Policy Prediction: The Missing Tool in Experimental Econometrics and a Roadmap to Fix It

Nandan Rao

September 9, 2019

Abstract

Applied economics research is most often "applied" to policy making. In the literature on experimental and quasi-experimental econometrics, however, there are few formal frameworks that use the results of experimental studies to predict the effects of policies in new contexts. I explain the existence of this gap from a historical perspective, review the current methods that do exist, and set a research agenda to fill it using recent ideas from econometrandics and machine learning.

1 Introduction

Randomized control trials (RCTs), or natural experiments that replicate them, have become the official gold standard of empirical work in economics and many related fields. More and more, policy makers are encouraged to look to RCTs to make "evidence-based" policy decisions (Charles F. Manski 2012; Cartwright and Hardie 2013).

Many prominent economists have expressed a concern that RCTs, and the quasi-experimental methods that seek to replicate them (i.e. natural experiments, intrumental variables, regression discontinuity, etc.) are particularly difficult to generalize to new contexts due to their overarching concern for internal validity, often at the expense of external validity (Heckman and Smith 1995; Heckman 2008; Deaton 2010; Charles F. Manski 2012; Deaton and Cartwright 2018).

Making evidence-based policy decisions is an act of generalizing from previous studies to decisions about the future. Charles Manski (2012) voices

the concern that experimental studies have tended to "be silent" on the question of external validity and that "from the perspective of policy choice… What matters is the informativeness of a study for policy making, which depends jointly on internal and external validity."

This silence on external validity means that, despite the huge field of research and techniques for ensuring causal identification (internal validity) via experimental or quasi-experimental methods, very little has been done to either A) create tools to prove that the same results will apply generally in all contexts or B) create tools to predict the results in a specific context, given proven results from one or more experiments.

Abhijit Banjerjee himself (A. V. Banerjee, Chassang, and Snowberg 2016), a large proponent of RCTs, has recently echoed the current lack of and need for more formal systems for the generalization of experimental studies, stating "it is our belief that creating a rigorous framework for external validity is an important step in completing an ecosystem for social science field experiments, and a complement to many other aspects of experimentation."

In response to such theoretical concerns, a small but growing literature has sprung up around empirically proving that results from prominent RCT or other causally identified studies do not easily extrapolate to new contexts (Pritchett and Sandefur 2015, Allcott 2015, Bisbee et al. 2017, Rosenzweig and Udry 2019). These results emphasize a need for powerful extrapolation tools if policy makers are realistically expected to make decisions based on "evidence."

Predicting the results of a policy in a new location is, tautologically, a prediction problem. Machine learning is a field which has been very successful at formalizing prediction problems. Domain adaptation, a subfield of machine learning, formalizes the problem of moving from one (or more) domains with labelled data to a new domain where only unlaballed data is available (for a survey, see Pan and Yang 2010).

Formulating the problem of policy prediction in terms of domain adaptation gives us a rigorous way to think about the assumptions that may or may not hold, as well as a rich, if short, history of techniques that have been used to solve the problems associated with the assumptions. I will show how policy prediction can be formulated as a covariate shift problem and propose that the associated assumptions could be tested empirically in a treatment effects framework.

This article will take the following form: Section 2 will define external and internal validity, their role in scientific inference, review the development of the Fisherian experimental model of statistical inference, and set out the need for a formal framework of extrapolation. Section 3 will

argue that a target context is necessary by reviewing the development of such an argument in mid-century econometrics and formalizing their arguments in probabilistic terms. Section 4 will introduce the concept of domain adaptation in machine learning, review econometric formulations of the same problem, and present recent work in machine learning and econometrics that can be used to solve the problem implied by the formulations. Section 5 will conclude.

2 Validity and Counterfactual Identification

2.1 A Taxonomy of Validities

Critiques against the holy status of RCTs have focused on their preference for internal validity and tendency to completely ignore external validity. It is worth defining these terms and reflecting on exactly what it is that experimental methods provide us. The most agreed upon definition of these terms comes from William R. Shadish, T. D. Cook, and D. T. Campbell 2001 (an update of their previously popular taxonomy of validities from T. Cook and D. Campbell 1979):

> **Statistical Conclusion Validity:** The validity of inferences about the correlation (covariation) between treatment and outcome.
>
> **Internal Validity:** The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.
>
> **Construct Validity:** The validity of inferences about the higher order constructs that represent sampling particulars.
>
> **External Validity:** The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

Many authors use the term external validity to refer to what Shadish, Cook, and Campbell separate into construct and external validity. As both are involved in generalization, and both are external to the particulars of the sample, I will follow that abuse of terminology and refer to all the challenges of both as "external validity."

What, then, is meant by the term "inference?"

2.2 Inference and Statistics

Mary S. Morgan, in her *History of Econometric Ideas* (1991), lays out the tension between mathematics and statistics in economics of the nineteenth-century: mathematics was seen as a tool for deductive reasoning in economics, used to derive logical conclusions from known laws. Statistics, on the other hand, was viewed as a tool for inductive reasoning, used to establish economic regularities.

This distinction between deduction and induction follows from the Aristotelian tradition of logic, where inference is the process of reasoning that can be broken into two parts: reasoning from particulars to generals (induction) and reasoning from generals to particulars (deduction). Once one has induced a general law ("all men are mortal"), one can deduce facts that might otherwise not yet be apparent ("Socrates will die"). There is, of course, a natural contradiction in this process: how can one know that all men are mortal, if one did not already know that Socrates will die? In other words, how can one ever claim that "all men are mortal" until one has seen every man die?

This problem forms the foundation of David Hume's *Problem of Induction*:

> "As to past Experience, it can be allowed to give direct and certain information of those precise objects only, and that precise period of time, which fell under its cognizance: but why this experience should be extended to future times, and to other objects, which for aught we know, may be only in appearance similar, this is the main question on which I would insist" (Hume 1748)

There are two distinct problems Hume raises, that of generalizing from one object to another (from seeing some men die to assuming all men have died) and that of generalizing from the past to the future (because all men have died, thus, all men will die). He goes on to explain that this extrapolation is only valid under strong assumptions as to the "uniformity of nature." One must assume nature is uniform in such a way as to enable extrapolation from one object to another or from the past to the future.

R.A. Fisher also framed his statistical techniques in terms of "inductive inference." In the introduction to The Design of Experiments (1935), he explicitly frames his canonical book and its techniques in the terms of logicians such as Hume, arguing for the possibility of induction via statistical methods:

> "...it is possible to draw valid inferences from the results of experimentation...as a statistician would say, from a sample to the population from which the sample was drawn, or, as a logician might put it, from the particular to the general." (Fisher 1935)

Fisherian statistical inference is a tool in the process of induction that seeks to address Hume's problem of extending experience with one object to that of similar objects. In particular, the "similar objects" to which conclusions are extended are not just similar in appearance, but rather have a distinct relationship to the experienced objects: they are the population from which the experienced objects represent a sample.

