Maximilian Kasy
Associate Professor of Economics, Harvard University
homepage: https://maxkasy.github.io/home/
email: maximiliankasy@fas.harvard.edu
telephone: 310-666-8071

Research statement

My research is guided by the goal of providing methodological advice for empiricists. The process of generating, interpreting, communicating, and using empirical evidence has many steps. Some of these steps are central topics in the econometrics literature, in particular estimation and inference. Other steps have received less attention, including experimental design, the interaction of statistical inference with the publication process, and policy choice based on data. My research provides guidance for problems arising in these areas, and draws on a number of domains, including mathematical statistics, applied microeconomics, economic theory, and machine learning.

In my research I have for instance addressed questions such as how to assign treatments in experiments, how to choose among alternative machine learning estimators in empirical economics, how to interpret empirical research that is subject to publication bias, and how to pick optimal tax rates based on noisy data. In ongoing research I build on this work and consider questions such as how to design adaptive field experiments for policy choice, how to reform the publication process in light of limited attention and constraints on social learning, and how to understand the large sample behavior of machine learning estimators.

Summary of research agendas

My recent and ongoing research is based on four intersecting agendas. The first agenda is to develop a deeper understanding of **machine learning** methods, their connection to economic theory, and conditions under which they do or do not perform well in economic applications. Can we combine insights from machine learning and economic theory to construct estimators that perform particularly well when the theory is approximately correct? And can we give guidance to applied researchers as to when machine learning methods are expected to be useful, and which method to prefer in which context? These questions are considered in *How to use economic theory to improve estimators: Shrinking toward theoretical restrictions* (forthcoming, REStat), and *Choosing among regularized estimators in empirical economics – The risk of machine learning* (forthcoming, REStat).

The second agenda is to explicitly recast problems faced by empirical researchers as **decision problems**. This involves careful specification of objective functions, action spaces, identifying assumptions, and prior beliefs. This approach can yield surprising but practical answers to familiar questions such as how to optimally assign treatments in a field experiment, or how to choose optimal tax rates based on empirical evidence. Papers following this approach include *Why experimenters might not always want to randomize, and what they could do instead* (Political Analysis, 2016), and *Optimal taxation and insurance using machine learning – sufficient statistics and beyond* (forthcoming, Journal of Public Economics).

The third agenda builds on recent debates regarding p-hacking, publication bias, and replicability, and aims to develop a perspective on **statistics in its social context**. Statistics (and empirical research more generally) is a social endeavor, involving different researchers, journal editors and referees, readers, policymakers, and others. Researchers and journals necessarily select which findings to report. How can we find out what form this selection process takes, and to what extent findings are selected? Can we correct for such selection when interpreting

published findings? And which findings should be published – should we aim to eliminate selection in the publication process, or does it serve a valuable purpose?[1] These questions are addressed in *Identification of and correction for publication bias* (revise and resubmit, AER), and *Which findings should be published?* (submitted).

The fourth agenda considers more classic questions of **identification**, motivated by empirical research on urban segregation and **economic inequality**.[2] I address questions such as the following. How can we distinguish between alternative explanations of urban segregation in models of sorting with externalities? What are the testable implications of path dependence in employment trajectories? How can we identify the distributional welfare impact of tax changes when there are equilibrium responses in the labor market? What is the identified set for welfare-rankings of alternative treatment assignment policies? What are the necessary and sufficient conditions for identifying causal effects in continuous instrumental variables settings? Examples of papers in this agenda are *Identification in Triangular Systems Using Control Functions* (Econometric Theory, 2011), *Identification in a model of sorting with social externalities and the causes of urban segregation* (Journal of Urban Economics, 2015), and *Partial identification, distributional preferences, and the welfare ranking of policies* (REStat, 2016).

Road map

The rest of this research statement is structured as follows. I will discuss each of my agendas in greater detail. For each of the first three agendas, I will review two papers that are published or forthcoming, and one project that is work in progress. For the fourth agenda I provide shorter summaries of a larger number of papers. I conclude with a brief discussion on why drawing on multiple domains is helpful in addressing the questions posed in my research.

