *Journal of Applied Econometrics*, *Journal of Econometric Methods*, *Review for Real Estate Economics*, *Journal of Urban Economics*, and *Annals of Operations Research* Co-organizer, Harvard Econometrics Reading Group, 2020-2022
Languages	English (fluent); Chinese (native)
Software skills	Python, R
Personal information	Chinese citizenship