Fisher's methodologies for significance testing relate to drawing conclusions about populations given a sample. The validity of the use of such techniques in a study falls squarely under the category of "statistically conclusion validity" in the taxonomy of Shadish, Cook, and Campbell.

Fisher's theory of experiments (in particular, the RCT) addresses internal validity and causality. The connection between the RCT and the identification of a causal relationship comes straight from John Stuart Mill's (1884) *Method of Difference* for causal discovery:

> "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon."

Fisher (1935) had this in mind when he defended randomization, claiming that holding all other variables constant was not feasible, and thus, holding them to the same distribution, by making the assignment of treatment independent of those variables, was desirable (Rosenbaum 2005). It should be noted, however, that for the subjects of interest to Fisher, the difference between something being a "cause" or "a necessary part of the cause" was not especially important. What Fisher was essentially interested in was interventional prediction: what is the effect when one makes individual changes to the growing conditions of these plants.

Holland (1986) begins his landmark paper, before formulating the Rubin-Neyman causal model, by framing his goal:

> "Others are concerned with deducing the causes of a given effect. Still others are interested in understanding the details of

> causal mechanisms. The emphasis here will be on measuring the effects of causes because this seems to be a place where statistics, which is concerned with measurement, has contributions to make."

The success of Fisher's framework of randomization and the Rubin-Neyman causal model comes down to this razor focus in purpose: they make no claims to discover all the causes of a given effect, to discover the mechanism of the cause, or even to separate between a cause or a necessary part of a cause. They allow us to reason about counterfactuals: what would have happened, on average and in the past, had we treated our entire population rather than a randomized part of a randomized sample drawn from that population. It operationalizes Mill's method of differences, creating a pathway to internal validity that is achievable in the real world.

The counterfactual causal model, however, provides no framework for generalizing from a specific population to a more general one or from the past to the future (Heckman 2008). The "effects of causes" is a black-box methodology for causal identification (Heckman and Smith 1995). It does not require one to answer the question "why" and it is therefore inherently context specific: it asks, "what happened when I did X in context D?" In Fisher's line of work (agriculture), none of those shortcomings were problematic. He was able to randomly sample from the exact population (seeds of grain) to which his inferences needed to generalize.

The success of Fisher's randomization in his field of agriculture, and the subsequent success of the Rubin-Neyman causal model in epidemiology, is in no way incidental. These are fields that deal with encapsulated biological units where agreed-upon scientific theory tells us that the relationships in these systems will be invariant to a wide degree of changes that happen in the world from one year to the next or from one country to the next.

For example, it is implicitly assumed that the growth of corn will not be affected by its expectations of how it will be cut down and processed during the harvest. Corn grows according to the properties of the soil and the sun it receives. Its growth will be independent of its destination in corn flakes or arepas, conditional on those properties. Opioids will inhibit pain in humans, regardless of their political ideology, faith in their governmental institutions, or their love of Shark Tank.

These arguments are not made explicitly, but their validity is absolutely necessary to enable the application of their findings: they create laws defined in relation to the entire range of contextual changes that one might

want for prediction- and decision-making and that allows the laws to be applied deductively in a wide array of useful situations.

While the validity of these arguments is taken for granted in contained biological processes, this is simply not the case in the social sciences. This is why Shadish, Cook, and Campbell lay out 19 different threats to construct and external validity of social science experiments. In the case of growing corn, the threats either simply do not apply or are implicitly neutralized through basic scientific understanding of plant growth.

In the case of economics, the outcomes are regularly based on individuals decisions to consume, work, study, invest, or move. The treatments are regularly subjected to a gauntlet of mediating factors and interacting variables that are highly context-dependent and correlated across individuals in a single place and time. The population for which one wants to draw actionable inference is in the future and it is fundamentally different from the population the sample from drawn from. Internal validity of a study in economics is not sufficient for actionable inference to take place in a new context.

2.3 The Danger of Informal Inference

It might be argued that it is, and should be, up to the sound judgement and expert opinion of the policy maker to determine if a given counterfactual analysis should extrapolate to their context or not. Empirical studies only need to provide internal validity, according to this argument, as the extrapolation is done by experts who know their target domain.

While incorporating expert domain knowledge can only help predictions, we can differentiation between using a formal statistical framework and using an informal framework of judgement. John Stuart Mill (1884), reflects on these difference and the dangers of inferring without formal frameworks.

He begins by denying that the only process of inference consists of separately applying induction (particulars to generals) and then deduction (general law to particular context):

> "All inference is from particulars to particulars: General propositions are merely registers of such inferences already made, and short formulae for making more: The major premise of a syllogism, consequently, is a formula of this description: and the conclusion is not an inference drawn from the formula, but an inference drawn according to the formula: the real logical

> antecedent, or premise, being the particular facts from which the general proposition was collected by induction." (Mill 1884)

In other words, in the process of creating the general law, one has created a series of particular laws, and only once assured that all the particular laws are valid can one be assured of the more general law. He goes on to caution against the direct reasoning of particulars to particulars because it is informal and we are likely to bring our own biases into the process and make mistakes:

> "In reasoning from a course of individual observations to some new and unobserved case, which we are but imperfectly acquainted with (or we should not be inquiring into it), and in which, since we are inquiring into it, we probably feel a peculiar interest; there is very little to prevent us from giving way to negligence, or to any bias which may affect our wishes or our imagination, and, under that influence, accepting insufficient evidence as sufficient." (Mill 1884)

John Stuart Mill argues that formal procedures, such as that implied by the framework of induction and deduction, allow individuals to avoid these biases. In his world, there was no formal procedure for reasoning from particulars to particulars, but there was for reasoning from particulars to general. As such, he recommends the latter as a way to avoid biases of "wishes" and "imagination."

In an ideal world, the research community has discovered a set of general laws that are invariant to all contexts and the policy maker can apply them, via the process of deduction, to get the desired result in their circumstance. But what if that general law has not yet been discovered? Then the policy maker must look at individual studies (particulars) and attempt to infer a prediction for the result of a similar policy in their context. This is exactly the situation that John Stuart Mill has warned is fertile ground for bias and imagination.