The research described in this statement benefited from support by grants by the **National Science Foundation** (grant "Statistical Decisions and Policy Choice," 2014-2017), and the **Sloan Foundation** (grant "Publication bias and specification searching. Identification, correction, and reform proposals", 2018-2021)

Agenda 1: Shrinkage estimation and machine learning

Optimal decision procedures minimize some aggregate notion of risk, such as Bayes (average) risk or minimax (worst case) risk. In order to get a better understanding of the performance of some decision procedure, it is often useful to also study a more disaggregated notion of risk, i.e., the risk function itself: For which parameters does a procedure perform well, for which does it perform poorly?

This is particularly pertinent in the context of machine learning estimators and related shrinkage procedures. How should an empirical researcher decide which machine learning method to use in the context of their application? The use of machine learning procedures is often justified by reference to their good performance in specific empirical examples. Not much is known, in general, about how their performance varies across possible alternative data generating processes.

Many machine learning estimators combine two key features, regularization and tuning. Regularization serves to reduce the estimator variability (often described as "over-fitting"), possibly at the cost of introducing systematic distortions (bias). Tuning, using procedures such as cross-validation, serves to optimize this trade-off between bias and variance. As famously

[1] I am also organizing a conference on these topics at Harvard in May 2019. The announcement can be found at `https://maxkasy.github.io/home/StatisticsSocialConference/`.

[2] Relatedly, I have also written an open online textbook on empirical research on economic inequality, available at `http://inequalityresearch.net/`.

shown by James and Stein (James and Stein, 1961; Stein, 1981), the combination of regularization and tuning can lead to uniform improvements of estimator performance, relative to unregularized alternatives, for multi-dimensional estimation problems.

These improvements of performance are concentrated in a vicinity of the point (or set) in the parameter space that the regularization procedure shrinks toward. Conventional procedures shrink toward 0 or some other arbitrary point. The first paper discussed in this section proposes an alternative, shrinking toward the restrictions implied by economic theory. The second paper discussed in this section focuses on the question how to regularize, comparing alternatives such as Ridge, Lasso, and Pretest. Which of these methods performs better for what kind of application? This paper also provides sufficient conditions for tuning to work well, which it generally does for high-dimensional estimation problems. The third project will take the literature full-circle, by showing that a wide array of machine learning procedures in large samples effectively reduce to James-Stein shrinkage.

How to use economic theory to improve estimators: Shrinking toward theoretical restrictions (with Pirmin Fessler, forthcoming, REStat)

This paper starts from the presumption that theories in economics will usually only be approximately true. In physics, statements such as "the speed of light is constant" might be true over a wide range of settings; by contrast, theories such as "people maximize expected utility" have many counterexamples, even if the statement is often approximately true. If that is the case, testing such theories will always lead to rejection in large enough samples, and imposing such theories in estimation will lead to inconsistent estimates. What role should theory then play in empirical research in economics? In this paper, we suggest a framework for the construction of estimators which perform particularly well when the empirical implications of a theory under consideration are approximately correct.

Building on the empirical Bayes paradigm Robbins (1964); Morris (1983), we consider families of priors that are centered on the set of parameters consistent with the predictions of the theory. These priors are governed by a parameter of dispersion, providing a measure for how well the theory appears to describe the data. Estimation proceeds in three steps. In a first step, the parameters of interest are estimated in an unrestricted way, ignoring the predictions of economic theory. In a second step, the hyper-parameters governing the family of priors are estimated. The hyper-parameters include both the parameters of the restricted model and the measure of dispersion. In a third step, "posterior means" for the parameters of interest are calculated, conditioning on the preliminary estimates and on the estimated values for the hyper-parameters.

