Chernozhukov, Victor; Demirer, Mert; Duflo, Esther; Fernandez-Val, Ivan

Working Paper

Generic machine learning inference on heterogenous treatment effects in randomized experiments

cemmap working paper, No. CWP61/17

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Chernozhukov, Victor; Demirer, Mert; Duflo, Esther; Fernandez-Val, Ivan (2017) : Generic machine learning inference on heterogenous treatment effects in randomized experiments, cemmap working paper, No. CWP61/17, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2017.6117

This Version is available at:
https://hdl.handle.net/10419/189806

cemmap
centre for microdata methods and practice

Generic machine learning inference on heterogenous treatment effects in randomized experiments

Victor Chernozhukov
Mert Demirer
Esther Duflo
Ivan Fernandez-Val

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP61/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

GENERIC MACHINE LEARNING INFERENCE ON HETEROGENOUS TREATMENT EFFECTS IN RANDOMIZED EXPERIMENTS

V. CHERNOZHUKOV, M. DEMIRER, E. DUFLO, I. FERNANDEZ-VAL

Abstract. We propose strategies to estimate and make inference on key features of heterogeneous effects in randomized experiments. These key features include *best linear predictors of the effects* using machine learning proxies, *average effects sorted by impact groups*, and *average characteristics of most and least impacted units*. The approach is valid in high dimensional settings, where the effects are proxied by machine learning methods. We post-process these proxies into the estimates of the key features. Our approach is generic, it can be used in conjunction with penalized methods, deep and shallow neural networks, canonical and new random forests, boosted trees, and ensemble methods. Our approach is agnostic and does not make unrealistic or hard-to-check assumptions; we don't require conditions for consistency of the ML methods. Estimation and inference relies on repeated data splitting to avoid overfitting and achieve validity. For inference, we take medians of p-values and medians of confidence intervals, resulting from many different data splits, and then adjust their nominal level to guarantee uniform validity. This variational inference method is shown to be uniformly valid and quantifies the uncertainty coming from both parameter estimation and data splitting. The inference method could be of substantial independent interest in many machine learning applications. An empirical application to the impact of micro-credit on economic development illustrates the use of the approach in randomized experiments. An additional application to the impact of the gender discrimination on wages illustrates the potential use of the approach in observational studies, where machine learning methods can be used to condition flexibly on very high-dimensional controls.

Key words: Agnostic Inference, Machine Learning, Confidence Intervals, Causal Effects, Variational P-values and Confidence Intervals, Uniformly Valid Inference, Quantification of Uncertainty, Sample Splitting, Multiple Splitting, Assumption-Freeness

1. Introduction

Randomized experiments play an important role in the evaluation of social and economic programs and medical treatments (e.g., [35, 27]). Researchers and policy makers are often interested in features of the impact of the treatment that go beyond the simple average treatment effects. In particular, very often, they want to know whether treatment effect depends on basic covariates, such as gender, age, etc. It is essential to assess if the impact of the program would generalize to a different population with different characteristics, and for economists, to better understand the driving mechanism behind a particular program effects.

One issue with reporting treatment effects split by subgroups, however, is that there are often a large number of potential sample splits: choosing subgroups ex-post opens the possibility

Date: December, 2017.

We thank Moshe Buchinsky, Denis Chetverikov, Siyi Luo, Susan Murphy, Whitney Newey, and seminar participants at UCLA for valuable comments.

of overfitting. To solve this problem, medical journals and the FDA require pre-registering subsample of interest in medical trials *in advance*. In economics, this approach has gained some traction, with the adoption of pre-analysis plan (which can be filed in the AEA registry for randomized experiments). Restricting heterogeneity analysis to pre-registered subgroups, however, amounts to throwing away a large amount of potentially valuable information, especially now that many researchers collect large baseline data sets. It should be possible to use the data to discover *ex post* whether there is any relevant heterogeneity in treatment effect by co-variates.

To do this in a disciplined fashion and avoid the risk of overfitting, scholars have recently proposed using machine learning tools (see e.g. [6] and below for a review). Indeed, ML tools seem to be ideal to explore heterogeneity of treatment effect, when researchers have access to a potentially large array of baseline variables to form subgroup, and little guiding principles on which of those are likely to be relevant. Several recent papers, which we review below, develop methods for detecting heterogeneity in treatment effects. Empirical researchers have taken notice.[1]

This paper develops a generic approach to use any of the tools of machine learning to predict and make inference on heterogeneous treatment or policy effects. A core difficulty of applying ML tools to the estimation of heterogenous causal effects is that, while they are successful in prediction empirically, it is much more difficult to obtain uniformly valid inference. In fact, in high dimensional settings, absent strong assumptions, generic ML tools may not even produce consistent estimates of the *conditional average treatment effect* (or CATE) (the difference in the expected potential outcomes between treated and control groups conditional on the covariates Z).

Previous attempts to solve this problem focus on specific tools and situations where those assumptions are satisfied. Our approach to resolve the fundamental impossibilities in non-parametric inference is different. Motivated by [29], instead of attempting to get consistent estimate and uniformly valid inference on CATE itself, we focus on providing valid estimation and inference for *features* of CATE. We start by building a ML proxy predictor of CATE, and then develop valid inference for features of the CATE based on this proxy predictor. In particular, we develop valid inference for the following three objects, which are likely to be of interest to applied researchers and policy makers: First, the **Best Linear Predictor** (BLP) of the CATE based on the ML proxy predictor, second, the **Sorted Group Average Treatment Effects**, or the average treatment effect sorted induced by the groups of the ML proxy predictors, third, the **Classification Analysis** (CLAN): the average characteristics of the most and least affected units defined in terms of the ML proxy predictor. Thus, we can find out if there is detectable heterogeneity in the treatment effect based on observables, and if there is any, what is the treatment effect for different bins. And finally we can describe which of the covariates is correlated with this heterogeneity.

[1] In the last few months alone, several new empirical papers in economics used ML methods to estimate heterogenous effects. E.g. [44], which shows that villagers outperform the machine learning tools when they predict heterogeneity in returns to capital. [25] predict who benefits the most from a summer internship projects. The methodological papers reviewed later also contain a number of empirical applications.

Thus, in the trade off between more restrictive assumptions and a more ambitious estimation, we chose a different point than previous papers: focus on coarser objects of the function rather than the function itself, but make as little assumptions as possible. This seems to be a worthwhile sacrifice: the objects that for which we have developed inference appear to us at this point to be the most relevant, but in the future, one could easily develop methods for other objects of interest.

The Model and Key Causal Functions. Let $Y(1)$ and $Y(0)$ be the potential outcomes in the treatment state 1 and the non-treatment state 0; see [46]. Let Z be a vector of of covariates that characterize the observational units. The main causal functions are the baseline conditional average (BCA):

$$b_0(Z) := \mathrm{E}[Y(0) \mid Z], \tag{1.1}$$

and the conditional average treatment effect (CATE):

$$s_0(Z) := \mathrm{E}[Y(1) \mid Z] - \mathrm{E}[Y(0) \mid Z]. \tag{1.2}$$

Suppose the binary treatment variable D is randomly assigned conditional on Z, with probability of assignment depending only on a subvector of stratifying variables $Z_1 \subseteq Z$, namely

$$D \perp\!\!\!\perp (Y(1), Y(0)) \mid Z, \tag{1.3}$$

and the propensity score is known and is given by

$$p(Z) := \mathrm{P}[D = 1 \mid Z] = \mathrm{P}[D = 1 \mid Z_1], \tag{1.4}$$

which we assume is bounded away from zero or one:

$$p(Z) \in [p_0, p_1] \subset (0, 1). \tag{1.5}$$

The observed outcome is $Y = DY(1) + (1 - D)Y(0)$. Under the stated assumption, the causal functions are identified by the components of the regression function of Y given D, Z:

$$Y = b_0(Z) + Ds_0(Z) + U, \;\; \mathrm{E}[U \mid Z, D] = 0,$$

that is,

$$b_0(Z) = \mathrm{E}[Y \mid D = 0, Z], \tag{1.6}$$

and

$$s_0(Z) = \mathrm{E}[Y \mid D = 1, Z] - \mathrm{E}[Y \mid D = 0, Z]. \tag{1.7}$$

We observe $\mathrm{Data} = (Y_i, Z_i, D_i)_{i=1}^N$, consisting of i.i.d. copies of the random vector (Y, Z, D) having probability law P. Expectation with respect to P is denoted by $\mathrm{E} = \mathrm{E}_P$. The probability law of the entire data is denoted by $\mathbb{P} = \mathbb{P}_P$ and the corresponding expectation is denoted by $\mathbb{E} = \mathbb{E}_P$.

Properties of Machine Learning Estimators of $s_0(Z)$ Motivating the Agnostic Approach. Machine learning (ML) is a name attached to a variety of new, constantly evolving statistical learning methods: Random Forest, Boosted Trees, Neural Networks, Penalized Regression, Ensembles, and Hybrids (see, e.g., [50] for a recent review, and [28] for a prominent textbook treatment). In modern high-dimensional settings, machine learning methods effectively explore the various forms of non-linear structured sparsity to yield "good" approximations to $s_0(z)$ whenever such assumptions are valid, based on equations (1.6) and (1.7). As a result these methods often work much better than classical methods in high-dimensional settings, and have found widespread uses in industrial and academic applications.

Motivated by the practical predictive success of the methods, it is really tempting to apply the ML methods directly to try to learn the CATE function $z \mapsto s_0(z)$ (by learning the two regression functions for treated and untreated and taking the difference). However, it is hard, if not impossible, to obtain uniformly valid inference on $z \mapsto s_0(z)$ using generic ML methods, under credible assumptions and practical tuning parameter choices. There are several fundamental reasons as well as huge gaps between theory and practice that are responsible for this.

One fundamental reason is that the ML methods might not even produce consistent estimators of $z \mapsto s_0(z)$ in high dimensional settings. For example, if z has dimension d and the target function $z \mapsto s_0(z)$ is assumed to have p continuous and bounded derivatives, then the worst case (minimax) lower bound on the rate of learning this function cannot be better than $N^{-p/(2p+d)}$ as $N \to \infty$, as shown by Stone [47]. Hence if p is fixed and d is also small, slowly increasing with N, such as $d \geqslant \log N$, then there exists no consistent estimator of $z \mapsto s_0(z)$ generally.

Hence, ML estimators cannot be regarded as consistent, unless further very strong assumptions are made. Examples of such assumptions include structured forms of linear and non-linear sparsity and super-smoothness. While these (sometime believable and yet untestable) assumptions make consistent adaptive estimation possible (e.g.,[17]), inference remains a more difficult problem, as adaptive confidence sets do not exist even for low-dimensional nonparametric problems ([41, 29]). Indeed, adaptive estimators (including modern ML methods) have biases of comparable or dominating order as compared to sampling error. Further assumptions such as "self-similarity" are needed to bound the biases and expand the confidence bands by the size of bias (see [30, 22]) to produce partly adaptive confidence bands. For more traditional statistical methods there are constructions in this vein that make use of either undersmoothing or bias-bounding arguments ([30, 22]). These methods, however, are not yet available for ML methods in high dimensions (see, however, [31] for a promising approach called "targeted undersmoothing" in sparse linear models).

Suppose we did decide to be optimistic (or panglossian) and imposed the strong assumptions, that made the theoretical versions of the ML methods provide us with high-quality consistent estimators of $z \mapsto s_0(z)$ and valid confidence bands based on them. This would still not give as a practical construction we'd want for our applications. The reason is that there is a huge gap between

theoretical versions of the ML methods appearing in various theoretical papers and the practical versions (with the actual, data-driven tuning parameters) coded up in statistical computing packages used by practitioners[2].[3] The use of ML, for example, involves many tuning parameters with practical rules for choosing them, while theoretical works provide little guidance or backing for such practical rules; see e.g., the influential book [28] for many examples of such rules.[4]

In this paper we shall take an agnostic view and we shall not rely on any structured assumptions, which might be difficult to verify or believe in practice, and we don't impose conditions that make the ML estimators consistent. We simply treat ML as providing proxy predictors for the objects of interest.

Our Agnostic Approach. Our approach will be radically different – we propose strategies for estimation and inference on

$$\textit{key features of} \quad s_0(Z) \text{ rather than } s_0(Z) \text{ itself.}$$

Because of this difference in focus we can avoid making strong assumptions about the properties of the ML estimators.

Let (M, A) denote a random partition of the set of indices $\{1, \ldots, N\}$. The strategies that we consider rely on random splitting of

$$\mathrm{Data} = (Y_i, D_i, Z_i)_{i=1}^N$$

into a main sample, denoted by $\mathrm{Data}_M = (Y_i, D_i, Z_i)_{i \in M}$, and an auxiliary sample, denoted by $\mathrm{Data}_A = (Y_i, D_i, Z_i)_{i \in A}$. We will sometimes refer to these samples as M and A. We assume that the main and auxiliary samples are approximately equal in size, though this is not required theoretically.

