Research Statement

Ka Yan Cheng

Ph.D. Candidate, Department of Economics, Emory University

ka.yan.cheng@emory.edu · (404) 512-4439 · https://kayancheng.github.io/

I am an econometrician working in microeconometrics, with interests in semiparametric estimation, machine learning, and distributional analysis.

Today's computing power and rich data have made it easy to get *a* result, but the goal of econometrics has always been to get a *better* one: an estimate that answers the economic question we care about, rests on assumptions we can defend, is as precise as the data allow, and comes with inference we can trust. My research asks how economic structure, together with econometric theory, can guide machine learning methods toward such answers.

Job Market Paper: Cross-Validation in Debiased Machine Learning

In my job market paper, "Cross-Validation for Ridge Regularization in Debiased Machine Learning," I study how to tune the first step of debiased machine learning (DML). DML first estimates two auxiliary functions and then combines them to estimate a causal or structural parameter. Each first-step estimator has tuning parameters (such as how strongly to penalize complexity), which researchers commonly select by cross-validation, with each function tuned separately. The usual guarantees for valid inference require the two first-step estimators to be jointly accurate, in the sense that the product of their errors shrinks quickly enough. Whether this requirement can be achieved depends on the difficulty of the estimation problem as well as on the tuning choices. I ask when suitable tuning choices can meet this requirement, and how to construct candidate sets so that the tuning parameters chosen by cross-validation yield sufficiently accurate estimators.

Specifically, I analyze ridge-regularized linear estimators of the two functions, the outcome regression and the Riesz representer (the debiasing function), on a shared, high-dimensional dictionary. I derive learning-rate bounds for these estimators and use them to characterize a frontier in terms of the two functions' approximation exponents. Above this frontier, the bounds certify that appropriate tuning choices satisfy this product-rate condition and hence support valid inference. Without requiring knowledge of these exponents, I construct candidate sets of dictionary sizes and ridge penalties. Whenever the estimation problem lies above the frontier, these sets contain tuning choices that deliver the required estimation accuracy. I establish oracle inequalities comparing the choices made by separate cross-validation with the best candidate in each set, and analyze how these guarantees extend to refitting the estimators with the selected tuning parameters. Together, these results connect conditions on the underlying estimation problem to a practical tuning procedure with theoretical guarantees.

Economic Structure in Debiased Machine Learning

In "Debiased Machine Learning: Identification, Estimation, and Shape Constraints" (with Qihui Chen and Zheng Fang), we develop a framework for incorporating economic knowledge into DML.

Such knowledge matters because machine learning methods, however powerful, can still struggle to estimate first-step functions accurately enough for valid inference. Imposing shape restrictions implied by economic theory, such as monotonicity and convexity, can help guide estimation while retaining flexibility in functional form. However, placing these restrictions on the primary nuisance function also constrains the Riesz representer (the function used to construct the bias correction). We establish conditions for identifying the representer and characterize it as the unique solution to a quadratic optimization problem, which allows us to estimate the constrained representer without requiring its analytical form. Applications to Medicaid expansions and mortality, and to working hours and wage growth, illustrate how economic and institutional knowledge can inform estimation.

Distributional Methods and Inequality

Advances in econometric methods and computation allow us to examine economic outcomes in greater detail than their averages alone reveal. A policy may raise average income by benefiting the middle of the distribution while leaving the poorest households no better off. Understanding distributional changes is therefore central to evaluating policies intended to reduce disadvantage.

In ongoing work with Esfandiar Maasoumi, "Intergenerational Mobility Through the Lens of ϕ-Divergence," we develop an approach to measuring intergenerational dependence using the joint distribution of parents' and children's economic outcomes. We use bivariate distribution regression to estimate this distribution conditional on observed characteristics, and then construct an entropic measure of its departure from conditional independence. This benchmark describes a setting in which parents' outcomes provide no additional information about their children's outcomes once these characteristics are accounted for. The measure therefore summarizes how strongly children's outcomes depend on their parents' across the entire joint distribution.

Future Research

I plan to build on my job market paper by investigating whether cross-validated tuning guarantees extend to nonlinear nuisance estimators, including neural networks. I am interested in how data-driven choices of model complexity and regularization can deliver the joint estimation accuracy needed for valid inference. This direction would connect my current results for ridge-regularized estimators with a broader class of methods used in practice.

Alongside this work, I plan to study flexible estimation and inference under shape restrictions and for distributional targets, with a focus on specific economic models and applications. Building on my joint work, I want to understand how economic knowledge can guide nuisance estimation and bias correction in these settings. I will also continue to explore the potential of distributional methods for studying inequality and evaluating policy effects, with the aim of understanding differences in economic outcomes that averages alone cannot reveal.