This approach has a number of advantages. (i) The resulting estimates are consistent, i.e., converge to the truth as samples get large, for any parameter values, in contrast to estimation imposing the theory. (ii) The variance and mean squared error of the estimates is smaller than under unrestricted estimation. (iii) In contrast to a fully Bayesian approach, no tuning parameters (features of the prior) have to be picked by the researcher. (iv) Our empirical Bayes approach avoids the irregularities (poor mean squared error in intermediate parameter regions) which are associated with testing theories and imposing them if they are not rejected. (v) Counterfactual predictions and forecasts are driven by the data whenever the latter are informative.

We implement the proposed approach in a number of economic contexts. These contexts are distinguished by the type of "theory" considered, including parametric structural models of production and labor demand (CES and nested CES production functions), the theory of consumer demand (negative semi-definiteness of compensated quantile demand elasticities), equilibrium models of financial markets (capital asset pricing model), structural models of pref-

erences (dynamic multinomial logit), and abstract theories of decision making (the stochastic axiom of revealed preference, and related restrictions).

Choosing among regularized estimators in empirical economics – The risk of machine learning (with Alberto Abadie, forthcoming, REStat)

Many applied settings in empirical economics involve simultaneous estimation of a large number of parameters. In particular, applied economists are often interested in estimating the effects of many-valued treatments (like teacher effects or location effects), treatment effects for many groups, and prediction models with many regressors. In such settings, methods that combine regularized estimation and data-driven choices of regularization parameters are useful to avoid over-fitting; this insight is central to the machine learning literature (eg. Friedman et al., 2001). In this article, we analyze the performance of a class of estimators that includes ridge, lasso and pretest in contexts that require simultaneous estimation of many parameters. Our analysis aims to provide guidance to applied researchers on (i) the choice between regularized estimators in practice and (ii) data-driven selection of regularization parameters. To address (i), we characterize the risk (mean squared error) of regularized estimators and derive their relative performance as a function of simple features of the data generating process. To address (ii), we show that data-driven choices of regularization parameters, based on Stein's unbiased risk estimate or on cross-validation, yield estimators with risk uniformly close to the risk attained under the optimal (unfeasible) choice of regularization parameters. We use data from recent examples in the empirical economics literature to illustrate the practical applicability of our results.

Let me briefly summarize the intuition for some of our results. In the stylized setting we consider (estimation of many means), for any given data generating process there is an (infeasible) risk-optimal regularized estimator. This estimator has the form of a posterior expectation, where we take the empirical distribution of parameters across components of the parameter vector as the prior. The risk function of any regularized estimator can be expressed as a function of the distance between that regularized estimator and the optimal one.

For tuning, we consider choices of regularization parameters based on the minimization of a criterion function that estimates risk, such as cross-validation or Stein's unbiased risk estimate. Ideally, a machine learning estimator evaluated at a data-driven choice of the regularization parameter would have a risk function that is uniformly close to the risk function of the infeasible estimator using an oracle-optimal regularization parameter (which minimizes true risk). We show this type of uniform consistency can be achieved under fairly mild conditions whenever the dimension of the problem under consideration is large, and risk is defined as mean squared error averaged across components.

Approximate cross-validation, machine learning, and James-Stein shrinkage (with Lester Mackey, work in progress)

A large class of machine learning estimators can be written as penalized m-estimators. The penalization term ensures regularization of the estimator, reducing variance at the cost of adding some bias. To decide on the amount of regularization, a tuning parameter needs to be chosen. The most popular tuning method is cross-validation (CV); in this project we will focus on leave-one-out CV. An alternative which has received some attention recently is approximate cross-validation (ACV), Giordano et al. (2018). ACV approximates leave-one-out estimates using influence functions. Another alternative is Stein's Unbiased Risk Estimate (SURE).

It is useful to contrast high-dimensional settings to settings of more moderate dimension. As discussed above, in high-dimensional settings strong guarantees for optimality of leave-one-out cross-validation and related procedures are available. In such settings, variability of the

estimated tuning parameter becomes negligible when considering the variability of the estimates of interest. Tuning is more "interesting," from a statistical point of view, in settings of smaller dimension, where variability of the estimated tuning parameter is not negligible. This project is concerned with the latter case. Our goal is to characterize the prediction risk of penalized m-estimators. To do so, we will consider large sample asymptotics for sequences of parameters local to the minimizer of the penalty. Away from the minimizer of the penalty, any reasonable tuning procedure will not penalize in large samples.