From the auxiliary sample A, we obtain ML estimators of the baseline and treatment effects, which we call the proxy predictors,

$$z \mapsto B(z) = B(z; \mathrm{Data}_A) \text{ and } z \mapsto S(z) = S(z; \mathrm{Data}_A).$$

These are possibly biased and noisy predictors of $b_0(z)$ and $s_0(z)$, and in principle, we do not even require that they are consistent for $b_0(z)$ and $s_0(z)$. We simply treat these estimates as proxies, which we post-process to estimate and make inference on the features of the CATE $z \mapsto s_0(z)$. We condition on the auxiliary sample Data_A, so we consider these maps as frozen, when working with the main sample.

[2]There are cases where such gap does not exist, e.g., see [13, 15] for the lasso.

[3]For example, even the wide use of K-fold cross-validation in high-dimensional settings for machine learning remains theoretically unjustified (there do exist, however, related subsample-based methods [52, 40] that do achieve near-oracle performance for tuning selection).

[4]Unfortunately, theoretical work often only provides existence results: there exist theoretical ranges of the tuning parameters that make the simple versions of the methods work for predictive purposes (under very strong assumptions), leaving no satisfactory guide to practice.

Using the main sample and the proxies, we shall target and develop valid inference about *key features of* $s_0(Z)$ rather than $s_0(Z)$, which include

(1) **Best Linear Predictor** (BLP) of the CATE $s_0(Z)$ based on the ML proxy predictor $S(Z)$;

(2) **Sorted Group Average Treatment Effects** (GATES): average of $s_0(Z)$ (ATE) by heterogeneity groups induced by the ML proxy predictor $S(Z)$;

(3) **Classification Analysis** (CLAN): average characteristics of the most and least affected units defined in terms of the ML proxy predictor $S(Z)$.

Our approach is *generic* with respect to the ML method being used, and is *agnostic* about its formal properties.

Our point estimation will make use of many splits of the data into main and auxiliary samples to produce robust estimates. Our estimation and inference will systematically account for two sources of uncertainty:

(I) **Estimation uncertainty** conditional on the auxiliary sample.

(II) **Splitting uncertainty** induced by random partitioning of the data into the main and auxiliary samples.

Because we account for the second source, we call the resulting collection of methods as variational estimation and inference methods (VEINs). For point estimates we report the median of the estimated key features over different random splits of the data. For the confidence intervals we take the medians of many random conditional confidence sets and we adjust their nominal confidence level to reflect the splitting uncertainty. We construct p-values by taking medians of many random conditional p-values and adjust the nominal levels to reflect the splitting uncertainty. Note that considering many different splits and accounting for variability caused by splitting is very important. Indeed, with a single splitting practice, empiricists may unintentionally look for a "good" data split, which supports their prior beliefs about the likely results, thereby invalidating inference.[5]

Relationship to the Literature. We focus the review strictly on the literatures about estimation and inference on heterogeneous effects and inference using sample splitting.

We first mention work that uses linear and semiparametric regression methods. A semiparametric inference method for characterizing heterogeneity, called the sorted effects method, was given in [20]. This approach does provide a full set of inference tools, including simultaneous

[5]This problem is "solved" by fixing the Monte-Carlo seed and the entire data analysis algorithm before the empirical study. Even if such a huge commitment is really made and followed, there is a considerable risk that the resulting data-split may be non-typical. Our approach allows one to avoid taking this risk.

bands for percentiles of the CATE, but is strictly limited to the traditional semiparametric estimators for the regression and causal functions. [31] proposed a sparsity based method called "targeted undersmoothing" to perform inference on heterogeneous effects. This approach does allow for high-dimensional settings, but makes strong assumptions on sparsity as well as additional assumptions that enable the targeted undersmoothing. A related approach, which allows for simultaneous inference on many coefficients (for example, inference, on the coefficients corresponding to the interaction of the treatment with other variables) was first given in [14] using a Z-estimation framework, where the number of interactions can be very large; see also [26] for a more recent effort in this direction, focusing on de-biased lasso in mean regression problems. This approach, however, still relies on a strong form of sparsity assumptions. [53] proposed a post-selection inference framework within the high-dimensional linear sparse models for the heterogeneous effects. The approach is attractive because it allows for some misspecification of the model.

Next we discuss the use of tree-based and other methods. [34] discussed the use of a heuristic support-vector-machine method with lasso penalization for classification of heterogeneous treatments into positive and negative ones. They used the Horvitz-Thompson transformation of the outcome (e.g., as in [33, 1]) such that that the new outcome becomes an unbiased, noisy version of CATE. [5] made use of the Horvitz-Thompson transformation of the outcome variable to inform the process of building causal trees, with the main goal of predicting CATE. They also provide a valid inference result on average treatment effects for groups defined by the tree leaves, conditional on the data split in two subsamples: one used to build the tree leaves and the one to estimate the predicted values given the leaves. This type of inference, however, does not account for splitting uncertainty.[6] [49] provided a subsampling-based construction of a causal random forest, providing valid pointwise inference for CATE (see also the review in [49] on prior uses of random forests in causal settings) for the case when covariates are very low-dimensional (and essentially uniformly distributed).[7] Unfortunately, this rules out the typical high-dimensional settings that arise in many empirical problems, including the ones considered in this paper.

Our approach is radically different from these existing approaches, in that we are changing the target, and instead of hunting for CATE $z \mapsto s_0(z)$, we focus on key features of $z \mapsto s_0(z)$. Our approach is generic, it can be used in conjunction with penalized methods, deep and shallow neural networks, canonical and new random forests, boosted trees, and ensemble methods. Our approach is agnostic and does not make unrealistic or hard-to-check assumptions, for example, we don't even require conditions for consistency of the ML methods. We simply treat the ML methods as

[6] In principle, our (generic) inference methods described below applies to the [5]'s problem to provide valid inference that does account for sample splitting uncertainty.

[7] The dimension d is fixed in [49]; the analysis relies on the Stone's model with smoothness index $\beta = 1$, in which no consistent estimator exists once $d \geqslant \log n$. It'd be interesting to establish consistency properties and find valid inferential procedures for the random forest in high-dimensional ($d \propto n$ or $d \gg n$) *approximately sparse* cases, with continuous and categorical covariates, but we are not aware of any studies that cover such settings, which are of central importance to us.

providing a proxy predictor $z \mapsto S(z)$, which we post-process to estimate and make inference on the key features of the CATE $z \mapsto s_0(z)$. Some of our strategies rely on Horvitz-Thompson transformations of outcome and some do not. The inspiration for our approach draws upon an observation in [29], namely that some fundamental impossibilities in non-parametric inference could be avoided if we focus inference on coarser features of the non-parametric functions rather than the functions themselves.

Our inference approach is also of independent interest, and could be applied to many problems, where sample splitting is used to produce ML predictions, e.g. [2]. Related references include [51, 42], where the ideas are related but quite different in details, which we shall explain below. The premise is the same, however, as in [42, 45] – we should not rely on a single random split of the data and should adjust inference in some way. Our approach takes the medians of many conditional confidence intervals as the confidence interval and the median of many conditional p-values as the p-value, and adjusts their nominal levels to account for the splitting uncertainty. Our construction of p-values builds upon ideas in [16, 42], though what we propose is radically simpler, and our confidence intervals appear to be brand new. Of course sample splitting ideas are classical, going back to [32, 38, 12, 23, 43], though having been mostly underdeveloped and overlooked for inference, as characterized by [45]. The overlooked status perhaps allowed us to innovate on the inference side.

2. Main Identification Results and Estimation Strategies

2.1. **BLP of CATE.** We consider two strategies for identifying and estimating the best linear predictor of $s_0(Z)$ using $S(Z)$:

$$\mathsf{BLP}[s_0(Z) \mid S(Z)] := \arg \min_{f(Z) \in \mathrm{Span}(1, S(Z))} \mathrm{E}[s_0(Z) - f(Z)]^2,$$

which, if exists, is defined by projecting $s_0(Z)$ on the linear span of 1 and $S(Z)$ in the space $L^2(P)$.

BLP of CATE: The First Strategy. Here we shall identify the coefficients of the BLP from the weighted linear projection:

$$Y = \alpha' X_1 + \beta_1 (D - p(Z)) + \beta_2 (D - p(Z))(S - \mathrm{E}S) + \epsilon, \;\; \mathrm{E}[w(Z)\epsilon X] = 0, \tag{2.1}$$

where $S := S(Z)$,

$$w(Z) = \{p(Z)(1 - p(Z))\}^{-1}, \quad X := (X_1, X_2)$$

$$X_1 := X_1(Z), \quad \text{e.g.,} \quad X_1 = [1, B(Z)],$$

$$X_2 := [D - p(Z), (D - p(Z))(S(Z) - (S - \mathrm{E}S)].$$

Note that the above equation uniquely pins down β_1 and β_2 under weak assumptions.

The interaction $(D - p(Z))(S - \mathrm{E}S)$ is orthogonal to $D - p(Z)$ under the weight $w(Z)$ and to all other regressors that are functions of Z under any Z-dependent weight.[8]

A consequence is our first main identification result, namely that

$$\beta_1 + \beta_2(S(Z) - \mathrm{E}S) = \mathsf{BLP}[s_0(Z) \mid S(Z)],$$

in particular $\beta_1 = \mathrm{E}s_0(Z)$ and $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$.

Theorem 2.1 (BLP 1)**.** *Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y and X have finite second moments and that $\mathrm{E}XX'$ is full rank. Then, (β_1, β_2) defined in* (2.1) *also solves the best linear predictor/approximation problem for the target $s_0(Z)$:*

$$(\beta_1, \beta_2)' = \arg\min_{b_1, b_2} \mathrm{E}[s_0(Z) - b_1 - b_2 S(Z)]^2,$$

in particular $\beta_1 = \mathrm{E}S_0(Z)$ and $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$.

The identification result is constructive. We can base the corresponding estimation strategy on the empirical analog:

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\beta}_1(D_i - p(Z_i)) + \widehat{\beta}_2(D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\epsilon}_i, \quad i \in M,$$

$$\mathbb{E}_{N,M}[w(Z_i)\widehat{\epsilon}_i X_i] = 0,$$

where $\mathbb{E}_{N,M}$ denotes the empirical expectation with respect to the main sample, i.e.

$$\mathbb{E}_{N,M} g(Y_i, D_i, Z_i) := |M|^{-1} \sum_{i \in M} g(Y_i, D_i, Z_i).$$

The properties of this estimator, conditional on the auxilliary data, are well known and follow as a special case of Lemma B.1 in the Appendix.

Comment 2.1 (Main Implications of the result)**.** If $S(Z)$ is a perfect proxy for $s_0(Z)$, then $\beta_2 = 1$. In general, $\beta_2 \neq 1$, correcting for noise in $S(Z)$. If $S(Z)$ is complete noise, uncorrelated to $s_0(Z)$, then $\beta_2 = 0$. Furthermore, if there is no heterogeneity, that is $s_0(Z) = s$, then $\beta_2 = 0$. Rejecting the hypothesis $\beta_2 = 0$ therefore means that there is both heterogeneity and $S(Z)$ is its relevant predictor. ■

Figure 1 provides two examples. The left panel shows a case without heterogeneity in the CATE where $s_0(Z) = 0$, whereas there right panel shows a case with strong heterogeneity in the CATE where $s_0(Z) = Z$. In both cases we evenly split 1000 observations between the auxiliary and main

[8]The orthogonalization ideas embedded in this strategy do have classical roots in econometrics (going back to at least Frisch and Waugh in the 30s), and similar strategies underlie the orthogonal or double machine learning approach (DML) in [21]. Our paper has different goals than DML, attacking the problem of inference on heterogeneous effects without rate and even consistency assumptions. The strategy here is more nuanced in that we are making it work under misspecification or inconsistent learning, which is likely to be true in very high-dimensional problems.