3 The Origins of Structure

Economists during the same period as Fisher took a different approach to conceptualizing and thinking about the inference they were doing. Ragnar Frisch, theorizing about macro-dynamic analysis, set out several key ideas relating to the structure of a system and the autonomy of a structure

(Frisch 1995). For Frisch, the "structure" of a system was all the characteristics of the phenomena that could be quantitatively described. In his macrodynamic systems, the structure is defined by a set of functional (simultaneous) equations. He then poses the question: what would happen to the system due to an arbitrary change in a single variable? To do so could imply a different "structure" than the one which the equations describe, requiring a different set of equations altogether to describe the new system.

> "But when we start speaking of the possibility of a structure different from what it actually is, we have introduced a fundamentally new idea. The big question will now be in what directions should we conceive of a possibility of changing the structure?... To get a real answer we must introduce some fundamentally new information. We do this by investigating what features of our structure are in fact the most autonomous in the sense that they could be maintained unaltered while other features of the structure were changed... So we are led to constructing a sort of super-structure, which helps us to pick out those particular equations in the main structure to which we can attribute a high degree of autonomy in the above sense. The higher this degree of autonomy, the more fundamental is the equation, the deeper is the insight which it gives us into the way in which the system functions, in short, the nearer it comes to being a real explanation. Such relations form the essence of 'theory'."

This concept, that of "the essence of theory" being the discovery of some relationship that is autonomous and invariant to a great degree of changes we can imagine performing to a system, is taken up by Trygve Haavelmo (1944), who writes that:

> "The principal task of economic theory is to establish such relations as might be expected to possess as high a degree of autonomy as possible."

He then goes on to consider a distinction between the "invariance" of a relationship under hypothetical changes in structure versus the "persistence" of a relationship under observed changes in structure:

> "...if we always try to form such relations as are autonomous with respect to those changes that are in fact *most likely to occur*, and if we succeed in doing so, then, of course, there will be a

> very close connection between actual persistence and theoretical degree of autonomy."

This implies a connection between autonomy and the *type* of changes to which it is invariant *with respect to*. This point is made even more explicitly by Leonid Hurwicz (1966). Similar to his predecessors, his model consists of a system of equations that constrain the state of the world, given a history of states. He calls this system of equations a "behavior pattern." He states that:

> "A great deal of effort is devoted in econometrics and elsewhere to attempts at finding the behavior pattern of an observed configuration... But do we really need the knowledge of the behavior pattern of the configuration?... It will be approached here from the viewpoint of prediction... That is, the word 'need' in the above question will be understood as 'need' for purposes of prediction."

He then goes on to define what he calls a "structural form" as one which is identified and identical across all possible behavior changes that one *needs* to predict within. He stresses that:

> "The most important point is that the concept of structure is relative to the domain of modifications anticipated."

Thus, there is an inherent and irrevocable connection between what we consider a "law" and the degree of changes we require the law to persist across. The law is defined in relation to those changes. Folowing Hume, the performance of induction is always connected to a specific assumption about the uniformity of nature. Additionally, the exact way in which nature must be uniform, for a particular system under study, is defined by the predictions we need to make with the discovered relations in that system.

3.1 Invariant Conditionals

Robert Engle, building on the work of Frisch and Hurwicz regarding invariant/autonomous structures, creates a statistical definition of what he terms "super exogoneity" (Engle, Hendry, and Richard 1983).A good review of this historical evolution of autonomy can be found in Aldrich 1989.

Super exogoneity is defined by Engle as follows. Consider a model in which the "structure" (the functional form) of a relationship between outcome y and variable z is parameterized by "structural" parameters $\lambda_1, \lambda_2 \in \lambda$. If the joint density implied by the model can be factorized as follows:

$$P(y, z, \lambda) = P(y|z, \lambda_1)P(z|\lambda_2)$$

and λ_2 is independent of, hence contains no information about, the structural parameter of interest λ_1, then the variable z is said to be weakly exogenous to λ_1. If, additionally, the conditional distribution, $P(y|z, \lambda_1)$ remains invariant to changes in the marginal $p(z)$ (either caused by changes in it's generating process, parameterized by λ_2, or through other interventions that modify the values themselves), then z is super exogenous.

This definition allows super exogoneity to be refuted by data:

> "It is clear that any assertion concerning super exogeneity is refutable in the data for past changes in $D(z, |X_{t-1}, \lambda_2)$ by examining the behavior of the conditional model for invariance when the parameters of the exogenous process changed...However, super exogeneity for all changes in the distribution of z, must remain a conjecture until refuted, both because nothing precludes agents from simply changing their behavior at a certain instant and because only a limited range of interventions will have occurred in any given sample period."

This clarifies the difference between weak and super exogoneity: one is observational and the other interventional. Weak exogoneity is an observational characteristic, one must only show that structural parameters λ_2 provide no information for estimating λ_1. Super exogoneity implies invariance across any possible intervention which changes the distribution of z.

Writing a system in terms of super exogenous variables is akin to finding Frisch's "super-structure." This is the part of the system that stays invariant to a set of allowable modifications: changes in $P(z)$. This is the part of the system that must be known in order to make policy predictions, when λ_2 includes the changes in the policy, whose total causal effect (in the sense of Pearl 2000) is transmitted to the output variable y through z. The existence of such an invariant super-structure is a necessary prerequisite to successfully predict the effects of policy and therefore to successfully inform policy choice from empirical data.

This method of looking in the data for a conditional distribution that is invariant to "structural" changes that lead to differences in the marginal

distribution of its "inputs" has formed the basis of new line of work in statistics and machine learning around causal discovery (Peters, Bühlmann, and Meinshausen 2015; Heinze-Deml and Meinshausen 2017; Rojas-Carulla et al. 2018). This line of research, in the terms of Engle, can be thought of as methods for model selection by selecting covariates that are super exogenous to the output in question. I will return to some of these methods in section 4.

Additionally, the connection between weak and super exogoneity is exploited by K. Zhang, J. Zhang, and Schölkopf 2015, where the lack of weak exogoneity is used to determine the causal direction in the case of two correlated, un-confounded variables. Zhang's method points to the possibility of tests that exclude the assumptions of covariate shift with data from only a single domain. In other words, a general testing framework that works without testing directly for super exogoneity, which in theory requires proving the invariance of the conditional distribution $P(Y|Z)$ against all arbitrary manipulations of the marginal $P(Z)$.

4 Domain Adaptation

4.1 Definition and Formulation

Consider a domain, $\mathcal{D}$, which we define as consisting of a feature space, $\mathcal{X}$, and a marginal distribution $P(X)$ where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. A task, $\mathcal{T}$, consists of an outcome space, $\mathcal{Y}$ and a true generating mechanism $f : \mathcal{X} \to \mathcal{Y}$.