We are in the process of proving the following conjectures (under appropriate regularity conditions). (i) Any penalized m-estimator is asymptotically equivalent, in the sense of uniform convergence of its risk function, to a corresponding estimator in the normal-means model. (ii) In the normal-means model, ACV is numerically identical to SURE, plugging a sample analog of the variance matrix into the definition of SURE. (iii) For smooth penalty functions, CV, ACV, and SURE yield asymptotically identical estimators, in the sense of mean squared convergence. (iv) The asymptotic risk function of penalization using SURE (or CV or ACV) for tuning, for smooth penalties, is identical to the risk function of the (positive part) James-Stein estimator (in the appropriate heteroskedastic setting).

The combination of these results gives us a complete picture of the risk behavior of penalized m-estimation using CV or ACV in large samples. Among other things, it provides guarantees that ACV, which is computationally more tractable than its alternatives in many settings, performs well.

Agenda 2: Decision problems

Many methodological problems in empirical economics can usefully be recast as decision problems. This is an approach that comes naturally to economists, since it is closely connected to models of (expected) utility maximization. In order to restate methodological problems as decision problems, we need to specify an objective function, an action space, identifying assumptions and, for a Bayesian approach, prior beliefs. The most important task is careful specification of these items. Once they are specified, solving for optimal decisions becomes, in principle, straightforward. One tool that has proven particularly helpful in my work on such problems are Gaussian process priors for unknown functions. Such priors are popular in machine learning and in geostatistics. They allow for tractable closed-form solutions without relying on functional-form restrictions.

Estimation and hypothesis testing are the two canonical classes of decision problems in statistics. In the following I discuss different classes of decision problems, experimental design to minimize estimation error, policy choices that aim to maximize social welfare (defined in the utilitarian sense), and experimental design for policy choice.

Why experimenters might not always want to randomize, and what they could do instead (Political Analysis, 2016)

Researchers running fields experiments, for instance in development economics, often face the following situation, cf. Morgan and Rubin (2012). Suppose that an experimenter has collected a sample as well as baseline information about the units in the sample. How should she allocate treatments to the units in this sample? I approach this question as a decision problem.

I show first that experimenters might not want to randomize, in general. While maybe surprising, the basic intuition for this result is simple and holds for any statistical decision problem. The intuition is that the risk (Bayes or minimax) of any randomized decision procedure is a weighted average of the risk of the deterministic procedures that it is averaging

over, and can thus never be lower than the risk of the best deterministic procedure. If the latter is unique, randomization is strictly dominated. For this reason, we usually don't consider randomized estimators or inference procedures. More specifically, in the context of experimental design for causal inference, the purpose of randomization is to pick treatment and control groups which are similar ex-ante, before they are exposed to different treatments. Allowing for randomness in the treatment assignment to generate imbalanced distributions of covariates can only hurt the balance of the distribution of potential outcomes. The paper also discusses various counter-arguments, outside the decision-theoretic paradigm, in favor of randomization (mixed strategies against adversarial audiences, randomization inference).

The paper proceeds by discussing how to implement optimal and near-optimal designs in practice, where near-optimal designs might involve some randomization. The key problem is to derive tractable expressions for the expected mean squared error (MSE) of estimators for the average treatment effect, given a treatment assignment. Once we have such expressions we can numerically search for the best assignment, or for a set of assignments that are close to optimal. In order to calculate the expected MSE, we need to specify a prior distribution for the conditional expectation of potential outcomes given covariates. I provide simple formulas for the expected MSE for a general class of non-parametric Gaussian process priors.[3]

Optimal taxation and insurance using machine learning – sufficient statistics and beyond (forthcoming, Journal of Public Economics)

A central question of the field of public finance is the optimal choice of policy parameters such as tax rates, health insurance copay, or unemployment benefit levels, with the goal of maximizing social welfare. Social welfare is defined as a weighted sum of individual utilities. Optimal policies can often be expressed as solutions to first-order conditions involving only a few key behavioral elasticities. The "sufficient statistics" approach to optimal taxation proceeds by plugging in estimates of these elasticities into the first-order conditions to arrive at recommended policy levels, cf. Chetty (2009).