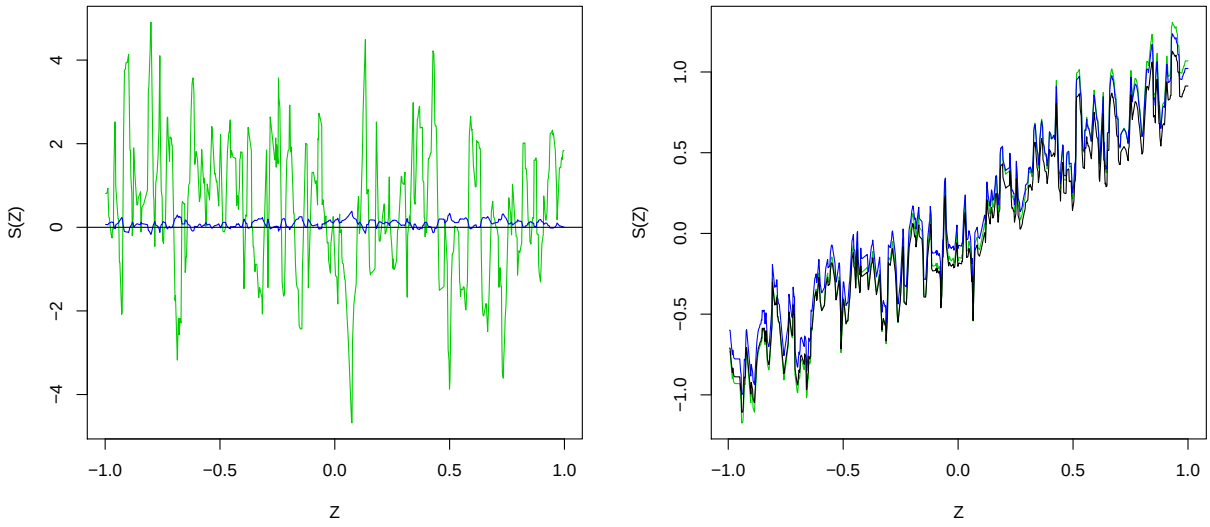

FIGURE 1. Example. In the left panel we have a homogeneous CATE $s_0(Z) = 0$; in the right panel we have heterogeneous CATE $s_0(Z) = Z$. The proxy predictor $S(Z)$ is produced by the Random Forest, shown by green line, the true BLP of CATE is shown by black line, and the estimated BLP of CATE is shown by blue line. The true and estimated BLP of CATE are more attenuated towards zero than the proxy predictor.

samples, Z is uniformly distributed in $(-1, 1)$, and the proxy predictor $S(Z)$ is estimated by random forest in the auxiliary sample following the standard implementation, see e.g. [28]. When there is no heterogeneity, post-processing the ML estimates helps reducing sampling noise bringing the estimated BLP close to the true BLP; whereas under strong heterogeneity the signal in the ML estimates dominates the sampling noise and the post-processing has little effect.

Comment 2.2 (Digression: Naive Strategy that is not Quite Right)**.** It is tempting and "more natural" to estimate

$$Y = \tilde{\alpha}_1 + \tilde{\alpha}_2 B + \tilde{\beta}_1 D + \tilde{\beta}_2 D(S - \mathrm{E}S) + \epsilon, \quad \mathrm{E}[\epsilon X] = 0.$$

This is a good strategy for predicting the conditional expectation of Y given Z and D. But, $\tilde{\beta}_2 \neq \beta_2$, and $\tilde{\beta}_1 + \tilde{\beta}_2(S - \mathrm{E}S)$ is not the best linear predictor of $s_0(Z)$. ■

BLP of CATE: The Second Strategy. The second strategy makes use of the Horvitz-Thompson transformation:

$$H = H(D, Z) = \frac{D - p(Z)}{p(Z)(1 - p(Z))}.$$

It is well known that the transformed response YH provides an unbiased signal about CATE:

$$\mathrm{E}[YH \mid Z] = s_0(Z)$$

and it follows that

$$\mathsf{BLP}[s_0(Z) \mid S(Z)] = \mathsf{BLP}[YH \mid S(Z)].$$

This simple strategy is completely fine for identification purposes, but can severely underperform in estimation and inference due to lack of precision. We can repair the deficiencies as follows.

We consider, instead, the linear projection:

$$YH = \mu' X_1 H + \beta_1 + \beta_2(S - \mathrm{E}S) + \tilde{\epsilon}, \quad \mathrm{E}\tilde{\epsilon}\tilde{X} = 0, \tag{2.2}$$

where $B := B(Z)$, $S := S(Z)$, and $\tilde{X} := (X_1'H, \tilde{X}_2')'$, $\tilde{X}_2 = (1, (S - \mathrm{E}S)')'$, where $X_1 = X_1(Z)$, e.g. $B(Z)$ or $(B(Z), S(Z), p(Z))'$. The terms X_1 are present in order to *reduce noise*.

We show that, as a complementary main identification result,

$$\beta_1 + \beta_2(S - \mathrm{E}S) = \mathsf{BLP}[s_0(Z) \mid S(Z)].$$

Theorem 2.2 (BLP 2)**.** *Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y has finite second moments and $\tilde{X} = (X_1 H, 1, (S - \mathrm{E}S))$ is such that $\mathrm{E}\tilde{X}\tilde{X}'$ is finite and full rank. Then, (β_1, β_2) defined in* (2.2) *solves the best linear predictor/approximation problem for the target $s_0(Z)$:*

$$(\beta_1, \beta_2)' = \arg\min_{b_1, b_2} \mathrm{E}[s_0(Z) - b_1 - b_2 S(Z)]^2,$$

in particular $\beta_1 = \mathrm{E}s_0(Z)$ and $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$.

The corresponding estimator is defined through the empirical analog:

$$Y_i H_i = \widehat{\mu}' X_{1i} H_i + \widehat{\beta}_1 + \widehat{\beta}_2(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\epsilon}_i, \quad \mathbb{E}_{N,M}\widehat{\epsilon}_i \tilde{X}_i = 0,$$

and the properties of this estimator, conditional on the auxiliary data, are well known and given in Lemma B.1.

2.2. **The Sorted Group ATE.** The target parameters are

$$\mathrm{E}[s_0(Z) \mid G],$$

where G is an indicator of group membership.

Comment 2.3. There are many possibilities for creating groups based upon ML tools applied to the auxiliary data. For example, one can group or cluster based upon predicted baseline response (as in the "endogenous stratification" analysis, [2]) or based upon actual predicted treatment effect S. We focus on the latter approach for defining groups, although our identification and inference ideas immediately apply to other ways of defining groups, and could be helpful in these contexts.

We build the groups to explain as much variation in $s_0(Z)$ as possible

$$G_k := \{S \in I_k\}, \quad k = 1, ..., K,$$

where $I_k = [\ell_{k-1}, \ell_k)$ are non-overlaping intervals that divide the support of S into regions $[\ell_{k-1}, \ell_k)$ with equal or unequal masses:

$$-\infty = \ell_0 < \ell_1 < \ldots < \ell_K = +\infty.$$

The parameters of interest are the Sorted Group Average Treatment Effects (GATES):

$$\mathrm{E}[s_0(Z) \mid G_k], \quad k = 1, \ldots, K.$$

Given the definition of groups, it is natural for us to impose the monotonicity restriction

$$\mathrm{E}[s_0(Z) \mid G_1] \leqslant ... \leqslant \mathrm{E}[s_0(Z) \mid G_K],$$

which holds asymptotically if $S(Z)$ is consistent for $s_0(Z)$ and the latter has an absolutely continuous distribution. Under the monotonicity condition, the estimates could be rearranged to obey the weak monotonicity condition, improving the precision of the estimator. The joint confidence intervals could also be improved by intersecting them with the set of monotone functions. Furthermore, as before, we can test for homogeneous effects, $s_0(Z) = s$, by testing whether,

$$\mathrm{E}[s_0(Z) \mid G_1] = ... = \mathrm{E}[s_0(Z) \mid G_K].$$

GATES: The First Strategy. Here we shall recover GATES parameters from the weighted linear projection equation:

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k \cdot (D - p(Z)) \cdot 1(G_k) + \nu, \quad \mathrm{E}[w(Z)\nu W] = 0, \tag{2.3}$$

for $B := B(Z)$, $S := S(Z)$, $W = (X_1', W_2')'$,

$$W_2 = (\{(D - p(Z))1(G_k)\}_{k=1}^K)'.$$

The presence of $D - p(Z)$ in the interaction $(D - p(Z))1(G_k)$ *orthogonalizes* this regressor relative to all other regressors that are functions of Z. The controls X_1, e.g. B, can be included to improve precision.

The second main identification result is that the projection coefficients γ_k are the GATES parameters:

$$\gamma = (\gamma_k)_{k=1}^K = (\mathrm{E}[s_0(Z) \mid G_k])_{k=1}^K.$$

Given the identification strategy, we can base the corresponding estimation strategy on the following empirical analogs:

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\gamma}' W_{2i} + \widehat{\nu}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[w(Z_i)\widehat{\nu}_i W_i] = 0. \tag{2.4}$$

The properties of this estimator, conditional on the auxilliary data, are well known and stated as a special case of Lemma B.1.

A formal statement appears below, together with a complementary result.

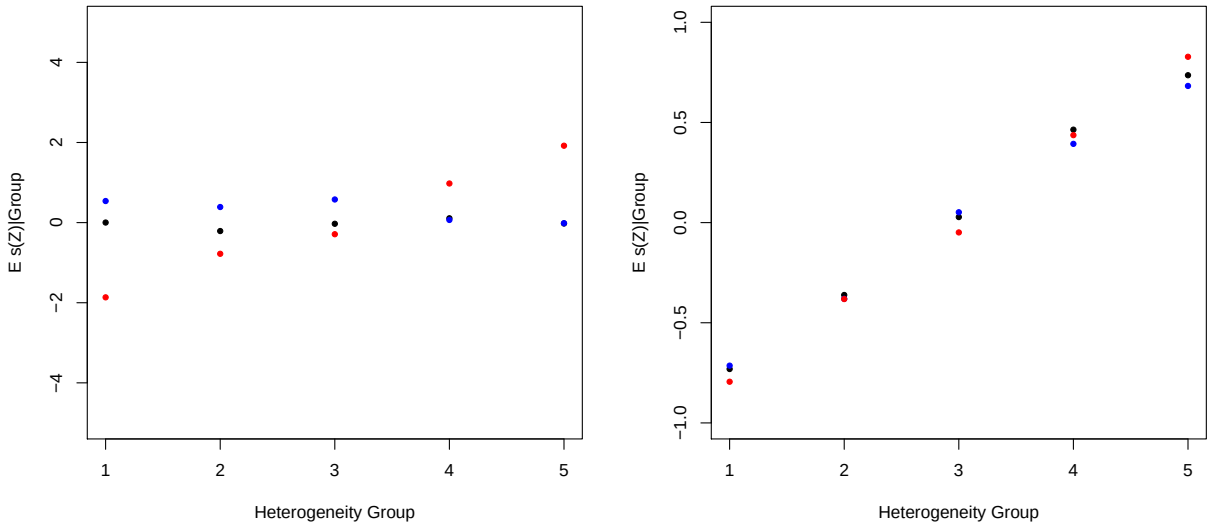

FIGURE 2. In the left panel we have the homogeneous CATE $s_0(Z) = 0$; in the right panel we have heterogeneous CATE $s_0(Z) = Z$. The proxy predictor $S(Z)$ for CATE is produced by the random forest, whose sorted averages by groups are shown as red dots, exhibiting large biases. These are the naive estimates. The true sorted group average treatment effects (GATES) $\mathrm{E}[s_0(Z) \mid G_k]$ are shown by black dots, and estimated GATES are shown by blue dots. The true and estimated GATES correct for the biases relative to the naive strategy shown in red. The estimated GATES shown by blue dots are always closer to the true GATEs shown by black dots than the naive estimates shown in red.

Figure 2 provides two examples using the same designs as in fig. 1. Post-processing the ML estimates again has stronger effect when there is no heterogeneity, but in both cases help bring the estimated GATES close to the true GATES.

GATES: The Second Strategy. Here we employ linear projections on Horvitz-Thompson transformed variables:

$$YH = \mu' X_1 H + \sum_{k=1}^{K} \gamma_k \cdot 1(G_k) + \nu, \quad \mathrm{E}[\nu \tilde{W}] = 0, \tag{2.5}$$

for $B := B(Z)$, $S := S(Z)$, $\tilde{W} = (X_1' H, \tilde{W}_2')$, $\tilde{W}_2' = (\{1(G_k)\}_{k=1}^K)$.

Again, we show that the projection parameters are GATES:

$$\gamma = (\gamma_k)_{k=1}^K = (\mathrm{E}[s_0(Z) \mid G_k])_{k=1}^K.$$

Given the identification strategy, we can base the corresponding estimation strategy on the following empirical analogs:

$$Y_i H_i = \widehat{\mu}' X_{1i} H_i + \widehat{\gamma}' \tilde{W}_{2i} + \widehat{\nu}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[\widehat{\nu}_i \tilde{W}_i] = 0. \tag{2.6}$$

The properties of this estimator, conditional on the auxiliary data, are well known and given in Lemma B.1. The resulting estimator has similar performance to the previous estimator, and under some conditions their first-order properties coincide.

The following is the formal statement of the identification result.

Theorem 2.3 (GATES)**.** *Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y has finite second moments and the W's and $\tilde{W}$ defined above are such that $\mathrm{E}WW'$ and $\mathrm{E}\tilde{W}\tilde{W}'$ are finite and have full rank. Consider $\gamma = (\gamma_k)_{k=1}^K$ defined by the weighted regression equation (2.3) or by the regression equation (2.5). These parameters defined in two different ways are equivalent and are equal to the expectation of $s_0(Z)$ conditional on the proxy group $\{S \in I_k\}$:*

$$\gamma_k = \mathrm{E}[s_0(Z) \mid G_k].$$

2.3. **Classification Analysis (CLAN).** When the BLP and GATES analyses reveal substantial heterogeneity, it is interesting to know the properties of the subpopulations that are most and least affected. Here we focus on the "least affected group" G_1 and "most affect group" G_K. Under the monotonicity assumption, it is reasonable that the first and the last groups are the most and least affected, where the labels "most" and "least" can be swapped depending on the context.

