JANN SPIESS
scholar.harvard.edu/spiess
jspiess@fas.harvard.edu

HARVARD UNIVERSITY

Placement Director: John Campbell	JOHN_CAMPBELL@HARVARD.EDU	617-496-6448
Placement Director: Nathan Hendren	NHENDREN@FAS.HARVARD.EDU	617-496-3588
Assistant Director: Brenda Piquet	BPIQUET@FAS.HARVARD.EDU	617-495-8927

Contact Information
Littauer Center 315
1805 Cambridge Street
Cambridge, MA 02138
857-284-9880

Undergraduate Studies:
BSc, Mathematics (TopMath), Technical University of Munich, with high distinction, 2010

Graduate Studies:
MASt, Mathematics (Part III of the Mathematical Tripos), University of Cambridge, with distinction, 2011
MPP, Public Policy, Harvard Kennedy School, 2013
AM, Economics, Harvard University, 2015

PhD Candidate, Economics, Harvard University, 2013 to present
Thesis Title: "Essays on the Econometrics of Machine Learning"
Expected Completion Date: May 2018

References:

Professor Sendhil Mullainathan
Department of Economics, Harvard University
617-496-2720, mullain@fas.harvard.edu

Professor Elie Tamer
Department of Economics, Harvard University
617-496-1526, elietamer@fas.harvard.edu

Professor Alberto Abadie
Department of Economics, MIT
617-715-2047, abadie@mit.edu

Professor Gary Chamberlain
Department of Economics, Harvard University
617-495-1869, gary_chamberlain@harvard.edu

Teaching and Research Fields:
Econometrics, Applied Microeconomics, Machine Learning, Behavioral Economics (teaching)

Teaching Experience:

2016	Economics 2140: Econometric Methods (PhD class in econometrics), Teaching fellow for Mikkel Plagborg-Møller, Harvard University
2014, 16	Economics 2150: Machine Learning in Econometrics (PhD class in econometrics), Teaching fellow for Sendhil Mullainathan, Harvard University
2012	Economics 2030: Psychology and Economics (PhD class in behavioral economics), Teaching fellow for David Laibson and Sendhil Mullainathan, Harvard University
2009/10	Discrete Mathematics (undergraduate class in mathematics), Teaching fellow for Anusch Taraz, Technical University of Munich

Research Experience:

2013	Research assistant to Alberto Abadie, Harvard Kennedy School
2012	Research assistant to David Laibson, Harvard University

Honors, Scholarships, and Fellowships:

2015–17	Machine Learning and Economics Fellowship, Harvard University
2012, 14, 16	Harvard University Certificate of Distinction in Teaching (awarded four times)
2011–13	McCloy Fellow, Harvard Kennedy School
2011	Bateman Scholar, Trinity Hall, University of Cambridge
2010–11	Kurt Hahn Scholar, University of Cambridge

Research Papers:

"Optimal Estimation when Researcher and Social Preferences are Misaligned" (Job Market Paper)

Econometric analysis typically focuses on the statistical properties of fixed estimators and ignores researcher choices. In this article, I approach the analysis of experimental data as a mechanism-design problem that acknowledges that researchers choose between estimators, sometimes based on the data and often according to their own preferences. Specifically, I focus on covariate adjustments, which can increase the precision of a treatment-effect estimate, but open the door to bias when researchers engage in specification searches. First, I establish that unbiasedness is a requirement on the estimation of the average treatment effect that aligns researchers' preferences with the minimization of the mean-squared error relative to the truth, and that fixing the bias can yield an optimal restriction in a minimax sense. Second, I provide a constructive characterization of all treatment-effect estimators with fixed bias as sample-splitting procedures. Third, I show that a researcher restricted specifically to the class of unbiased estimators of the average treatment effect solves a prediction problem. The equivalence of unbiased estimation and prediction across sample splits characterizes all admissible unbiased procedures in finite samples, leaves space for beneficial specification searches, and offers an opportunity to leverage machine learning. As a practical implication, I describe flexible pre-analysis plans for randomized experiments that achieve efficiency without bias.

"Unbiased Shrinkage Estimation" (arXiv:1708.06436)

Shrinkage estimation usually reduces variance at the cost of bias. But when we care only about some parameters of a model, I show that we can reduce variance without incurring bias if we have additional information about the distribution of covariates. In a linear regression model with homoscedastic Normal noise, I consider shrinkage estimation of the nuisance parameters associated with control variables. For at least three control variables and exogenous treatment, I establish that the standard least-squares estimator is dominated with respect to squared-error loss in the treatment effect even among unbiased estimators and even when the target parameter is low-dimensional. I construct the dominating estimator by a variant of James–Stein shrinkage in a high-dimensional Normal-means problem. It can be interpreted as an invariant generalized Bayes estimator with an uninformative (improper) Jeffreys prior in the target parameter.