This approach has been very successful in connecting the theory of optimal taxation to empirical research. There are some shortcomings to this approach, however, in terms of how it deals with sampling uncertainty, and in terms of its implicit reliance on functional form assumptions. In this paper, I build on the insights of the sufficient statistics literature, and propose an alternative to the plug-in approach which aims to address these shortcomings. This alternative is based on maximizing posterior expected social welfare, combining insights from (i) optimal policy theory as developed in the field of public finance, and (ii) machine learning using Gaussian process priors. I provide explicit formulas for posterior expected social welfare and optimal policies in a wide class of policy problems.

This approach yields different policy recommendations for several reasons. A first reason is that I take a different approach to decision making under uncertainty. The conventional approach plugs noisy estimates into nonlinear functionals, leading to systematic bias in finite samples. This bias is avoided by maximizing posterior expected welfare. A second reason is that the conventional approach implicitly relies on functional form assumptions, when estimating elasticities using regressions that are linear in logarithms, for instance. The proposed nonparametric approach in this paper avoids the distortions resulting from incorrect functional form choices.

The proposed methods are applied to the choice of coinsurance rates in health insurance, using data from the RAND health insurance experiment. The key trade-off in this setting is between transfers towards the sick and insurance costs. The key empirical relationship the

[3]A web app implementing the proposed procedures is available at `https://maxkasy.github.io/home/treatmentassignment/`.

policy maker needs to learn about is the response of health care expenditures to coinsurance rates. Holding the economic model and distributive preferences constant, I obtain much smaller point estimates of the optimal coinsurance rate (18% vs. 50%) when applying my estimation method instead of the conventional "sufficient statistic" approach.

Adaptive experimental design for policy choice (with Anja Sautmann, work in progress)
Following up on this line of research, I am currently working on a project on adaptive experimental design for policy choice. The paper discussed above focuses on experimental design for the purpose of point estimation of treatment effects. By contrast, the goal of many field experiments is to evaluate alternative policies for the purpose of choosing a policy which will be implemented more widely. Examples are medical treatments or educational interventions. This suggests a different objective function. Rather than minimizing the mean squared error, we might aim to maximize expected social welfare, assuming policy is chosen optimally conditional on the outcome of the experiment. With multiple treatment options, this results in qualitatively different recommendation for experimental design, relative to conventional approaches. Experimenters should assign most units to the treatment options that have the highest chance of being optimal, rather than wasting resources to pin down the effect of treatments which won't be optimal anyway. In practice, this is of particular interest when we can run an experiment in several waves, adapting treatment in subsequent waves based on earlier outcomes.

We intend to show theoretically, using simulations, and an actual implementation in the field, that we can substantially improve upon conventional experimental designs in such settings. The optimal experimental design is given by the solution to a dynamic stochastic optimization problem. Solving such problems can be numerically difficult. Easily computable approximate solutions that perform well can be constructed, however, drawing on insights from the literature on online learning and bandit problems, e.g. Russo et al. (2018).

Agenda 3: Statistics in its social context

The projects discussed thus far take a decision-theoretic perspective. Decision theory is a rather solitary affair. It involves just you, the data, and a decision to be made. In practice, however, statistics (and empirical research more generally) is a social endeavor, involving different researchers, journal editors and referees, readers, policymakers, and others. This social dimension of statistics has come to the fore in recent debates about publication bias, p-hacking, and the perceived replicability crisis of various empirical fields.