Let $g(Y, Z)$ be a vector of characteristics of an observational unit. The parameters of interest are the average characteristics of the most and least affected groups:

$$\delta_1 = \mathrm{E}[g(Y, Z) \mid G_1] \quad \text{and} \quad \delta_K = \mathrm{E}[g(Y, Z) \mid G_K].$$

The parameters δ_K and δ_1 are identified because they are averages of variables that are directly observed. We can compare δ_K and δ_1 to quantify differences between the most and least affected groups. We call this type of comparisons as classification analysis or CLAN.

3. "Variational" Estimation and Inference Methods

3.1. **Estimation and Inference: The Generic Targets.** Let θ denote a generic target parameter or functional, for example,

- $\theta = \beta_2$ is the heterogeneity predictor loading parameter;
- $\theta = \beta_1 + \beta_2(S(z) - \mathrm{E}S)$ is the "personalized" prediction of $s_0(z)$;
- $\theta = \gamma_k$ is the expectation of $s_0(Z)$ for the group G_k;
- $\theta = \gamma_K - \gamma_1$ is the difference in the expectation of $s_0(Z)$ between the most and least affected groups;

- $\theta = \delta_K - \delta_1$ is the difference in the expectation of the characteristics of the most and least impacted groups.

3.2. **Quantification of Uncertainty: Two Sources.** There are two principal sources of sampling uncertainty:

(I) Estimation uncertainty regarding the parameter θ, conditional on the data split;

(II) Uncertainty or "variation" induced by the data splitting.

Conditional on the data split, quantification of estimation uncertainty is standard. To account for uncertainty with respect to the data splitting, it makes sense to examine the robustness and variability of the estimates/confidence intervals with respect to different random splits. One of our goals is to develop methods, which we call "variational estimation and inference" (VEIN) methods, for quantifying this uncertainty, which can be of independent interest in many settings where the sample splitting is used.

Quantifying Source (I): Conditional Inference. We first recognize that the parameters implicitly depend on

$$\mathrm{Data}_A := \{(Y_i, D_i, X_i)\}_{i \in A},$$

the auxiliary sample, used to create the ML proxies $B = B_A$ and $S = S_A$. Here we make the dependence explicit: $\theta = \theta_A$.

All of the examples admit an estimator $\widehat{\theta}_A$ such that under mild assumptions,

$$\widehat{\theta}_A \mid \mathrm{Data}_A \sim_a N(\theta_A, \widehat{\sigma}_A^2),$$

in the sense that, as $|M| \to \infty$,

$$\mathbb{P}(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A) \leqslant z \mid \mathrm{Data}_A) \to_P \Phi(z).$$

Implicitly this requires the auxiliary data Data_A to be "sufficiently regular", and this should happen with high probability.

As a consequence, the confidence interval (CI)

$$[L_A, U_A] := [\widehat{\theta}_A \pm \Phi^{-1}(1 - \alpha/2)\widehat{\sigma}_A]$$

covers θ_A with approximate probability $1 - \alpha$:

$$\mathbb{P}[\theta_A \in [L_A, U_A] \mid \mathrm{Data}_A] = 1 - \alpha - o_P(1).$$

This leads to straighforward conditional inference, which does not account for the sample splitting uncertainty.

Quantifying Source (II): "Variational" Inference. Different partitions (A, M) of $\{1, ..., N\}$ yield different targets θ_A. Conditional on the data, we treat θ_A as a random variable, since (A, M) are random sets that form random partitions of $\{1, \ldots, N\}$ into samples of size n and $N - n$. Different partitions also yield different estimators $\widehat{\theta}_A$ and approximate distributions for these estimators. Hence we need a systematic way of treating the randomness in these estimators and their distributions.

Comment 3.1. In cases where the data sets are not large, it may be desirable to restrict attention to balanced partitions (A, M), where the proportion of treated units is equal to the designed propensity score.

We want to quantify the uncertainty induced by the random partitioning. Conditional on Data, the estimated $\widehat{\theta}_A$ is still a random variable, and the confidence band $[L_A, U_A]$ is a random set. For reporting purposes, we instead would like to report an estimator and confidence set, which are non-random conditional on the data.

Adjusted Point and Interval Estimators. Our proposal is as follows. As a point estimator, we shall report the median of $\widehat{\theta}_A$ as (A, M) vary (as random partitions):

$$\widehat{\theta} := \mathrm{Med}[\widehat{\theta}_A \mid \mathrm{Data}].$$

This estimator is more robust than the estimator based on a single split. To account for partition uncertainty, we propose to report the following confidence interval (CI) with the nominal confidence level $1 - 2\alpha$:

$$[l, u] := [\overline{\mathrm{Med}}[L_A \mid \mathrm{Data}], \underline{\mathrm{Med}}[U_A \mid \mathrm{Data}]].$$

Note that the price of splitting uncertainty is reflected in the discounting of the confidence level from $1 - \alpha$ to $1 - 2\alpha$. Alternatively, we can report the confidence interval based on inversion of a test based upon p-values, constructed below.

The above estimator and confidence set are non-random conditional on the data. The confidence set reflects the uncertainty created by the random partitioning of the data into the main and auxilliary data.

Comment 3.2. For a random variable X with law P_X we define

$$\underline{\mathrm{Med}}(X) := \inf\{x \in \mathbb{R} : \mathrm{P}_X(X \leqslant x) \geqslant 1/2\},$$

$$\overline{\mathrm{Med}}(X) := \sup\{x \in \mathbb{R} : \mathrm{P}_X(X \geqslant x) \geqslant 1/2\},$$

$$\mathrm{Med}(X) := (\underline{\mathrm{Med}}(X) + \overline{\mathrm{Med}}(X))/2.$$

Note that the lower median is the usual definition of the median. The upper median is the next distinct quantile of the random variable (or it is the usual median after reversing the order on $\mathbb{R}$). For example, when X is uniform on $\{1, 2, 3, 4\}$, then $\underline{\mathrm{Med}}(X) = 2$ and $\overline{\mathrm{Med}}(X) = 3$; and if X is

uniform on $\{1, 2, 3\}$, then $\overline{\text{Med}}(X) = \underline{\text{Med}}(X) = 2$. For continuous random variables the upper and lower medians defined above coincide. For discrete random variables they can differ, but the differences will be small for variables that are close to being continuous. ■

Suppose we are testing $H_0 : \theta_A = \theta_0$ against $H_1 : \theta_A < \theta_0$, conditional on the auxiliary data, then the p-value is given by

$$p_A = \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_0)).$$

The p-value for testing $H_0 : \theta_A = \theta_0$ against $H_1 : \theta_A > \theta_0$, is given by $p_A = 1 - \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_0))$.

Under the null hypothesis p_A is approximately distributed as the uniform variable, $p_A \sim U(0, 1)$, conditional on Data_A. Note that, conditional on Data, p_A still has randomness induced by random partitioning of the data, which we need to address.

Adjusted P-values. We say that testing the null hypothesis, based on the p-values p_A, that are random conditional on data, has significance level α if

$$\mathbb{P}(p_A \leqslant \alpha/2 \mid \text{Data}) \geqslant 1/2 \quad \text{or } p_{.5} = \underline{\text{Med}}(p_A \mid \text{Data}) \leqslant \alpha/2.$$

That is, for at least 50% of the random data splits, the realized p-value p_A falls below the level $\alpha/2$. Hence we can call $p = 2p_{.5}$ the *sample splitting-adjusted p-value*, and consider its small values as providing evidence against the null hypothesis.

Comment 3.3. Our construction of p-values builds upon the false-discovery-rate type adjustment ideas in [16, 42], though what we propose is radically simpler, and is minimalistic for our problem, whereas the idea of our confidence intervals below appears to be completely new. ■

The main idea behind this construction is simple: the p-values are distributed as marginal uniform variables $\{U_j\}_{j \in J}$, and hence obey the following property.

Lemma 3.1 (A Property of Uniform Variables)**.** *Consider M, the (usual, lower) median of a sequence $\{U_j\}_{j \in J}$ of uniformly distributed variables, $U_j \sim U(0, 1)$ for each $j \in J$, where variables are not necessarily independent. Then,*

$$\mathbb{P}(M \leqslant \alpha/2) \leqslant \alpha.$$

Proof. Let M denote the median of $\{U_j\}_{j \in J}$. Then $M \leqslant \alpha/2$ is equivalent to $|J|^{-1} \sum_{j \in J}[1(U_j \leqslant \alpha/2)] - 1/2 \geqslant 0$. So

$$\mathbb{P}[M \leqslant \alpha/2] = \mathbb{E}1\{|J|^{-1} \sum_{j \in J}[1(U_j \leqslant \alpha/2)] \geqslant 1/2\}.$$

By Markov inequality this is bounded by

$$2\mathbb{E}|J|^{-1} \sum_{j \in J}[1(U_j \leqslant \alpha/2)] \leqslant 2\mathbb{E}[1(U_j \leqslant \alpha/2)] \leqslant 2\alpha/2 = \alpha.$$

where the last inequality holds by the marginal uniformity. ■

Main Inference Result: Variational P-values and Confidence Intervals. We present a formal result on adjusted p-values using this condition:

PV. Suppose that $\mathcal{A}$ is a set of regular auxiliary data configurations such that for all $x \in [0,1]$, under the null hypothesis:

$$\sup_{P \in \mathcal{P}} |\mathbb{P}_P[p_A \leqslant x \mid \text{Data}_A \in \mathcal{A}] - x| \leqslant \delta = o(1),$$

and $\inf_{P \in \mathcal{P}} \mathbb{P}_P[\text{Data}_A \in \mathcal{A}] =: 1 - \gamma = 1 - o(1)$. In particular, suppose that this holds for the p-values

$$p_A = \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A)) \text{ and } p_A = 1 - \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A)).$$

Lemma B.1 shows that this condition is plausible for the least squares estimators defined in the previous section under mild conditions.

Theorem 3.1 (Uniform Validity of Variational P-Value)**.** *Under condition* PV *and the null hypothesis holding,*

$$\mathbb{P}_P(p_{.5} \leqslant \alpha/2) \leqslant \alpha + 2(\delta + \gamma) = \alpha + o(1),$$

uniformly in $P \in \mathcal{P}$.

In order to establish the properties of the confidence interval $[l, u]$, we first consider the properties of the related confidence interval, which is based on the inversion of the p-value based tests:

$$\text{CI} := \{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2,\ p_l(\theta) > \alpha/2\}, \tag{3.1}$$

for $\alpha < .25$, where, for $\widehat{\sigma}_A > 0$,

$$p_l(\theta) := \underline{\text{Med}}(1 - \Phi[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta)] \mid \text{Data}), \tag{3.2}$$

$$p_u(\theta) := \underline{\text{Med}}(\Phi[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta)] \mid \text{Data}). \tag{3.3}$$

The confidence interval CI has the following representation in terms of the medians of t-statistics implied by the proof Theorem 3.2 stated below:

$$\text{CI} = \left\{ \theta \in \mathbb{R} : \begin{array}{c} \overline{\text{Med}} \left[\frac{\theta - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \mid \text{Data} \right] < 0 \\ \underline{\text{Med}} \left[\frac{\theta - \widehat{\theta}_A}{\widehat{\sigma}_A} + \Phi^{-1}(1-\alpha/2) \mid \text{Data} \right] > 0 \end{array} \right\}. \tag{3.4}$$

This CI can be (slightly) tighter than $[l, u]$, while the latter is much simpler to construct.

The following theorem establishes that both confidence sets maintain the approximate confidence level $1 - 2\alpha$.

Theorem 3.2 (Uniform Validity of Variational Confidence Intervals)**.** CI *can be represented as (3.4) and* $\text{CI} \subseteq [l, u]$*, and under condition* PV*,*

$$\mathbb{P}_P(\theta_A \in \text{CI}) \geqslant 1 - 2\alpha - 2(\delta + \gamma) = 1 - 2\alpha - o(1),$$

uniformly in $P \in \mathcal{P}$.

4. Other Considerations and Extensions

1. Choosing the Best ML Method Targeting CATE in Stage 1. There are several options. The best ML method can be chosen using the auxiliary sample, based on either (A) the ability to predict YH using BH and S or (B) the ability to predict Y using B and $(D - p(Z))(S - \mathrm{E}(S))$ under the weight $w(Z)$ (as in the first type of strategies we developed in the first stage). To be specific, we can solve either of the following problems:

(A) minimize the errors in the prediction of YH on BH and S:

$$(B, S) = \arg \min_{B \in \mathcal{B}, S \in \mathcal{S}} \quad \sum_{i \in A} [Y_i H_i - B(Z_i) H_i - S(Z_i)]^2,$$

where $\mathcal{B}$ and $\mathcal{S}$ are parameter spaces for $z \mapsto B(z)$ and $z \mapsto S(z)$; or

(B) minimize the errors in the weighted prediction of Y on B and $(D - p(Z))(S - \mathrm{E}(S))$:

$$(B, S) = \arg \min_{B \in \mathcal{B}, S \in \mathcal{S}} \quad \sum_{i \in A} w(Z_i) [Y_i - B(Z_i) - (D_i - p(Z_i)) \{S(Z_i) - \bar{S}(Z_i)\}]^2,$$

where $\bar{S}(Z_i) = |A|^{-1} \sum_{i \in A} S(Z_i)$ and $\mathcal{B}$ and $\mathcal{S}$ are parameter spaces for $z \mapsto B(z)$ and $z \mapsto S(z)$.