"Bias Reduction in Instrumental Variable Estimation through First-Stage Shrinkage" (arXiv:1708.06443)

The two-stage least-squares (2SLS) estimator is known to be biased when its first-stage fit is poor. I show that better first-stage prediction can alleviate this bias. In a two-stage linear regression model with Normal noise, I consider shrinkage in the estimation of the first-stage instrumental variable coefficients. For at least four instrumental variables and a single endogenous regressor, I establish that the standard 2SLS estimator is dominated with respect to bias. The dominating IV estimator applies James–Stein type shrinkage in a first-stage high-dimensional Normal-means problem followed by a control-function approach in the second stage. It preserves invariances of the structural instrumental variable equations.

"Robust Post-Matching Inference" (with Alberto Abadie)

Nearest-neighbor matching (Cochran, 1953; Rubin, 1973) is a popular nonparametric tool to create balance between treatment and control groups in observational studies. As a preprocessing step before regression analysis, matching reduces the dependence on parametric modeling assumptions (Ho et al., 2007). Moreover, matching followed by regression allows estimation of elaborate models that are useful to describe heterogeneity in treatment effects. In current empirical practice, the matching step is often ignored for the estimation of standard errors and confidence intervals. That is, to do inference, researchers proceed as if matching did not take place. We show that ignoring the matching first step produces valid standard errors if matching is done without replacement and if the regression model is correctly specified relative to the population regression function of the outcome variable on the treatment variable and all the covariates used for matching. However, standard errors that ignore the matching step are not valid if matching is conducted with replacement or, more crucially, if the second step regression model is misspecified. We show that two easily implementable alternatives, (i) clustering the standard errors at the level of the matches, or (ii) a nonparametric block bootstrap procedure, produce approximations to the distribution of the post-matching estimator that are robust to misspecification, provided that matching is done without replacement. These results allow robust inference for post-matching methods that use regression in the second step. A simulation study and an empirical example demonstrate the empirical relevance of our results.

"Machine Learning Tests for Effects on Multiple Outcomes" (with Jens Ludwig and Sendhil Mullainathan; arXiv:1707.01473)

A core challenge in the analysis of experimental data is that the impact of some intervention is often not entirely captured by a single, well-defined outcome. Instead there may be a large number of outcome variables that are potentially affected and of interest. We propose a data-driven approach rooted in machine learning to the problem of testing effects on such groups of outcome variables. It is based on two simple observations. First, the "false-positive" problem that a group of outcomes is similar to the concern of "over-fitting," which has been the focus of a large literature in statistics and computer science. We can thus leverage sample-splitting methods from the machine-learning playbook that are designed to control over-fitting to ensure that statistical models express generalizable insights about treatment effects. The second simple observation is that the question whether treatment affects a group of variables is equivalent to the question whether treatment is predictable from these variables better than some trivial benchmark (provided treatment is assigned randomly). This formulation allows us to leverage data-driven predictors from the machine-learning literature to flexibly mine for effects, rather than rely on more rigid approaches like multiple-testing corrections and pre-analysis plans. We formulate a specific methodology and present three kinds of results. First, our test is exactly sized for the null hypothesis of no effect. Second, a specific version is asymptotically equivalent to a benchmark joint Wald test in a linear regression. Third, this methodology can guide inference on where an intervention has effects. Finally, we argue that our approach can naturally deal with typical features of real-world experiments, and be adapted to baseline balance checks.

Publication:

"Machine Learning: An Applied Econometric Approach" (with Sendhil Mullainathan)
Journal of Economic Perspectives (Spring 2017)

Machines are increasingly doing "intelligent" things: Facebook recognizes faces in photos, Siri understands voices, and Google translates websites. The fundamental insight behind these breakthroughs is as much statistical as computational. Face recognition algorithms, for example, use a large dataset of photos labeled as having a face or not to estimate a function $f(x)$ that predicts the presence y of a face from pixels x. This similarity to econometrics raises questions: How do these new empirical tools fit with what we know? As empirical economists, how can we use them? We present a way of thinking about machine learning that clarifies its place in the econometric toolbox. Machine learning not only provides new tools, it solves a specific problem. Machine learning revolves around prediction on new sample points from the same distribution, while many economic applications revolve around parameter estimation and counterfactual prognosis. So applying machine learning to economics requires finding relevant prediction tasks.

Book:

"Machine Learning for Economics" (with Sendhil Mullainathan)
Princeton University Press (in preparation)