The first paper to be discussed in this section contributes to this debate by providing tools to assess the status-quo of selectivity in published research. How can we identify what determines the probability of publication of particular findings, and how can we adjust our interpretation of published findings to the resulting distortions?

The second paper provides a normative counterpart to this inquiry, asking which findings should be published. This paper studies statistical reporting as a form of communication, and considers the role of published findings in informing policy decisions. As it turns out, from such an instrumental perspective it is optimal to publish surprising findings, despite the resulting distortions of statistical inference.

Following up on these results, this section concludes by sketching an agenda for future research on optimal statistical reporting when taking seriously the psychological and sociological constraints on social learning based on published research.

Identification of and correction for publication bias (with Isaiah Andrews, revise and resubmit, AER)

Despite following the same protocols, replications of published experiments frequently find effects of smaller magnitude or opposite sign than those in the initial studies (Open Science Collaboration, 2015; Camerer et al., 2016). One leading explanation for replication failure is publication bias, Ioannidis (2005). Journal editors and referees may be more likely to publish results that are statistically significant, that confirm some prior belief or, conversely, that are surprising. Researchers in turn face strong incentives to select which findings to write up and submit to journals, based on the likelihood of ultimate publication.

How can we find out how selective the publication process is? And once we have an answer, how can we use it in interpreting published findings? We consider two approaches to identification of selectivity in the publication process, where by selectivity we mean the dependence of publication probabilities on the findings of a study.

The first approach uses data from systematic replications of a collection of original studies. By replication we mean a study which applies the same same experimental protocol to a new sample. If the original and replication have the same sample size, absent selectivity the joint distribution of initial and replication estimates is symmetric. Asymmetries in this joint distribution nonparametrically identify conditional publication probabilities, assuming the latter depend only on the initial estimate. This result generalizes to the case of differing sample sizes.

Our second approach uses data from meta-studies, by which we mean studies that collect estimates and standard errors from multiple (published) studies. Under an independence assumption common in the meta-studies literature, absent selectivity the distribution of estimates for high variance studies is a noisier version of the distribution for low variance studies. Deviations from this prediction again identify conditional publication probabilities.[4]

When conditional publication probabilities are known (up to scale), we propose median unbiased estimators and valid confidence sets for scalar parameters. These results allow valid inference on the parameters of each study.

We then apply the proposed procedures to three empirical literatures. Our first two applications use data from systematic replication studies in experimental economics and in psychology. Estimates based on our replication approach suggest that results significant at the 5% level are over 30 times more likely to be published than are insignificant results, providing strong evidence of selectivity. Estimation based on our meta-study approach, which uses only the originally published results, yields the same conclusions. Our third application considers the literature on the impact of minimum wages on employment, where no replication estimates are available. Our estimates suggest that results corresponding to a negative significant effect of the minimum wage on employment are about 3 times more likely to be published than are insignificant results. Our point estimates suggest that positive and significant effects might be less likely to be published than negative and significant effects.

Which findings should be published? (with Alexander Frankel, submitted)

The publication process thus appears to be quite selective, resulting in important distortions to statistical inference and in lack of replicability of published findings. In response to these concerns, there have been calls for reforms in the direction of non-selective publication; see for instance Christensen and Miguel (2016). One proposal is to promote statistical practices that de-emphasize statistical significance, for instance by banning "stars" in regression tables. Another proposal is for journals to adopt Registered Reports, in which pre-registered analysis plans are reviewed and accepted prior to data collection.

[4]A web app implementing this approach is available at `https://github.com/maxkasy/MetaStudiesApp`.

Are these recommendations warranted? In our paper we argue that the answer depends on what we consider to be the objective of the research and publication process. If the objective is relevance for policy decisions, in particular, than these recommendations are not optimal.

In this paper, contributing to the economic theory literature on the value of information, we seek the optimal rule for determining whether a study should be published, given both its design and its findings. Our analysis is from an instrumental perspective: the value of a study is that it informs the public about some policy-relevant state of the world before the public chooses a policy action. The optimal publication rule defined in this manner selects on a study's findings. To understand why, note that there is no instrumental value from publishing a study with a "null result" that doesn't move the policy away from the default action. The same policy would have been chosen even if the study weren't published, so publishing would incur a cost without a benefit. The studies that are worth publishing are the ones that show that there is some payoff gain from taking an action other than the default.