Many techniques of traditional statistics and machine learning make the assumption that the domain and the task are the same in the training data and the prediction context. Transfer learning is the generic name for frameworks that work outside these assumptions, when either the domain, the task, or both are allowed to change.

I will consider the term domain adaptation in the sense of Ben-David (2006) and Pan and Yang (2010). Under this definition, domain adaptation is a specific form of transfer learning where the task $\mathcal{T}$ is constant and the feature space, $\mathcal{X}$ is the same across all domains. There is a target domain, $\mathcal{D}_T$ from which one has samples $\{x_1, \ldots, x_i\} \in X$ and (one or more) source domain(s), $\mathcal{D}_S$, from which one has tuples $\{(y_1, x_1), \ldots, (y_n, x_n)\} \in (\mathcal{Y}, \mathcal{X})$.

The consistency of the task relates to the construct validity in the taxonomy of William R. Shadish, T. D. Cook, and D. T. Campbell 2001. If the outcome measured in one experiment might be considered to measure a different construct than the outcome measured in another experiment,

then the task can not be said to be the same and this problem formulation fails.

Allow $\mathcal{X} = \{\mathcal{W}, \mathcal{Z}\}$, where $\mathcal{W} = \{W_0, W_1\}$ a binary treatment and $\mathcal{Z} = Z_1, \ldots, Z_P$ a set of covariates. The assumption of the consistency of the feature space relies on the construct validity of both the treatment and the covariates. If this validity does not hold, if the covariates or the treatment across domains represent a different construct, then this formulation is not valid.

If one considers a random-variable formulation of the true generating mechanism, $f(\cdot)$, then the assumption of task consistency implies the conditional distribution is consistent across tasks, thus $P_S(Y|X) = P_T(Y|X)$. This assumption is referred to as "covariate shift", as the only difference between domains $\mathcal{D}_S$ and $\mathcal{D}_T$ is that of the marginal covariate distribution, $P_S(X) \neq P_T(X)$.

Under the assumptions of covariate shift, Shimodaira (2000) shows that the optimum likelihood function for a maximum likelihood estimator of $P_T(Y|X)$, given sufficiently large sample size, is obtained by minimizing the weighted loss function trained on labeled data from the source domain and weighted by the ratio $w(x) = \frac{P_T(x)}{P_S(x)}$:

$$\underset{\beta}{\operatorname{argmin}} \sum_i -w(x_i) \log P_S(y_i|x_i, \beta)$$

The use of this weighting ratio, $w(x)$ (referred to as the *importance*) has led to many other importance estimation techniques to deal with covariate shift (Suigyama et al. 2007; Pan and Yang 2010).

For the rest of this article, I will focus on the goal of formulating policy prediction problems such that the covariate shift assumption. If that assumption holds, the task of policy prediction in a new context can be solved with a relatively straight-forward application of importance estimation techniques.

4.2 Domain Adaptation in Econometrics

The problem of covariate shift has a related, long history in the problem of overcoming sample selection bias in econometrics (C. F. Manski and Lerman 1977). While the fundamental statistical problem is the same, the applications are specific.

The most obvious economic approach which addresses the problem of external validity is that of structural economics. Heckman (2008) explains that RCTs solve internal validity, while the problem that "economic policy

analysts have to solve daily" involves more, and not just the generalization of previous experiments but also the prediction of the effects of policies "never historically experienced," which he argues only structural models based on behavioral choices are capable of doing.

The tension between structural and experimental economics is long standing and I will not succeed in resolving it here. This article seeks to lay out a research agenda which is firmly planted within the field of experimental econometrics, however, I do wish to acknowledge that structural econometrics does indeed solve the same problem with a different set of assumptions. While I do not believe that can be easily brushed off, I also do not have the space to address it here.

Hotz et al (2005) provide one of the only canonical models in the experimental econometrics literature, that I am aware of, for extrapolating from a source context with experimental data and measuring predictions in a target context where such data is not available.

Their technical methodology follows Abadie and G. W. Imbens 2006, using matching with replacement from the target to the source domain to create a bootstrapped dataset on which to run a regression. This can be seen as a reinvention of importance estimation (Shimodaira 2000; Suigyama et al. 2007). The implication of this choice of techniques, as shown previously, is the assumption of covariate shift. The failure of their technique to work in certain contexts implies a violation of the covariate shift assumptions. They try adding additional personal covariates to their model and find that doing so has little effect on their predictive ability. Their models are linear and additive and they deal with treatment heterogeneity by splitting the sample into two groups, according to a variable previously identified, and running all analyses separately. They predict the outcomes of control and treatment groups separately and introduce t-tests to compare the distributions of the predicted outcomes with that of the actual outcomes. This leads to a potential test of transportability by looking at the t-test of the control group outcomes.

Gechter 2015 builds on the technique of Hotz, but rather than using a statistical test to determine if the model is transportable or not, he creates bounds for prediction in a new domain, using the difference between outcomes for the control groups $P_S(Y|W_0, Z)$ and $P_T(W_0, Z)$ to create the bounds.

I propose that these methodologies can be improved in the following areas:

1. A theoretical justification for the technique in terms of the assumptions involved, related to observable and unobservable variables, that

may or may not fail in practice.

2. A formal framework for model selection that determines which covariates to condition on and which to marginalize out to enable a transportable prediction.

3. A focus on outcomes rather than treatment effects.

The following sections will be devoted to showing why these features matter and how they can be addressed with recent ideas from both machine learning and econometrics.

4.3 Exogoneity and Covariate Shift

Pearl and Bareinboim 2014 offers a full set of conditions in which conditional distributions are transportable. As their setup relies on the do-calculus of Pearl 2000 and it is more than we need for our illustrative purposes here, I leave the reader to follow the references if they are of interest. I will, however, provide a small set of propositions to connect exogoneity and the concept of covariate shift, whose connection to their derived conditions will be obvious to anyone familiar.

The following propositions will help us understand, in terms of exogoneity, different ways by which the covariate shift assumption can fail.

Consider a source domain, $\mathcal{D}_S$ and a target domain $\mathcal{D}_T$, with feature space defined as $\mathcal{X} = \mathcal{Z} \cup \mathcal{H}$. $Z \in \mathcal{Z}$ is a set of observable covariates, where $P_S(Z) \neq P_T(Z)$, and $H \in \mathcal{H}$ a set of unobservable covariates.