This idea improves over simple but inefficient strategy of predicting YH just using S, which have been suggested before for causal inference. It also improves over the simple strategy that predicts Y using B and DS (which chooses the best predictor for $\mathrm{E}[Y \mid D, Z]$ in a given class but not necessarily the best predictor for CATE $s_0(Z)$). Note that this idea is new and is of major independent interest.

2. Choosing the Best ML Method BLP Targeting CATE in Stage 2. The best ML method can also be chosen in the main sample by maximizing

$$\Lambda := |\beta_2|^2 \mathrm{Var}(S(Z)) = \mathrm{Corr}^2(s_0(Z), S(Z)) \mathrm{Var}(s_0(Z)). \tag{4.1}$$

Maximizing Λ is equivalent to maximizing the correlation between the ML proxy predictor $S(Z)$ and the true score $s_0(Z)$, or equivalent to maximizing the R^2 in the regression of $s_0(Z)$ on $S(Z)$.

3. Choosing the Best ML Method GATES Targeting CATE in Stage 2. Analogously, for GATES the best ML method can also be chosen in the main sample by maximizing

$$\bar{\Lambda} = \mathrm{E} \left(\sum_{k=1}^{K} \gamma_k 1(S \in I_k) \right)^2 = \sum_{k=1}^{K} \gamma_k^2 \mathrm{P}(S \in I_k). \tag{4.2}$$

This is the part of variation $\mathrm{E} s_0^2(Z)$ of $s_0(z)$ explained by $\bar{S}(Z) = \sum_{k=1}^{K} \gamma_k 1(S(Z) \in I_k)$. Hence choosing the ML proxy $S(Z)$ to maximize $\bar{\Lambda}$ is equivalent to maximizing the R^2 in the regression

of $s_0(Z)$ on $\bar{S}(Z)$ (without a constant). If the groups $G_k = \{S \in I_k\}$ have equal size, namely $\mathrm{P}(S(Z) \in I_k) = 1/K$ for each $k = 1, ..., K$, then

$$\bar{\Lambda} = \frac{1}{K} \sum_{k=1}^{K} \gamma_k^2.$$

4. Stratified Splitting. The idea is to balance the proportions of treated and untreated in both A and M samples, so that the proportion of treated is equal to the experiment's propensity scores across strata. This formally requires us to replace the i.i.d. assumption by the i.n.i.d. assumption (independent but not identically distributed observations) when accounting for estimation uncertainty, conditional on the auxiliary sample. This makes the notation more complicated, but the results in Lemma B.1 still go through with notational modifications.

5. When Proxies have Little Variation. The analysis may generate proxy predictors S that have little variation, so we can think of them as "weak", which makes the parameter β_2 weakly identified. We can either add small noise to the proxies, which is called jittering, so that inference results go through, or we may switch to testing rather than estimation. For practical reasons, we prefer the jittering approach.

5. Further Potential Applications to Prediction and Causal Inference Problems

Our inference approach generalizes to any problem of the following sort.

Generalization. Suppose we can construct an *unbiased signal* $\tilde{Y}$ such that

$$\mathrm{E}[\tilde{Y} \mid Z] = s_0(Z),$$

where $s_0(Z)$ is now a generic target function. Let $S(Z)$ denote an ML proxy for $s_0(Z)$. Then, using previous arguments, we immediately can generate the following conclusions:

(1) The projection of $\tilde{Y}$ on the ML proxy $S(Z)$ identifies the BLP of $s_0(Z)$ using $S(Z)$.
(2) The grouped averages of the target (GAT) $\mathrm{E}[s_0(Z) \mid G_k]$ are identified by $\mathrm{E}[\tilde{Y} \mid G_k]$.
(3) Using ML tools we can train proxy predictors $S(Z)$ to predict $\tilde{Y}$ in auxiliary samples.
(4) We post-process $S(Z)$ in the main sample, by estimating the BLP and GATs.
(5) We apply variational inference on functionals of the BLP and GATs.

The noise reduction strategies, like the ones we used on the context of H-transformed outcomes, can be useful in these cases, but their construction could depend on the context.

Example 1. Forecasting or Predicting Regression Functions using ML proxies. This is the most common type of the problem arising in forecasting. Here the target is the best predictor of Y using Z, namely $s_0(Z) = \mathrm{E}[Y \mid Z]$, and $\tilde{Y} = Y$ trivially serves as the unbiased signal. The interesting part

here is the use of variational inference tools developed in this paper for constructing confidence intervals for the predicted values produced by the estimated BLP of $s_0(Z)$ using $S(Z)$.

Example 2. Predicting Structural Derivatives using ML proxies. Suppose we are interested in best predicting the partial derivative $s_0(z)$, where $s_0(z) = \partial_x g(x,z)$, where $g(x,z) = \mathrm{E}[Y \mid X = x, Z = z]$. In the context of demand analysis, Y is the log of individual demand, X is the log-price of a product, and Z includes prices of other products and characteristics of individuals. Then the unbiased signal is given by $\tilde{Y} = Y[\partial_x \log p(X \mid Z)]$, where $p(x \mid z)$ is the conditional density function of x given z. That is, $\mathrm{E}[\tilde{Y} \mid Z] = s_0(Z)$ under mild conditions on the density using the integration by parts formula.

6. Empirical Applications and Implementation Algorithms

We consider two empirical applications. The first application is an RCT conducted in Morocco, which investigates the effects of microfinance access on several outcomes. The second application is an observational study, which analyzes the gender wage gap using the U.S. Current Population Survey (CPS). While the results of the second study should not be interpreted as causal, we aim to obtain the best approximation to the causal effects of gender-based discrimination on wages using observational data, by flexibly conditioning on many controls using ML methods. We conclude this section by providing the implementation algorithms for both examples.

6.1. **Heterogeneity in the Effect of Microcredit Availability.**

We analyze a randomized experiment designed to evaluate the impact of microcredit availability on borrowing and self-employment activities, which was previously studied in [24]. The experiment was conducted in 162 villages in Morocco, divided into 81 pairs of villages with similar observable characteristics (number of households, accessibility to the center of the community, existing infrastructure, type of activities carried out by the households, and type of agriculture activities). One of the villages in each pair was randomly assigned to treatment and the other to control. Between 2006 and 2007 a microfinance institution started operating in the treated villages. Two years after the intervention an endline household survey was conducted with 5,551 households, which constitute our sample. There was no other microcredit penetration in these villages, before and for the duration of the study. Therefore, we interpret the treatment as the availability of microcredit.

Recent randomized evaluations of access to microcredit at the community level have found limited impacts of microcredit[9]. Despite evidence that access to microfinance leads to an increase in borrowing ([4], [10], [48]) and business creation or expansion ([4], [7], [10], [48]), most studies have found that this does not translate into an increase in economic outcomes such as profit, income, labor supply and consumption ([4], [10], [24]). Moreover, there is also no evidence of substantial gains in human development outcomes, such as education and health ([10],[48]). Studies which

[9]See [11] for a summary of the recent literature

estimate the impact of microfinance by randomizing microcredit at the individual level confirm these findings ([8], [36], [37]).

One question that remains elusive is whether the lack of evidence on the average effects masks heterogeneity, in which there are potential winners and losers of the microcredit expansion. Understanding this heterogeneity can have important implications for evaluating the welfare effects of microcredit, designing policies and targeting the groups that would benefit from microfinance. Indeed, the idea that there might be heterogeneity in the impact of microcredit has been a common theme among RCTs evaluating microfinance programs. Having found mostly positive but insignificant coefficients, the papers cited above attempt to explore heterogeneous treatment effects, mostly using quantile treatment effects. For profits, most studies seem to find positive impact at the higher quantiles (and in the data set we study here, [24] actually find *negative* impacts at lower level). Using Bayesian hierarchical methods to aggregate the evidence across studies, Meager (2017) cautions that these results on quantiles may not be generalizable: the profit variables seems to have too much noise to lend itself to quantile estimation.

A number of recent papers also consider heterogeneous treatment effects by studying the effect of microfinance on a subpopulation. In a follow-up study of [10], [9] investigates whether the heterogeneity is persistent six years after the microfinance was introduced. They find that credit has a much bigger impact on the business outcomes of those who started a business before microfinance entered than of those without prior businesses. Using the same dataset as in this application [24] classify households into three categories in term of their probability to borrow before the intervention and find that microcredit access has a significant impact on investment and profit, but still no impact on income and consumption among those who are most likely to borrow. It is worth noting that the original strategy for this study was to construct groups which, *ex ante* had different probability to borrow, in order to separately estimate direct effect of microcredit on those most likely to borrow, and indirect on those very unlikely to borrow. The researchers had initially collected a short survey on the entire village, and had then tried to predict the probability to borrow using a model fitted in a first group of villages. But the model proved to have low predictive power in the other villages, and in their paper then end up predicting the probability to borrow ex-post. In the paper, they worry that this ex-post classification leads them to over-fit.

The identification strategy developed in this paper provides several advantages in studying heterogeneity in the treatment effects of microfinance. First, contrary to the literature, which relies on ad hoc subgroup analysis across a few baseline characteristics, we are agnostic about the source of heterogeneity. While the variable "had a prior business" has proven to be a robust and generalizable predictors of differences in treatment effect (Meager, 2017) and could therefore be pre-specified in future pre-analysis plans, we have little idea about what else predict heterogeneity. Second, our approach is valid in high dimensional settings, allowing us to include a rich set of characteristics in an unspecified functional form. Finally, using the CLAN estimation we are able to identify the

characteristics of the most and least affected subpopulations, which could be an important input for a welfare analysis or targeting households who are likely to benefit from access to microfinance.

We focus on heterogeneity in treatment effects on four household outcome variables, Y: the amount of money borrowed, the output from self-employment activities, profit from self-employment activities, and monthly consumption. The treatment variable, D, is an indicator for the household residing in a treated village. The covariates, Z, include some baseline household characteristics such as number of members, number of adults, head age, indicators for households doing animal husbandry, doing other non-agricultural activity, having an outstanding loan over the past 12 months, household spouse responded to the survey, another household member (excluding the HH head) responded to the survey, and 81 village pair fixed effects (these are the variables that are available for all households). We also include indicators for missing observation at baseline as controls. Table 6 shows some descriptive statistics for the variables used in the analysis (all monetary variables are expressed in Moroccan Dirams, or MAD). Treated and control households have similar characteristics and the unconditional average treatment effect on loans, output, profit and consumption are respectively 1,128, 5,237, 1,844 and -31.

TABLE 1. Descriptive Statistics of Households

	All	Treated	Control
Outcome Variables			
Total Amount of Loans	2,359	2,930	1,802
Total output from self-employment activities (past 12 months)	32,499	35,148	29,911
Total profit from self-employment activities (past 12 months)	10,102	11,035	9,191
Total monthly consumption	3,012	2,996	3,027
Baseline Covariates			
Number of Household Members	3.879	3.872	3.886
Number of Members 16 Years Old or Older	2.604	2.601	2.607
Head Age	35.976	35.937	36.014
Declared Animal Husbandry Self-employment Activity	0.415	0.426	0.404
Declared Non-agricultural Self-employment Activity	0.146	0.129	0.164
Borrowed from Any Source	0.210	0.224	0.196
Spouse of Head Responded to Self-employment Section	0.067	0.074	0.061
Member Responded to Self-employment Section	0.044	0.048	0.041

We implement our methods using the algorithm described in Section 6.3. By design the propensity score $p(Z_i) = 1/2$ for all the households. Table 2 compares the four ML methods for producing the proxy predictors $S(Z_i)$ considered in Stage 1. We find that the Random Forest and Elastic Net

TABLE 2. Comparison of ML Methods: Microfinance Availability

	Elastic Net	Boosting	Neural Network	Random Forest
Amount of Loans				
Best BLP (Λ)	2,808,960	2,086,256	2,464,276	2,706,767
Best GATES ($\bar{\Lambda}$)	875	421	630	1253
Output				
Best BLP (Λ)	143,620,159	78,127,786	76,945,520	133,825,537
Best GATES ($\bar{\Lambda}$)	8,697	2,823	4,219	5,229
Profit				
Best BLP (Λ)	32,307,828	17,105,855	20,404,000	39,286,050
Best GATES ($\bar{\Lambda}$)	4,595	2,319	1,847	4,478
Consumption				
Best BLP (Λ)	44,583	26,387	34,019	36,927
Best GATES ($\bar{\Lambda}$)	101	101	96	109

Notes: Medians over 100 splits in half.

outperform the Boosted Tree and Neural Network across all outcome variables for both metrics. Accordingly, we focus on these two methods for the rest of the analysis.[10]

Table 3 presents results of the BLP of CATE using the ML proxies $S(Z)$ for the four outcome variables. We report estimates of the coefficients β_1 and β_2, which correspond to the ATE and heterogeneity loading (HET) parameters in the BLP, respectively. In parentheses, we report confidence intervals adjusted for variability across the sample splits using the median method; and in brackets, we report adjusted p-values. The estimated ATEs of microfinance availability are consistent with the findings of [24] and are similar to the unconditional ATE, as expected by virtue of the randomization. The ATE on the amount of loans and output are positive and statistically significant at least at the 10% level with both ML methods. Microfinance availability does not have a significant impact on profit and consumption.