After characterizing optimal publication rules, we return to the distortions caused by selective publication. It is immediate that common forms of inference are valid if the publication probability does not depend on the point estimate given the standard error. We show that the inverse is also true: Under any policy in which the publication probability depends on the point estimate, common forms of frequentist inference will be invalid conditional on publication. Point estimates are no longer unbiased, for instance, and uncorrected likelihood-based inference will not be accurate. Putting these results together, we see that selectively publishing extreme results is better for policy-relevance but leads to distorted inference. To the extent that the current (selective) publication regime qualitatively resembles the optimal rules we derive, then, a move towards non-selective publication in order to improve statistical credibility might have costs as well as benefits.

An abstraction in the model described above is that it considers a "static" environment with a single paper to be published and a single action to be taken. We next consider a dynamic extension to our model. Just as before, we find a benefit of publishing extreme results. But we also find a benefit of publishing precise results – even precise null results that don't change the current action. Publishing a precise result today helps avoid future mistakes arising from the noise in the information that has yet to arrive.

Limited attention and statistical reporting as constrained communication (work in progress)

A key puzzle, when thinking about which findings should be published, and more generally which statistics should be reported when analyzing data, is why there should be selection at all. From a (naive) Bayesian perspective, it would be optimal to simply communicate all available data to readers of scientific studies, rather than reporting selected statistics and tests. It would furthermore be optimal to communicate the results of all studies to policymakers.

Such an approach is evidently far from scientific practice, and is not practically feasible. A possible justification for the reporting of selected statistics and tests is limited bandwidth in the processing of information, an important topic in the psychology and economics literature on limited attention. I plan to consider optimal procedures for statistical reporting subject to constraints on the processing of information identified by the literature on limited attention.

I am currently organizing a conference, to take place at Harvard in May 2019, which will further explore these issues.Invited participants are from the fields of statistics, economic theory, psychology, philosophy, and history of science. The hope is to spark a productive exchange on the foundations of statistical practice.

Agenda 4: Identification problems motivated by applied work on economic inequality

Lastly, let me discuss some research on questions of identification motivated in particular by empirical work on urban segregation, inequality and distributional policy evaluation. In light of the larger number of my papers in this area, I will just provide brief summaries here.

Identification problems

In **Partial identification, distributional preferences, and the welfare ranking of policies** (Review of Economics and Statistics, 2016), I consider the problem of ranking and choosing policies when the policy-relevant parameters are only partially identified, for instance due to less-than-perfect compliance or attrition in an experiment. In many cases the data are informative for policy choices, even if treatment effects or similar objects can only be bounded. In **Identification in Triangular Systems Using Control Functions** (Econometric Theory, 2011) I derive necessary and sufficient conditions for validity of the control function approach that is popular in applied work using instrumental variables. As it turns out, the control function approach yields valid estimates only in the (unlikely) case that unobserved heterogeneity is one-dimensional. In **A nonparametric test for path dependence in discrete panel data** (Economics Letters, 2011), where I explored observable implications of duration dependence of transition probabilities (for instance between employment and unemployment) in the presence of arbitrary heterogeneity. Absent duration dependence, the distribution of individual trajectories would follow a mixture of Markov chains, which can be characterized using an extension of de Finetti's theorem. I propose a test based on this characterization. The paper **Uniformity and the delta method** (forthcoming, Journal of Econometric Methods, 2018) explores the underlying reasons behind inference difficulties in various settings with weak or partial identification. As argued in this paper, in almost all cases these issues are due to a failure of the delta-method to deliver accurate approximations.