Proposition 1. *The covariate shift assumption is violated if the variable set Z is not super exogenous.*

Proof. This follows directly from the definition of super exogoneity: a failure of super exogoneity implies that the conditional distribution $P(Y|Z)$ is not invariant to changes in $P(Z)$, which implies that $P_T(Y|Z) \neq P_S(Y|Z)$ given the assumption that $P_S(Z) \neq P_T(Z)$. □

Proposition 2. *The covariate shift assumption holds for $P(Y|Z)$ if the variable set Z is super exogenous and $P_S(Y|Z, H) = P_S(Y|Z)$.*

Proof. As Z and H encompass all variables and are thus the only possible changes across domains, this follows directly from super exogoneity of Z, $P_S(Y|Z, H) = P_S(Y|Z) = P_T(Y|Z)$. □

Proposition 3. *The covariate shift assumption holds for $P(Y|Z)$ if the variable set $\{Z \cup H\}$ is super exogenous and $P_S(H|Z) = P_T(H|Z)$.*

Proof. Super exogoneity of $\{H \cup Z\}$, given they encompass all variables, implies $P_S(Y|Z,H) = P_T(Y|Z,H)$. Then $P_S(Y|Z) = \int P_S(Y|Z,H)P_S(H|Z)dH = \int P_T(Y|Z,H)P_T(H|Z)dH = P_T(Y|Z)$. □

Proposition 4. *The covariate shift assumption holds for $P(Y|Z)$ if the variable set $\{Z \cup H\}$ is super exogenous and the unobservable variables are independent of the observables: $P_S(H|Z) = P_S(H) = P_T(H)$.*

Proof. Super exogoneity of $\{H \cup Z\}$, given they encompass all variables, implies $P_S(Y|Z,H) = P_T(Y|Z,H)$. Then $P_S(Y|Z) = \int P_S(Y|Z,H)P_S(H)dH = \int P_T(Y|Z,H)P_T(H)dH = P_T(Y|Z)$. □

I will provide a concrete example to underscore the relevance of these propositions.

4.4 An Example Application: Microcredit

As a motivating example, I will refer to a set of recent microcredit RCTs that have been used as a dataset to examine external validity of individual studies (Pritchett and Sandefur 2015; Meager 2018). The studies were all published in 2015, all carried out all within a few years of each other, and all in different locations: Mexico, Mongolia, India, Bosnia and Herzegovina, Ethiopia, Morocco (Attanasio et al. 2015; Angelucci, Karlan, and Zinman 2015; Augsburg et al. 2015; A. Banerjee et al. 2015; Crépon et al. 2015; Tarozzi, Desai, and Johnson 2015).

All the studies involved some randomization through which individuals were randomly given "greater access" to microcredit lending, with the exact way in which greater access is defined and the exact terms of the microcredit lending differing across sites. The Bosnian study randomizes at individual level, taking those who were just below the cutoff to being accepted and offering them loans, while the other studies randomize at regional levels, marketing or offering their products in new regions. In line with the frameworks outlined above, before applying any techniques of knowledge transfer between these domains, we must ask ourselves: is the feature space the same? Do these treatments represent the same construct or not? In each case, various output variables were recorded, including household consumption in the months following the intervention and business profits, for which it is easier to argue that they measure the same construct.

Let us assume, for the sake of using this as an example, that the feature space is the same and that variables all measure the same constructs. It is easy to imagine unobservable variables that might effect the profit of the

business and additionally might interact with the treatment of access to microcredit. One such unobservable was hypothesized by Abhijit Banerjee (2011): he proposes that many of the individuals in these studies who have their own business do not actually romanticize entrepreneurship in the way that economists do. Indeed, many would prefer to have salaried jobs than to run their own small business. Given credit, they may not invest very vigorously in their business because they do not actually believe it can grow, nor do they spend any time imagining it growing to be anything more than it is.

Let us call such an unobservable "entrepreneurialness." As it is unobservable, we will refer to it with the variable H. This characteristic, however, certainly effects other characteristics that we do observe (Z) and might try to include in our predictive model $P(Y|Z)$. For example, the number of past loans or number of previous businesses.

Let $Z = \{W, Z_2\}$ where $W \in \{W_0, W_1\}$ will represent the treatment (some increased availability of microcredit to the individual) and Z_2 will consists of the observed variables that proxy entrepreneurialness but do not effect profits in any direct way, thus $P(Y|H, W, Z_2) = P(Y|H, W)$. We will also assume that the treatment, W, is randomized (with full participation for the sake of simplicity) such that $P(H|W, Z_2) = P(H|Z_2)$.

It should be clear, based on our propositions, that the covariate shift assumptions can fail based on two conditions that may or may not hold between the source and the target domain:

1. $P_S(H) = P_T(H)$
2. $P_S(H|Z_2) = P_T(H|Z_2)$

We take each possibility in turn (the technical conditions are taken directly from Pearl and Bareinboim 2014 and adapted to this scenario):

$P_S(H) = P_T(H)$ **and** $P_S(H|Z_2) \neq P_T(H|Z_2)$

In this case, $P_S(Y|W) = \int P_S(Y|W, H, Z_2)P_S(H)dH = P_T(Y|W)$, thus the invariant condition is given by $P(Y|W)$. In contrast, $P(Y|W, Z_2)$ will not be invariant: $P_S(Y|W, Z_2) = \int P_S(Y|W, H, Z_2)P_S(H|Z_2)dH \neq P_T(Y|W, Z_2)$.

$P_S(H) \neq P_T(H)$ **and** $P_S(H|Z_2) = P_T(H|Z_2)$

This will provide an outcome that is the reverse of the above: $P_S(Y|W, Z_2) = \int P_S(Y|W, H, Z_2)P_S(H|Z_2)dH = P_T(Y|W, Z_2)$, thus the invariant condition is given by $P(Y|W, Z_2)$. In contrast, $P(Y|W)$ will not be invariant: $P_S(Y|W) = \int P_S(Y|W, H, Z_2)P_S(H)dH \neq P_T(Y|W)$.

$P_S(H) = P_T(H)$ **and** $P_S(H|Z_2) = P_T(H|Z_2)$

In this case, both the conditionals will be invariant, the proof is the same as above.

$P_S(H) \neq P_T(H)$ **and** $P_S(H|Z_2) \neq P_T(H|Z_2)$

In this case, neither of the conditionals will be invariant, the proof being again the same as above.

Given that H is unobserved, the empirical evidence for the distribution of $P(H)$ and $P(H|Z_2)$ across domains is not easily recovered given a single source domain and a single target domain[1] Similarly, one might not be able to theoretically prove that there is not another unobserved variable, H_2, for which no proxy was measured and whose marginal distribution might differ across domains ($P_S(H_2) \neq P_T(H_2)$). Given data from only one domain, we do not have any clear way to empirically show that H_2 does not exist and does not throw off the predictive distribution by a potentially large degree.