Turning to the heterogeneity results, we reject the hypothesis that HET is zero at the 10% level for the amount of loans, output and profit with the elastic net method, suggesting the presence of heterogeneity in the effect of microfinance availability. The results are consistent across both ML methods except for output, for which HET coefficient on the Random Forest proxy is not significantly different from zero at the 10%. Finally, the BLP analysis does not reveal any significant heterogeneity in the effect on consumption. Overall, these results suggest that microfinance availability has heterogenous impacts on business-related outcomes that does not seem to translate into

[10]The results obtained using Boosted Tree and Neural Network are similar to the results reported, but they are slightly less precise. These results are not reported but are available from the authors upon requests.

TABLE 3. BLP of Microfinance Availability

	Elastic Net		Random Forest	
	ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
Amount of Loans	1,163	0.238	1,185	0.375
	(544,1736)	(0.021,0.448)	(561,1771)	(0.028,0.774)
	[0.000]	[0.060]	[0.000]	[0.069]
Output	5,095	0.262	5,027	0.192
	(232,10033)	(0.085,0.433)	(-89,10194)	(-0.100,0.508)
	[0.079]	[0.008]	[0.109]	[0.391]
Profit	1,553	0.244	1,603	0.279
	(-1344,4389)	(0.079,0.416)	(-1276,4536)	(0.046,0.518)
	[0.584]	[0.008]	[0.521]	[0.039]
Consumption	-59.1	0.157	-58.6	0.196
	(-161.5,44.2)	(-0.058,0.385)	(-166.6, 43.3)	(-0.160,0.574)
	[0.514]	[0.278]	[0.508]	[0.553]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.

P-values for the hypothesis that the parameter is equal to zero in brackets.

a detectable contemporaneous effect on the standard of living as represented by consumption, even for the most positively affected households.

We next estimate the GATES. We divide the households into $K = 5$ groups based on the quintiles of the ML proxy predictor $S(Z)$ and estimate the average effect for each group. Figures 3-6 present the estimated GATES coefficients $\gamma_1 - \gamma_5$ along with joint confidence bands. We also report the ATE and its confidence interval that were obtained in the BLP analysis for comparison. The GATES provide a richer understanding of the heterogeneity. In particular, the figures reveal that there are groups of winners, the most affected groups, for which the GATES on amount of loans, output and profit are significantly different from zero. These groups are likely to drive the heterogeneity in the treatment effect that we find in the BLP analysis. We further investigate the GATES by comparing the most and least affected groups in Table 4. Here we find that the difference of GATES of these two groups is significantly different from zero at least at the 10% level on amount of loans, level on output and profit, whereas we fail to reject the hypothesis that this difference is zero at conventional levels on consumption. Looking at the least affected group, it is reassuring to see that we have no evidence of negative impact on profit and income, mitigating the concerns that there are adversely affected households. However, there is negative and insignificant effect on consumption for the same group. A possible explanation for this result is that investment is lumpy and some households cut back consumption to increase investment. All the results for the GATES are fairly robust to the ML method.

TABLE 4. GATES of 20% Most and Least Affected Groups

	Elastic Net			Random Forest		
	20% Most (γ_5)	20% Least (γ_1)	Difference ($\gamma_5 - \gamma_1$)	20% Most (γ_5	20% Least (γ_1)	Difference ($\gamma_5 - \gamma_1$)
Amount of Loans	2,677	-197	2,995	2,870	94.707	2,814
	(1298,4076)	(-1835,1307)	(945,5103)	(1149,4587)	(-1663,1723)	(503,5193)
	[0.000]	[1.000]	[0.008]	[0.002]	[1.000]	[0.032]
Output	22,367	-3,039	25,088	21,606	626	21,035
	(7678,36920)	(-12546,6535)	(7028,42698)	(5862,38022)	(-11871,13529)	(125,43170)
	[0.007]	[1.000]	[0.015]	[0.015]	[1.000]	[0.097]
Profit	10,644	-1,152.242	11,768	11,540	-2,031	14,037
	(2146,19096)	(-7250,4952)	(1077,22422)	(2965,20955.576)	(-8721,4796)	(2459,25833)
	[0.028]	[1.000]	[0.061]	[0.014]	[1.000]	[0.037]
Consumption	66.4	-333	383	62	-300	332
	(-166.2,289.8)	(-695.6,23.2)	(-38.0,805.6)	(-271,346)	(-683,66)	(-196,835)
	[1.000]	[0.140]	[0.152]	[1.000]	[0.228]	[0.429]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

We conclude by looking at the average characteristics of the most and least affected groups to understand what generates heterogeneity in the treatment effects. We omit the results for consumption as we do not detect heterogeneity for this outcome. We focus on three characteristics in this analysis: Age of household head, non-agricultural self-employment activity and whether the household borrowed from any source. Table 5 reports the CLAN for the 10% least and most affected groups defined by the deciles of the CATE proxy $S(Z)$ as well as the difference between the two. We find households with young heads, non-agriculture self-employment and that borrowed less from any source are likely to borrow more from the microfinance institution, suggesting that formal loans are substitutes rather than complements. For output only the average self-employment sector indicator is found to be significantly different between the most and least affected groups. Finally, the estimates for profit are similar to the estimates of amount of loans. It is important to note that the significance of some of these differences is sensitive to the ML method used to generate the proxy predictors, while the sign of the differences is robust.

6.2. **Gender Wage Gap in the U.S..** We consider an application to the gender wage gap using data from the U.S. March Supplement of the Current Population Survey (CPS) in 2015. We select white, non-hispanic individuals who are aged 25 to 64 years and work more than 35 hours per week during at least 50 weeks of the year. We exclude self-employed workers; individuals living in group quarters; individuals in the military, agricultural or private household sectors; individuals with inconsistent reports on earnings and employment status; individuals with allocated or missing

TABLE 5. CLAN of Microfinance Availability

	Elastic Net			Random Forest		
	10% Most (δ_{10})	10% Least (δ_1)	Difference ($\delta_{10} - \delta_1$)	10% Most (δ_{10})	10% Least (δ_1)	Difference ($\delta_{10} - \delta_1$)
Amount of Loans						
Head Age	29.3	35.2	-6.6	23.0	33.8	-10.4
	(26.3,32.4)	(32.2,38.2)	(-10.9,-2.4)	(19.8,26.2)	(30.5,36.9)	(-14.9,-6.0)
	-	-	[0.004]	-	-	[0.000]
Non-agricultural self-emp.	0.199	0.068	0.123	0.134	0.118	0.022
	(0.159,0.238)	(0.030,0.108)	(0.069,0.178)	(0.096,0.173)	(0.076,0.156)	(-0.033,0.075)
	-	-	[0.000]	-	-	[0.875]
Borrowed from Any Source	0.144	0.169	-0.038	0.109	0.217	-0.107
	(0.099,0.189)	(0.124,0.212)	(-0.101,0.025)	(0.064,0.153)	(0.175,0.262)	(-0.164,-0.050)
	-	-	[0.448]	-	-	[0.001]
Output						
Head Age	36.280	36.708	-0.896	29.090	30.831	-1.925
	(33.4,39.1)	(33.6,39.6)	(-5.242,3.432)	(25.8,32.3)	(27.5,34.1)	(-6.648,2.799)
	-	-	[1.000]	-	-	[0.849]
Non-agricultural self-emp.	0.275	0.050	0.226	0.215	0.088	0.130
	(0.233,0.315)	(0.007,0.093)	(0.169,0.285)	(0.172,0.257)	(0.045,0.129)	(0.070,0.190)
	-	-	[0.000]	-	-	[0.000]
Borrowed from Any Source	0.193	0.215	-0.033	0.165	0.189	-0.024
	(0.142,0.241)	(0.167,0.262)	(-0.102,0.034)	(0.121,0.208)	(0.146,0.234)	(-0.086,0.039)
	-	-	[0.687]	-	-	[0.895]
Profit						
Head Age	34.1	40.4	-6.5	29.2	33.7	-5.8
	(31.2,37.0)	(37.5,43.4)	(-10.7,-2.5)	(25.7,32.6)	(30.390,37.108)	(-10.566,-1.217)
	-	-	[0.003]	-	-	[0.029]
Non-agricultural self-emp.	0.181	0.108	0.082	0.153	0.099	0.051
	(0.140,0.222)	(0.068,0.149)	(0.022,0.138)	(0.113,0.192)	(0.058,0.139)	(-0.003,0.105)
	-	-	[0.014]	-	-	[0.129]
Borrowed from Any Source	0.180	0.257	-0.091	0.144	0.162	-0.032
	(0.130,0.230)	(0.207,0.307)	(-0.160,-0.022)	(0.098,0.190)	(0.122,0.206)	(-0.095,0.029)
	-	-	[0.020]	-	-	[0.578]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

information in any of the variables used in the analysis; and individuals with hourly wage rate below $\$3$. The resulting sample consists of $32,523$ workers including $18,137$ men and $14,382$ of women.

We will estimate the BLP, GATES and CLAN of the conditional gender wage gap, i.e., the difference in the average wage between female and male workers with the same observable characteristics; see [18] for a recent survey on this topic. The outcome variable Y is the logarithm of the hourly wage rate constructed as the ratio of the annual earnings to the total number of hours worked, which is constructed in turn as the product of number of weeks worked and the usual

number of hours worked per week. The treatment D is an indicator for female worker, and the observable characteristics Z include 5 marital status indicators (widowed, divorced, separated, never married, and married); 5 educational attainment indicators (less than high school graduate, high school graduate, some college, college graduate, and advanced degree); 4 region indicators (midwest, south, west, and northeast); potential experience constructed as the maximum of age minus years of schooling minus 7 and zero, i.e., $experience = \max(age - education - 7, 0)$; 5 occupation indicators (management, professional and related; service; sales and office; natural resources, construction and maintenance; and production, transportation and material moving); and 12 industry indicators (mining, quarrying, and oil and gas extraction; construction; manufacturing; wholesale and retail trade; transportation and utilities; information; financial services; professional and business services; education and health services; leisure and hospitality; other services; and public administration).[11]

TABLE 6. Descriptive Statistics of Workers

	All	Women	Men		All	Women	Men
Log wage	3.16	3.02	3.26	O.manager	0.49	0.56	0.43
Female	0.44	1.00	0.00	O.service	0.09	0.10	0.09
MS.married	0.70	0.65	0.73	O.sales	0.22	0.30	0.16
MS.widowed	0.01	0.02	0.01	O.construction	0.09	0.01	0.15
MS.separated	0.02	0.02	0.01	O.production	0.11	0.04	0.17
MS.divorced	0.12	0.15	0.09	I.minery	0.03	0.01	0.04
MS.Nevermarried	0.16	0.16	0.16	I.construction	0.06	0.01	0.09
E.lhs	0.02	0.01	0.03	I.manufacture	0.13	0.08	0.18
E.hsg	0.25	0.21	0.28	I.retail	0.13	0.11	0.14
E.sc	0.28	0.30	0.27	I.transport	0.04	0.02	0.06
E.cg	0.28	0.29	0.27	I.information	0.02	0.02	0.03
E.ad	0.17	0.19	0.15	I.finance	0.08	0.10	0.07
R.northeast	0.26	0.26	0.26	I.professional	0.11	0.10	0.12
R.midwest	0.27	0.28	0.27	I.education	0.25	0.41	0.12
R.south	0.33	0.34	0.33	I.leisure	0.04	0.05	0.04
R.west	0.22	0.21	0.23	I.services	0.03	0.03	0.04
Experience	21.21	21.03	21.35	I.public	0.07	0.07	0.08

Source: March Supplement CPS 2015

Table 6 reports sample means of the variables used in the analysis. Working women are more highly educated than working men, have about the same potential experience, and are less likely to be married and more likely to be divorced. They work relatively more often in managerial and sales occupations and in the industries providing education and health services. Working men are relatively more likely to work in construction and production occupations within non-service industries. The unconditional gender wage gap is 24%.

[11]The sample selection and variable construction is the same as in [20].

In this application we do not observe the propensity scores. We estimate them from a large sample of 129,983 workers from the U.S. March Supplement of the CPS pooling the years 2012 to 2017, excluding 2015, as described in Section 6.3. We ignore the sampling error in the estimated propensity scores for the rest of the analysis because the sample size to estimate these scores is much larger than the sample size to estimate the parameters of interest.

TABLE 7. Comparison of ML Methods

	Elastic Net	Boosting	Neural Network	Random Forest
Best BLP (Λ)	0.084	0.079	0.076	0.082
Best GATES ($\bar{\Lambda}$)	0.060	0.058	0.058	0.059

Notes: Medians over 1,000 splits in half.