Urban segregation

My dissertation was motivated by questions of urban segregation and its causes. In **Nonparametric inference on the number of equilibria** (The Econometrics Journal, 2015), I developed tests for the hypothesis of "neighborhood tipping" advanced in the urban economics literature. The key technical difficulty here is inference on the number of roots of functions that are nonparametrically estimated. In a related spirit, **Identification in a model of sorting with social externalities and the causes of urban segregation** (Journal of Urban Economics, 2015) considered models of spatial equilibrium to develop approaches that allow to distinguish between alternative causes of urban segregation: Spatial spillovers or homophily on the one hand, unobserved exogenous heterogeneity of locations on the other hand.

Inequality and distributional policy evaluation

Related concerns motivated several papers on economic inequality and distributional decompositions. In these papers, we study the entire distribution of outcomes, rather than just conditional means or some other summary statistic. In **The impact of changing family structures on the income distribution among Costa Rican women 1993-2009** (joint with Alvaro Ramos-Chaves, Feminist Economics, 2014) we study the effect that mandatory DNA-testing for presumptive fathers of out-of-wedlock in children had on the distribution of women's economic resources. We consider the effect on the distribution of women's equivalent household incomes (i.e., individual incomes imputed based on household income and composi-

tion), and find a reduction of women's poverty due to reduced birth rates, increased marriage rates, and increased transfer payments. Using similar methods, in **Survey mode effects on income inequality measurement** (joint with Pirmin Fessler and Peter Lindner, forthcoming, Journal of Economic Inequality, 2018), we find that survey methods have quantitatively important causal effects on the distribution of measured incomes, with telephone-based modes leading to underestimation of the number of households in the upper and lower tail of the distribution. This finding calls into question comparisons of inequality based on surveys using different methodologies. In **Who wins, who loses? Identification of the welfare impact of changing wages** (submitted, 2018) I extend the methodology of distributional decompositions to explicitly consider welfare effects (rather than effects on observable outcomes). This is technically challenging since it requires the identification of causal effects conditional on vectors of endogenous outcomes. In an application to the expansion of the Earned Income Tax Credit in the 1990s, I find that a sizable share of the incidence of this expansion went to employers.

Intersection of domains

The research described in this statement draws on multiple domains. My future research will continue to do so. To conclude, let me briefly describe why I believe that these domains can be helpful for developing the methodology of empirical economics. There is, of course, mathematical statistics, which often takes center-stage in econometrics.

These domains further include actual empirical research, with the goal to develop methodological tools in tandem with the questions and challenges faced by empirical researchers, as well as the data and computational resources available to them. As empirical research develops, the methodological toolkit needs to follow suit in order to remain helpful.

These domains also include economic theory, applied both to the objects of empirical research, and to the activity of doing empirical research itself. Empirical research is produced and selectively communicated by actors with potentially divergent objectives and information; economic theory helps to understand this process.

Next, I draw on the neighboring domains of biostatistics and of machine learning, which have many insights to contribute to econometrics. The booming field of machine learning is driven by the fast growing availability of "big data," characterized by a large number of observations, a very rich set of available variables or "features," and non-traditional forms of data such as text, networks, or images. This has prompted innovations in areas such as regularization and tuning for prediction using highly non-linear models, as well as adaptive experimentation for online learning. Econometrics can profit by judiciously adapting these methods, and can contribute to an understanding of their benefits and shortcomings.

Lastly, there are conceptual and foundational considerations as discussed in the philosophy and history of science. For example, to what extent is our notion of causality tied to the notion of randomization, as opposed to the notion of controlled (exogenous) interventions? Or, to what extent does economics have theories with similar claims to universal truth as the "hard" sciences, and what does that imply for notions such as the testing of theories?

References

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Chetty, R. (2009). Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Annual Review of Economics*, 1(1):451–488.

Christensen, G. S. and Miguel, E. (2016). Transparency, reproducibility, and the credibility of economics research. NBER Working Paper No. 22989.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2018). Return of the infinitesimal jackknife. *arXiv preprint arXiv:1806.00550*.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8).

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):pp. 47–55.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, pages 1–20.

Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018). A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.