A reasonable seeming alternative would then be to try and find labeled data from more than one domain. If one has such data from two domains, $\mathcal{D}_{S1}$ and $\mathcal{D}_{S2}$, then one could directly check the conditional invariance of $P_{S1}(Y|W) = P_{S2}(Y|W)$ and $P_{S1}(Y|W, Z_2) = P_{S2}(Y|W, Z_2)$. Under the assumption that any changes to unobservable variables between $\mathcal{D}_S$ and $\mathcal{D}_T$ will also exist between $\mathcal{D}_{S1}$ and $\mathcal{D}_{S2}$, one can thus be said to be confirming the transportability of the two possible predictive conditional distributions.

This is very close to the idea pursued by Rojas-carulla et al (Rojas-Carulla et al. 2018). They operationalize the testing of invariance of the conditional to it's domain through an independence test between the residuals of a parametric model for the output and the index label of the domain: $P(Y - f(X^*, S)) = P(Y - f(X^*))$, where $S = \{1, 2, \ldots, K\}$ for K source domains. They use this constraint to look for a subset of the features $X^* \in X$ for which conditional invariance holds.

Christina Henze-Deml (2017) proposes a similar solution to the problem of domain adaptation in image recognition. Using a series of images

[1] It is potentially possible to say something about $P(H|Z_2)$. Under the assumption that H is the direct cause of Z_2, without any unobserved confounding variables, we can say that $P(Z_2|H)$ will be invariant and independent from $P(Z_2)$ but $P(H|Z_2)$ will not (see Daniusis et al. 2010; Schölkopf et al. 2012; Peters, Janzing, and Schölkopf 2017, for an exposition of this feature which is closely related to weak exogoneity). Thus, $P(H|Z_2)$ will be invariant across domains if $P_S(Z_2) = P_T(Z_2)$. This is a testable assumption!

in which the same individual, with the same causal characteristics, is captured across multiple domains subject to shifts in orthogonal features. A regularization term is added to the optimization problem of the neural network trained on the source domains. The term penalizes the conditional variance of the prediction $\mathbb{V}\big[\hat{Y}|g(X)\big]$ for feature representation function $g(\cdot)$ applied to features where the "causal characteristics" are known to be the same. Through the regularization, the function $g(\cdot)$ learns a representation for which the conditional distribution of the outcome is invariant across domains.

Thus, the general formulation of these techniques is to consider the problem one of feature representation, where $g : \mathcal{Z} \to \mathbb{R}^D$ consists of a representation function that might search among variables Z or search over latent representations of variables Z to discover a model that is predictive of the outcome, subject to the constraint of a conditional distribution that is invariant across domains. Quite naturally, as done in both of the above references, one must do this in a way to avoid problems of multiple testing and overfitting either through levels of significance for variable inclusion or cross validation.

4.5 Treatment Effects

In both the econometric formulations of domain adaptation as well as those from machine learning, the focus of the literature has been on predicting the outcome itself. In many cases, a policy maker might be more interested in predicting the treatment effect, which can be thought of as a relaxation of the problem of predicting outcomes. Indeed, if one had the predicted outcomes, one would have the treatment effect, but not the other way around.

What implication does treatment effect prediction have for the extension of the frameworks introduced above? Consider the case where the true data generating process of the outcome, Y, is linearly separable into two parts:

$$Y(W, Z) = f_t(W, Z_1) + f_e(Z_1, Z_2)$$

Where $Z = \{Z_1, Z_2\}$, the set of observed covariates and $W \in \{W_0, W_1\}$ a binary treatment variable, as above. Let $\tau(z)$ be the conditional average treatment effect:

$$\tau(z) = \mathbb{E}\big[Y(W_0, Z) - Y(W_1, Z)|Z = z\big] = \tau(z_1)$$

Thus, $\tau(\cdot)$ only relies on a subset of $Z_1 \subset Z$. Treating τ_i as a random variable (dropping the explicit dependence on z_1) this allows us to achieve the covariate shift assumption with the invariance of the conditional distribution $P(\tau|Z_1)$ across domains, which relies on weakly fewer variables. Given all the threats to transportability outlined by our propositions above, this could be a major benefit in practice if the outcome is a linear combination of causes.

Estimating the treatment effect, however, comes with its own complications. Due to Holland's (1986) *Fundamental problem of Causal Inference*, τ_i is not actually observed and estimating its conditional distribution relies on additional assumptions (Firpo 2007). One must also also employ methods designed to find the heterogeneous treatment effect variables that make up Z_1 in a way which does not run into multiple testing problems. Luckily, there has been recent interest in developing powerful nonparametric techniques, such as causal trees (Athey and G. Imbens 2016), that perform exactly this function.

Thus, by combining these recent techniques from econometrics and machine learning, I argue that it is both desirable and likely possible to develop techniques to predict the effect of a policy, W, on a new domain $\mathcal{D}_T$, given conditional treatment effect distributions, $P_1(\tau|g(Z_1)), \ldots, P_S(\tau|g(Z_1))$, from source domains $\mathcal{D}_S$ and a learned feature representation function $g(\cdot)$ that enforces conditional invariance such that the covariance shift assumption holds and importance estimation methods can be used to predict effects on a target population.

5 Conclusions

Banerjee and Duflo (2014) discuss concerns to experimental validity and generalization. As a primary tool to address external validity of RCTs, they emphasize the need for replication studies. Indeed, they posit what can be thought of as asymptotic theory of domain adaptation for RCTs:

> "If we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any given set of covariates."

Asymptotic theory can be comforting, it's nice to know that the road we are on leads somewhere. However, the key term "any given set of covariates" must be restricted for this to be actionable in the finite lifespan

of our species. The definition of those covariates comes down to defining the uniformity of nature assumption posed by Hume. If the covariate set is infinite, then we assume nothing regarding the nature of uniformity, and induction is impossible.

How, then, do we define those covariates? Which parts of nature must be uniform and which parts are allowed to change? How many studies are enough studies and does that depend on where on intends to implement a new policy?

I have laid out a research agenda to answer those questions, and have shown that:

1. If the covariate shift assumption holds, importance estimation techniques can be used to predict in new contexts given data in a previous one.

2. In the face of unobservables, the covariate shift assumption must be proven to hold empirically by using labeled data from multiple contexts.

3. If that labeled data is experimental, it can be reasonable to estimate directly the treatment effect, potentially reducing the feature space and increasing the possibility of transportability.