Table 7 compares the ML methods. In this case the elastic net comes out as the winner both based on BLP and GATES targeting of CATE, followed closely by the random forest. As in the previous section, we focus on these two ML methods for the rest of the analysis. Table 8 shows the results for the BLP coefficients. The average conditional gender wage gap of -22% comes close to the unconditional gap, but the slope of the BLP uncovers substantial heterogeneity across workers. The estimates and joint confidence intervals for the GATES in Figure 7 confirm this finding. Here we divide the sample in $K = 5$ groups defined by the quintiles of the CATE proxy $S(Z)$ and show that average conditional gender wage gap ranges from about -33% to -9% across the groups. Table 9 compares the averages of the gender wage gap in the first (most affected) and fifth (least affected) groups. We reject that the average gaps for the 20% most and least affected are equal at any conventional level. The results for the BLP and GATES are robust to the ML method used in Stage 1.

TABLE 8. BLP of Gender Wage Gap

Elastic Net		Random Forest	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
-0.229	0.726	-0.228	0.552
(-0.251,-0.207)	(0.540,0.912)	(-0.249,-0.207)	(0.409,0.694)
[0.000]	[0.000]	[0.000]	[0.000]

Notes: Medians over 1,000 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

Table 10 reports the results of a classification analysis where we divide the workers in $K = 10$ groups defined by the deciles of the CATE proxy $S(Z)$ and compare the average characteristics of

TABLE 9. GATES of Gender Wage Gap

	20% Least Affected (γ_5)	20% Most Affected (γ_1)	Difference ($\gamma_5 - \gamma_1$)
Elastic Net	-0.088	-0.328	0.240
	(-0.136,-0.040)	(-0.377,-0.279)	(0.171,0.308)
	[0.001]	[0.000]	[0.000]
Random Forest	-0.095	-0.329	0.234
	(-0.142,-0.048)	(-0.377,-0.280)	(0.166,0.301)
	[0.000]	[0.000]	[0.000]

Notes: Medians over 1,000 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

the most and least affected group. This analysis reveals another dimension of the heterogeneity in the gender wage gap by showing that the 10% most affected workers are more likely to have either of the following characteristics relative to the 10% less affected workers on average: high wages, high experience, not advanced degree, married, not divorced or never married. These differences are robust to the two ML method. The elastic net and random forest, however, disagree in the comparison of the averages of some of the education variables such as high school graduate and some college. Due to this sensitivity, we recommend reporting results from at least the two best ML methods. Overall, the heterogeneity patterns that we find are consistent with both glass ceiling theories for the gender wage gap [3], and with heterogenous individual preferences where young, single, and highly educated women are more career-oriented. [20] found similar patterns applying semiparametric regression methods to the same data.

6.3. **Implementation Algorithm.** In this section we describe an algorithm based on the first identification strategy and provide some specific implementation details for the two empirical examples.

Algorithm 1 (**Inference Algorithm**). The inputs are given by the data on units $i \in [N] = \{1, ..., N\}$.

Step 0. Fix the number of splits S and the significance level α, e.g. $S = 100$ and $\alpha = 0.05$.

Step 1. Compute the propensity scores $p(Z_i)$ for $i \in [N]$.

Step 2. Consider S splits in half of the indices $i \in \{1, ..., N\}$ into the main sample, M, and the auxiliary sample, A. Over each split $s = 1, .., S$, apply the following steps:

a. Tune and train each ML method separately to learn $B(\cdot)$ and $S(\cdot)$ using A. For each $i \in M$, compute the predicted baseline effect $B(Z_i)$ and predicted treatment effect $S(Z_i)$. If there is zero variation in $B(Z_i)$ and $S(Z_i)$ add Gaussian noise with a variance of 0.1 to the proxies.
b. Estimate the BLP parameters by weighted OLS in M, i.e.,

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\beta}_1(D_i - p(Z_i)) + \widehat{\beta}_2(D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\epsilon}_i, \;\; i \in M$$

TABLE 10. CLAN of Gender Wage Gap

	Elastic Net			Random Forest		
	10% Least (δ_{10})	10% Most (δ_1)	Difference ($\delta_{10}-\delta_1$)	10% Least (δ_{10})	10% Most (δ_1)	Difference ($\delta_{10}-\delta_1$)
Log Wage	2.949	3.436	-0.484	3.074	3.233	-0.157
	(2.918,2.981)	(3.404,3.467)	(-0.528,-0.440)	(3.043,3.106)	(3.201,3.264)	(-0.201,-0.113)
	-	-	[0.000]	-	-	[0.000]
Experience	12.902	25.664	-12.778	14.176	25.109	-10.893
	(12.386,13.419)	(25.14,26.187)	(-13.513,-12.055)	(13.666,14.687)	(24.602,25.622)	(-11.615,-10.176)
	-	-	[0.000]	-	-	[0.000]
Less than H. School	0.039	0.023	0.017	0.025	0.039	-0.013
	(0.030,0.048)	(0.014,0.032)	(0.004,0.028)	(0.016,0.034)	(0.030,0.047)	(-0.025,-0.001)
	-	-	[0.016]	-	-	[0.078]
High School	0.246	0.167	0.080	0.162	0.268	-0.106
	(0.226,0.266)	(0.147,0.187)	(0.052,0.109)	(0.141,0.182)	(0.248,0.288)	(-0.134,-0.078)
	-	-	[0.000]	-	-	[0.000]
Some College	0.184	0.094	0.089	0.186	0.238	-0.050
	(0.166,0.201)	(0.076,0.110)	(0.066,0.112)	(0.166,0.206)	(0.217,0.258)	(-0.079,-0.021)
	-	-	[0.000]	-	-	[0.001]
College Graduate	0.326	0.570	-0.235	0.361	0.354	0.010
	(0.303,0.35)	(0.546,0.594)	(-0.269,-0.202)	(0.337,0.384)	(0.33,0.377)	(-0.023,0.043)
	-	-	[0.000]	-	-	[1.000]
Advanced Degree	0.201	0.135	0.066	0.259	0.096	0.164
	(0.182,0.22)	(0.116,0.154)	(0.04,0.091)	(0.241,0.278)	(0.078,0.115)	(0.137,0.19)
	-	-	[0.000]	-	-	[0.000]
Married	0.101	0.953	-0.851	0.359	0.861	-0.499
	(0.088,0.114)	(0.94,0.966)	(-0.869,-0.832)	(0.338,0.38)	(0.84,0.881)	(-0.529,-0.47)
	-	-	[0.000]	-	-	[0.000]
Widowed	0.010	0.008	0.003	0.008	0.011	-0.002
	(0.005,0.015)	(0.003,0.012)	(-0.004,0.009)	(0.004,0.013)	(0.006,0.016)	(-0.009,0.005)
	-	-	[0.728]	-	-	[0.981]
Divorced	0.092	0.019	0.072	0.095	0.083	0.012
	(0.08,0.103)	(0.008,0.03)	(0.056,0.088)	(0.081,0.109)	(0.069,0.097)	(-0.008,0.031)
	-	-	[0.000]	-	-	[0.478]
Separated	0.028	0.010	0.017	0.013	0.021	-0.008
	(0.021,0.035)	(0.003,0.017)	(0.008,0.027)	(0.006,0.019)	(0.014,0.027)	(-0.017,0.001)
	-	-	[0.001]	-	-	[0.170]
Nevermarried	0.764	0.003	0.76	0.523	0.022	0.500
	(0.749,0.779)	(-0.011,0.018)	(0.738,0.781)	(0.505,0.541)	(0.004,0.04)	(0.474,0.526)
	-	-	[0.000]	-	-	[0.000]

Notes: Medians over 1,000 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

such that $\mathbb{E}_{N,M}[w(Z_i)\widehat{\epsilon}_i X_i] = 0$ for $X_i = [X_{1i}', D_i - p(Z_i), (D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i)]'$, where $w(Z_i) = \{p(Z_i)(1-p(Z_i))\}^{-1}$ and X_{1i} includes a constant, $B(Z_i)$ and $S(Z_i)$.

c. Estimate the GATES parameters by weighted OLS in M, i.e.,

$$Y_i = \widehat{\alpha}' X_{1i} + \sum_{k=1}^{K} \widehat{\gamma}_k \cdot (D_i - p(Z_i)) \cdot 1(S_i \in I_k) + \widehat{\nu}_i, \;\; i \in M,$$

such that $\mathbb{E}_{N,M}[w(Z_i)\widehat{\nu}_i W_i] = 0$ for $W_i = [X_{i1}', \{(D_i - p(Z_i))1(S_i \in I_k)\}_{k=1}^K]'$, where $w(Z_i) = \{p(Z_i)(1-p(Z_i))\}^{-1}$, X_{1i} includes a constant, $B(Z_i)$ and $S(Z_i)$, $I_k = [\ell_{k-1}, \ell_k)$, and ℓ_k is the (k/K)-quantile of $\{S_i\}_{i \in M}$.

d. Estimate the CLAN parameters in M by

$$\widehat{\delta}_1 = \mathbb{E}_{N,M}[g(Y_i, Z_i) \mid S_i \in I_1] \quad \text{and} \quad \widehat{\delta}_K = \mathbb{E}_{N,M}[g(Y_i, Z_i) \mid S_i \in I_K],$$

where $I_k = [\ell_{k-1}, \ell_k)$ and ℓ_k is the (k/K)-quantile of $\{S_i\}_{i \in M}$.

e. Compute the two performance measures for the ML methods

$$\widehat{\Lambda} = |\widehat{\beta}_2|^2 \widehat{\mathrm{Var}}(S(Z)) \qquad \widehat{\bar{\Lambda}} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\gamma}_k^2.$$

Step 3: Choose the best ML methods based on the medians of $\widehat{\Lambda}$ and $\widehat{\bar{\Lambda}}$ over the splits.

Step 4: Compute the estimates, $(1-\alpha)$-level conditional confidence intervals and conditional p-values for all the parameters of interest. Monotonize the confidence intervals if needed. For example, construct a $(1-\alpha)$ joint confidence interval for the GATES as

$$\{\widehat{\gamma}_k \pm \widehat{c}(1-\alpha)\widehat{\sigma}_k, \;\; k = 1, \ldots, K\}, \tag{6.1}$$

where $\widehat{c}(1-\alpha)$ is a consistent estimator of the $(1-\alpha)$-quantile of $\max_{k \in 1,\ldots,K} |\widehat{\gamma}_k - \gamma_k|/\widehat{\sigma}_k$ and $\widehat{\sigma}_k$ is the standard error of $\widehat{\gamma}_k$ conditional on the data split. Monotonize the band (6.1) with respect to k using the rearrangement method of [19].

Step 5: Compute the adjusted $(1-2\alpha)$-confidence intervals and adjusted p-values using the VEIN methods described in Section 3.

Comment 6.1 (ML Methods)**.** We consider four ML methods to estimate the proxy predictors: elastic net, boosted trees, neural network with feature extraction, and random forest. The ML methods are implemented in R using the package caret [39]. The names of the elastic net, boosted tree, neural network with feature extraction, and random forest methods in caret are glmnet, gbm, pcaNNet and rf, respectively. For each split of the data, we choose the tuning parameters separately for $B(z)$ and $S(z)$ based on mean squared error estimates of repeated 2-fold cross-validation, except for random forest, for which we use the default tuning parameters to reduce the computational time.[12] In tuning and training the ML methods we use only the auxiliary sample. In all the methods we rescale the outcomes and covariates to be between 0 and 1 before training.

[12]We have the following tuning parameters for each method: Elastic Net: alpha (Mixing Percentage), lambda (Regularization Parameter), Boosted trees: n.trees (Number of Boosting Iterations), interaction.depth (Max Tree Depth), shrinkage (Shrinkage), n.minobsinnode (Min. Terminal Node Size), size (Number of Hidden Units) , decay (Weight Decay), mtry (Number of Randomly Selected Predictors).

Comment 6.2 (Microfinance Application)**.** We adopt two strategies to improve precision. First, the linear projections of the BLP and GATES control for village pair fixed effects along with the predicted baseline effect, $B(z)$ and predicted treatment effect, $S(z)$. Second, as suggested in Section 4, we use stratified sample splitting where the strata are village pairs. We cluster the standard errors at the village level to account for potential correlated shocks within each village. All reported results are medians over $S = 100$ splits and $\alpha = 0.05$.

Comment 6.3 (Gender Wage Gap Application)**.** We estimate the propensity scores using a neural network with feature extraction from a large sample of 129,983 workers from the U.S. March Supplement of the CPS pooling the years 2012 to 2017, excluding 2015. We select the neural network among elastic net, boosted tree, neural network, and random forest, and the tuning parameters by repeated 2-fold cross validation on the mean squared error. To avoid tuning ML methods for each data split separately, we tune ML methods by using 20% random extract from the 2012-2017 pooled sample and use those tuning parameters for all splits. This allows us to use $S = 1,000$ splits in this empirical application. Finally, to reduce the disproportionate impact of extreme propensity score weights we trimmed the predicted propensity scores between 0.01 and 0.99 in the main sample. These dropped observations make up approximately 4% of the sample.