References

[1] Alberto Abadie and Guido W. Imbens. "Large Sample Properties of Matching Estimators". In: *Econometrica* 74.1 (2006), pp. 235–267. ISSN: 0012-9682. DOI: 10.1111/j.1468-0262.2006.00655.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2006.00655.x/abstract.

[2] John Aldrich. "Autonomy". In: *Oxford Economic Papers* 41.1 (1989), pp. 15–34. ISSN: 1464-3812. DOI: 10.1093/oxfordjournals.oep.a041889. URL: https://academic.oup.com/oep/article/2364178/AUTONOMY.

[3] Hunt Allcott. "Site Selection Bias in Program Evaluation". In: *The Quarterly Journal of Economics* 130.3 (Aug. 2015), pp. 1117–1165. ISSN: 0033-5533. DOI: 10.1093/qje/qjv015. URL: http://www.nber.org/papers/w18373.pdf%20https://academic.oup.com/qje/article-lookup/doi/10.1093/qje/qjv015.

[4] Manuela Angelucci, Dean Karlan, and Jonathan Zinman. "Microcredit impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco". In: *American Economic Journal: Applied Economics* 7.1 (2015), pp. 151–182. ISSN: 19457790. DOI: 10.1257/app.20130537.

[5] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.27 (2016), pp. 7353–7360. ISSN: 10916490. DOI: 10.1073/pnas.1510489113.

[6] Orazio Attanasio et al. "The impacts of microfinance: Evidence from joint-liability lending in Mongolia". In: *American Economic Journal: Applied Economics* 7.1 (2015), pp. 90–122. ISSN: 19457790. DOI: 10.1257/app.20130489.

[7] Britta Augsburg et al. "The impacts of microcredit: Evidence from Bosnia and Herzegovina". In: *American Economic Journal: Applied Economics* 7.1 (2015), pp. 183–203. ISSN: 19457790. DOI: 10.1257/app.20130272.

[8] Abhijit V. Banerjee, Sylvain Chassang, and Erik Snowberg. "Decision Theoretic Approaches to Experiment Design and External Validity". In: *SSRN Electronic Journal* (2016). DOI: 10.2139/ssrn.2770140.

[9] Abhijit V. Banerjee and Esther Duflo. *Poor Economics: rethinking poverty & the ways to end it*. 2011, p. 302.

[10] Abhijit V. Banerjee and Esther Duflo. "The experimental approach to development economics". In: *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences* (2014), pp. 78–114. DOI: 10.1146/annurev.economics.050708.143235.

[11] Abhijit Banerjee et al. "The miracle of microfinance? Evidence from a randomized evaluation". In: *American Economic Journal: Applied Economics* 7.1 (2015), pp. 22–53. ISSN: 19457790. DOI: 10.1257/app.20130533.

[12] Shai Ben-David et al. "Analysis of Representations for Domain Adaptation". In: (2006).

[13] James Bisbee et al. "Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect". In: *Journal of Labor Economics* 35.S1 (2017), S99–S147. ISSN: 0734-306X. DOI: 10.1086/691280.

[14] Nancy Cartwright and Jeremy Hardie. “Evidence-based policy: a practical guide to doing it better”. In: *Choice Reviews Online* 50.10 (2013), pp. 50–5831–50–5831. ISSN: 0009-4978. DOI: 10.5860/choice.50-5831.

[15] Thomas D Cook and D T Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. English. Houghton Mifflin, 1979.

[16] Bruno Crépon et al. “Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco”. In: *American Economic Journal: Applied Economics* 7.1 (Jan. 2015), pp. 123–150. ISSN: 1945-7782. DOI: 10.1257/app.20130535. URL: http://pubs.aeaweb.org/doi/10.1257/app.20130535.

[17] Povilas Daniusis et al. “Inferring deterministic causal relations”. In: *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010* (2010), pp. 143–150.

[18] Angus Deaton. “Instruments, Randomization, and Learning about Development”. In: *Journal of Economic Literature* 48.2 (June 2010), pp. 424–455. ISSN: 0022-0515. DOI: 10.1257/jel.48.2.424. URL: http://pubs.aeaweb.org/doi/10.1257/jel.48.2.424.

[19] Angus Deaton and Nancy Cartwright. “Understanding and misunderstanding randomized controlled trials”. In: *Social Science & Medicine* 210.October 2017 (2018), pp. 2–21. ISSN: 0277-9536. DOI: 10.1016/j.socscimed.2017.12.005. URL: https://doi.org/10.1016/j.socscimed.2017.12.005.

[20] Robert F Engle, David F Hendry, and Jean-francois Richard. “Exogeneity”. In: *Econometrica* 51.2 (Mar. 1983), p. 277. ISSN: 00129682. DOI: 10.2307/1911990. URL: https://www.jstor.org/stable/1911990?origin=crossref.

[21] B Y Sergio Firpo. “Efficient Semiparametric Estimation of Quantile Treatment Effects”. In: 75.1 (2007), pp. 259–276.

[22] R.A. Fisher. “The Design of Experiments”. In: (1935). ISSN: 00218790.

[23] Ragnar Frisch. *The Foundations of Econometric Analysis*. Ed. by David F. Hendry and Mary S. Morgan. Cambridge: Cambridge University Press, 1995, pp. 407–419. ISBN: 9781139170116. DOI: 10.1017/CBO9781139170116. URL: http://ebooks.cambridge.org/ref/id/CBO9781139170116.

[24] Michael Gechter. “Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India”. 2015.

[25] Trygve Haavelmo. “The Probability Approach in Econometrics”. In: *Econometrica* 12 (July 1944), p. iii. ISSN: 00129682. DOI: 10.2307/1906935. URL: https://www.jstor.org/stable/1906935?origin=crossref.

[26] James J Heckman. “Econometric Causality”. In: *International Statistical Review* 76.1 (Apr. 2008), pp. 1–27. ISSN: 0306-7734. DOI: 10.1111/j.1751-5823.2007.00024.x. URL: http://doi.wiley.com/10.1111/j.1751-5823.2007.00024.x.

[27] James J Heckman and Jeffrey A Smith. “Assessing the Case for Social Experiments”. In: *Journal of Economic Perspectives* 9.2 (1995), pp. 85–110. ISSN: 0895-3309. DOI: 10.1257/jep.9.2.85.

[28] Christina Heinze-Deml and Nicolai Meinshausen. “Conditional Variance Penalties and Domain Shift Robustness”. In: 2009 (Oct. 2017). arXiv: 1710.11469. URL: http://arxiv.org/abs/1710.11469.

[29] Paul W Holland. “Statistics and Causal Inference: Rejoinder”. In: *J Am Stat Assoc* 81.396 (1986), p. 968. ISSN: 0162-1459. DOI: 10.2307/2289069.

[30] V. Joseph Hotz, Guido W. Imbens, and Julie H. Mortimer. “Predicting the efficacy of future training programs using past experiences at other locations”. In: *Journal of Econometrics* 125.1-2 SPEC. ISS. (2005), pp. 241–270. ISSN: 03044076. DOI: 10.1016/j.jeconom.2004.04.009.

[31] David Hume. *An enquiry concerning human understanding*. 1748.

[32] Leonid Hurwicz. “On the Structural Form of Interdependent Systems”. In: *Studies in Logic*. Vol. 44. Board of Trustees of the Leland Stanford Junior University, 1966, pp. 232–239. DOI: 10.1016/S0049-237X(09)70590-7. URL: http://dx.doi.org/10.1016/S0049-237X(09)70590-7%20https://linkinghub.elsevier.com/retrieve/pii/S0049237X09705907.

[33] C. F. Manski and Steven R. Lerman. “The Estimation of Choice Probabilities from Choice Based Samples Author (s): Charles F . Manski and Steven R . Lerman”. In: *Econometrica* 45.8 (1977), pp. 1977–1988.

[34] Charles F. Manski. *Public Policy in an Uncertain World*. Cambridge, MA and London, England: Harvard University Press, Jan. 2012. ISBN: 9780674067547. DOI: 10.4159/harvard.9780674067547. URL: http://www.degruyter.com/view/books/harvard.9780674067547/harvard.9780674067547/harvard.9780674067547.xml.

[35] Rachael Meager. "Aggregating Distributional Treatment Effects : A Bayesian Hierarchical Analysis of the Microcredit Literature". In: (2018). DOI: 10.2139/ssrn.2620834.

[36] John Stuart Mill. *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. Vol. 1. Longmans, green, and Company, 1884.

[37] Mary S. Morgan. "The History of Econometric Ideas." In: *The Economic History Review* 44.2 (1991), p. 391. ISSN: 00130117. DOI: 10.2307/2598321.

[38] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (Oct. 2010), pp. 1345–1359. ISSN: 1041-4347. DOI: 10.1109/TKDE.2009.191. arXiv: 1603.06111. URL: https://doi.org/10.1016/j.artmed.2018.03.006%7B%5C%%7D0Ahttp://arxiv.org/abs/1801.06146%7B%5C%%7D0Ahttp://arxiv.org/abs/1902.01382%7B%5C%%7D0Ahttp://arxiv.org/abs/1902.09092%7B%5C%%7D0Ahttp://arxiv.org/abs/1703.06345%7B%5C%%7D0Ahttp://dblp.org/rec/conf/aaai/ZhouPTH16%7B%5C%%7D0Awww.aaai.org%7B%5C%%7D0Ahttp://arxiv.or%20http://ieeexplore.ieee.org/document/5288526/.

[39] Judea Pearl. *Causality Second Edition*. 2000, p. 386. ISBN: 0521773628. DOI: citeulike-article-id:3888442.

[40] Judea Pearl and Elias Bareinboim. "External validity: From do-calculus to transportability across populations". In: *Statistical Science* 29.4 (2014), pp. 579–595. ISSN: 08834237. DOI: 10.1214/14-STS486. arXiv: arXiv:1503.01603v1.

[41] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. "Causal inference using invariant prediction : identification and confidence intervals". In: (2015), pp. 1–51. arXiv: arXiv:1501.01332v3.

[42] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. December. 2017, pp. 1214–1216. ISBN: 9780262037310. URL: file:///tmp/mozilla%7B%5C_%7Dkuebler0/11283.pdf%7B%5C%%7D0Ahttps://books.google.com/books?hl=en%7B%5C&%7Dlr=%7B%5C&%7Did=XPpFDwAAQBAJ%7B%5C&%7Doi=fnd%7B%5C&%7Dpg=PR7%7B%5C&%7Dots=GilrFtvtAZ%7B%5C&%7Dsig=KYfn3zHUHfiWXhBTvmvwX33bLlA.

[43] Lant Pritchett and Justin Sandefur. "Learning from Experiments when Context Matters". In: *American Economic Review* 105.5 (May 2015), pp. 471–475. ISSN: 0002-8282. DOI: 10.1257/aer.p20151016. URL: http://eds.b.ebscohost.com.ezproxy.lib.usf.edu/eds/pdfviewer/pdfviewer?vid=5%7B%5C&%7Dsid=aef21225-20b8-458e-842b-e860243ea5f7%7B%5C%%7D40sessionmgr104%7B%5C&%7Dhid=104%20http://pubs.aeaweb.org/doi/10.1257/aer.p20151016.

[44] Mateo Rojas-Carulla et al. "Invariant models for causal transfer learning". In: *Journal of Machine Learning Research* 19 (2018), pp. 1–34. ISSN: 15337928.

[45] Paul R. Rosenbaum. "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies". In: *American Statistician* 59.2 (2005), pp. 147–152. ISSN: 00031305. DOI: 10.1198/000313005X42831.

[46] Mark Richard Rosenzweig and Christopher Udry. "External Validity in a Stochastic World: Evidence from Low-Income Countries". In: *SSRN Electronic Journal* (2019). DOI: 10.2139/ssrn.3392657.

[47] Bernhard Schölkopf et al. "On causal and anticausal learning". In: *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*. Vol. 2. 2012, pp. 1255–1262. ISBN: 9781450312851.

[48] Hidetoshi Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of Statistical Planning and Inference* 90 (2000), pp. 227–244. URL: www.elsevier.com/locate/jspi.

[49] Masashi Suigyama et al. "Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation". In: *Proceedings of the 20th International Conference on Neural Information Processing Systems* (2007), pp. 1433–1440. ISSN: 00203157. DOI: 10.1007/s10463-008-0197-x. URL: http://eprints.pascal-network.org/archive/00003287/.

[50] Alessandro Tarozzi, Jaikishan Desai, and Kristin Johnson. "The impacts of microcredit: Evidence from Ethiopia". In: *American Economic Journal: Applied Economics* 7.1 (2015), pp. 54–89. ISSN: 19457790. DOI: 10.1257/app.20130475.

[51] William R. Shadish, Thomas D Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. 2001, p. 656. ISBN: 0395615569.

[52] Kun Zhang, Jiji Zhang, and Bernhard Schölkopf. “Distinguishing Cause from Effect Based on Exogeneity”. In: (Apr. 2015). arXiv: 1504.05651. URL: http://arxiv.org/abs/1504.05651.