7. Concluding Remarks

We propose to focus inference on key features of heterogeneous effects in randomized experiments, and develop the corresponding methods. These key features include best linear predictors of the effects and average effects sorted by groups, as well as average characteristics of most and least affected units. Our new approach is valid in high dimensional settings, where the effects are estimated by machine learning methods. The main advantage of our approach is its credibility: the approach is agnostic about the properties of the machine learning estimators, and does not rely on incredible or hard-to-verify assumptions. Estimation and inference relies on data splitting, where the latter allows us to avoid overfitting and all kinds of non-regularities. Our inference quantifies uncertainty coming from both parameter estimation and the data splitting, and could be of independent interest. Empirical applications illustrate the practical uses of the approach.

Appendix A. Proofs

Proof of Theorem 2.1. The subset of the the normal equations, which correspond to β_1 and β_2, are given by $\mathrm{E}[w(Z)(Y - \alpha' X_1 - \beta' X_2)X_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $X_2 = X_2(Z, D) = [D - p(Z), (D - p(Z)(S - \mathrm{E}S)]'$, that $X_1 = X_1(Z)$, and the law of

iterated expectations, we notice that:

$$\mathrm{E}[w(Z)b_0(Z)X_2] = \mathrm{E}[w(Z)b_0(Z)\underbrace{\mathrm{E}[X_2 \mid Z]}_{=0}] = 0,$$
$$\mathrm{E}[w(Z)UX_2] = \mathrm{E}[w(Z)\underbrace{\mathrm{E}[U \mid Z, D]}_{0}X_2(Z, D)] = 0,$$
$$\mathrm{E}[w(Z)X_1X_2] = \mathrm{E}[w(Z)X_1(Z)\underbrace{\mathrm{E}[X_2(Z, D) \mid Z]}_{=0}] = 0.$$

Hence the normal equations simplify to: $\mathrm{E}[w(Z)(s_0(Z)D - \beta' X_2)X_2] = 0$. Since

$$\mathrm{E}[\{D - p(Z)\}\{D - p(Z)\} \mid Z] = p(Z)(1 - p(Z)) = w^{-1}(Z),$$

and $S = S(Z)$, the components of X_2 are orthogonal by the law of iterated expectations:

$$\mathrm{E}w(Z)(D - p(Z))(D - p(Z))(S - \mathrm{E}S) = \mathrm{E}(S - \mathrm{E}S) = 0.$$

Hence the normal equations above further simplify to

$$\mathrm{E}[w(Z)\{s_0(Z)D - \beta_1(D - p(Z))\}(D - p(Z))] = 0,$$
$$\mathrm{E}[w(Z)\{s_0(Z)D - \beta_2(D - p(Z))(S - \mathrm{E}S)\}(D - p(Z))(S - \mathrm{E}S)] = 0.$$

Solving these equations and using the law of iterated expectations, we obtain

$$\beta_1 = \frac{\mathrm{E}w(Z)\{s_0(Z)D(D - p(Z))\}}{\mathrm{E}w(Z)(D - p(Z))^2} = \frac{\mathrm{E}w(Z)s_0(Z)w^{-1}(Z)}{\mathrm{E}w(Z)w^{-1}(Z)} = \mathrm{E}s_0(Z),$$
$$\beta_2 = \frac{\mathrm{E}w(Z)\{s_0(Z)D(D - p(Z))(S - \mathrm{E}S)\}}{\mathrm{E}w(Z)(D - p(Z))^2(S - \mathrm{E}S)^2}$$
$$= \frac{\mathrm{E}w(Z)s_0(Z)w^{-1}(Z)(S - \mathrm{E}S)}{\mathrm{E}w(Z)w^{-1}(Z)(S - \mathrm{E}S)^2} = \mathrm{Cov}(s_0(Z), S)/\mathrm{Var}(S).$$

The conclusion follows by noting that these coefficients also solve the normal equations

$$\mathrm{E}\{[s_0(Z) - \beta_1 - \beta_2(S - \mathrm{E}S)][1, (S - \mathrm{E}S)]'\} = 0,$$

which characterize the optimum in the problem of best linear approximation/prediction of $s_0(Z)$ using S. ∎

Proof of Theorem 2.2. The normal equations defining $\beta = (\beta_1', \beta_2')$ are given by $\mathrm{E}[(YH - \mu' X_1 H - \beta' \tilde{X}_2)\tilde{X}_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $\tilde{X}_2 = \tilde{X}_2(Z) = [1, (S(Z) - \mathrm{E}S(Z))]'$, that $X_1 = X_1(Z)$, and the law of iterated expectations, we notice that:

$$\mathrm{E}[b_0(Z)H\tilde{X}_2(Z)] = \mathrm{E}[b_0(Z)\underbrace{\mathrm{E}[H \mid Z]}_{=0}\tilde{X}_2(Z)] = 0,$$
$$\mathrm{E}[UH\tilde{X}_2(Z)] = \mathrm{E}[\underbrace{\mathrm{E}[U \mid Z, D]}_{0}H(D, Z)\tilde{X}_2(Z)] = 0,$$
$$\mathrm{E}[X_1(Z)H\tilde{X}_2(Z)] = \mathrm{E}[X_1(Z)\underbrace{\mathrm{E}[H \mid Z]}_{=0}\tilde{X}_2(Z)] = 0.$$

Hence the normal equations simplify to:

$$\mathrm{E}[(s_0(Z)DH - \beta' \tilde{X}_2)\tilde{X}_2] = 0.$$

Since 1 and $S - \mathrm{E}S$ are orthogonal, the normal equations above further simplify to

$$\mathrm{E}\{s_0(Z)DH - \beta_1\} = 0,$$
$$\mathrm{E}[\{s_0(Z)DH - \beta_2(S - \mathrm{E}S)\}(S - \mathrm{E}S)] = 0.$$

Using that

$$\mathrm{E}[DH \mid Z] = [p(Z)(1 - p(Z))]/[p(Z)(1 - p(Z))] = 1,$$

$S = S(Z)$, and the law of iterated expectations, the equations simplify to

$$\mathrm{E}\{s_0(Z) - \beta_1\} = 0,$$
$$\mathrm{E}\{s_0(Z) - \beta_2(S - \mathrm{E}S)\}(S - \mathrm{E}S) = 0.$$

These are normal equations that characterize the optimum in the problem of best linear approximation/prediction of $s_0(Z)$ using S. Solving these equations gives the expressions for β_1 and β_2 stated in the theorem. ■

Proof of Theorem 2.3. The proof is similar to the proof of Theorem 2.1- 2.2. Moreover, since the proofs for the two strategies are similar, we will only demonstrate the proof for the second strategy.

The subset of the normal equations, which correspond to $\gamma = (\gamma_k)_{k=1}^K$, are given by $\mathrm{E}[(YH - \mu'\tilde{W}_1 - \beta'\tilde{W}_2)\tilde{W}_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $\tilde{W}_2 = \tilde{W}_2(Z) = [1(S \in I_k)_{k=1}^K]'$, that $\tilde{W}_1 = X_1(Z)H$, and the law of iterated expectations, we notice that:

$$\mathrm{E}[b_0(Z)H\tilde{W}_2(Z)] = \mathrm{E}[b_0(Z)\underbrace{\mathrm{E}[H \mid Z]}_{=0}\tilde{W}_2(Z)] = 0,$$
$$\mathrm{E}[UH\tilde{W}_2(Z)] = \mathrm{E}[\underbrace{\mathrm{E}[U \mid Z, D]}_{0}H(D, Z)\tilde{W}_2(Z)] = 0,$$
$$\mathrm{E}[\tilde{W}_1\tilde{W}_2(Z)] = \mathrm{E}[X_1(Z)\underbrace{\mathrm{E}[H \mid Z]}_{=0}\tilde{W}_2(Z)] = 0.$$

Hence the normal equations simplify to:

$$\mathrm{E}[\{s_0(Z)DH - \beta'\tilde{W}_2\}\tilde{W}_2] = 0.$$

Since components of $\tilde{W}_2 = \tilde{W}_2(Z) = [1(G_k)_{k=1}^K]'$ are orthogonal, the normal equations above further simplify to

$$\mathrm{E}[\{s_0(Z)DH - \gamma_k 1(G_k)\}1(G_k)] = 0.$$

Using that

$$\mathrm{E}[DH \mid Z] = [p(Z)\{1 - p(Z)\}]/[p(Z)\{1 - p(Z)\}] = 1,$$

that $S = S(Z)$, and the law of iterated expectations, the equations simplify to

$$\mathrm{E}[\{s_0(Z) - \gamma_k 1(G_k)\}1(G_k)] = 0 \iff$$
$$\gamma_k = \mathrm{E}s_0(Z)1(G_k)/\mathrm{E}[1(G_k)] = \mathrm{E}[s_0(Z) \mid G_k].$$

■

Proof of Theorem 3.1. We have that $p_{.5} \leqslant \alpha/2$ is equivalent to $\mathbb{E}_P[1(p_A \leqslant \alpha/2) \mid \text{Data}] \geqslant 1/2$. So

$$\mathbb{P}_P[p_{.5} \leqslant \alpha/2] = \mathbb{E}_P 1\{\mathbb{E}_P[1(p_A \leqslant \alpha/2) \mid \text{Data}] \geqslant 1/2\}.$$

By Markov inequality,

$$\mathbb{E}_P 1\{\mathbb{E}_P[1(p_A \leqslant \alpha/2) \mid \text{Data}] \geqslant 1/2\} \leqslant 2\mathbb{P}_P[p_A \leqslant \alpha/2]$$

Moreover,

$$\mathbb{P}_P(p_A \leqslant \alpha/2) \leqslant \mathbb{E}_P[\mathbb{P}_P[p_A \leqslant \alpha/2 \mid \text{Data}_A \in \mathcal{A}] + \gamma] \leqslant \alpha/2 + \delta + \gamma.$$

■

Proof of Theorem 3.2. To show the second claim, we note that

$$\begin{aligned}\mathbb{P}_P(\theta_A \notin \text{CI}) &= \mathbb{P}_P(p_l(\theta_A) \leqslant \alpha/2) + \mathbb{P}_P(p_u(\theta_A) \leqslant \alpha/2) \\ &\leqslant \alpha + \delta + \gamma + \alpha + \delta + \gamma,\end{aligned}$$

where the inequality holds by Theorem 3.1 on the p-values. The last bound is upper bounded by $2\alpha + o(1)$ by the regularity condition PV for the p-values, uniformly in $P \in \mathcal{P}$.

To show the first claim, we need to show the following inequalities:

$$\sup\{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2\} \leqslant u, \quad \inf\{\theta \in \mathbb{R} : p_l(\theta) > \alpha/2\} \geqslant l.$$

We demonstrate the first inequality, and the second follows similarly.

We have that

$$\begin{aligned}\{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2\} &= \{\theta \in \mathbb{R} : \underline{\text{Med}}[\Phi\{\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta)\} \mid \text{Data}] > \alpha/2\} \\ &= \{\theta \in \mathbb{R} : \Phi\{\underline{\text{Med}}[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta) \mid \text{Data}]\} > \alpha/2\} \\ &= \{\theta \in \mathbb{R} : \underline{\text{Med}}[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta) \mid \text{Data}] > \Phi^{-1}(\alpha/2)\} \\ &= \{\theta \in \mathbb{R} : \overline{\text{Med}}[\widehat{\sigma}_A^{-1}(\theta - \widehat{\theta}_A) \mid \text{Data}] < \Phi^{-1}(1-\alpha/2)\} \\ &= \left\{\theta \in \mathbb{R} : \overline{\text{Med}}\left[\frac{\theta - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \mid \text{Data}\right] < 0\right\},\end{aligned}$$

where we have used the equivariance of $\overline{\text{Med}}$ and $\underline{\text{Med}}$ to monotone transformations, implied from their definition. We claim that by the definition of

$$u := \underline{\text{Med}}[\widehat{\theta}_A + \widehat{\sigma}_A \Phi^{-1}(1-\alpha/2) \mid \text{Data}],$$

we have

$$\overline{\text{Med}}\left[\frac{u - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \mid \text{Data}\right] \geqslant 0.$$

Indeed, by the definition of u we have that

$$\mathbb{E}\left(1(u - \widehat{\theta}_A - \widehat{\sigma}_A \Phi^{-1}(1-\alpha/2) \geqslant 0) \mid \text{Data}\right) \geqslant 1/2.$$

Since $\widehat{\sigma}_A > 0$ by assumption we have

$$1(u - \widehat{\theta}_A - \widehat{\sigma}_A \Phi^{-1}(1-\alpha/2) \geqslant 0) = 1\left(\frac{u - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \geqslant 0\right),$$

and it follows that

$$\mathbb{P}\Big(\frac{u - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \geqslant 0 \mid \mathrm{Data}\Big) \geqslant 1/2.$$

The claimed inequality $\sup\{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2\} \leqslant u$ follows. ∎

Appendix B. A Lemma on Uniform in P Conditional Inference

Lemma B.1. *Fix positive constants c and C, and a small constant $\delta > 0$. Let $\tilde{Y}$ and X denote generic outcome and generic d-vector of regressors, whose use and definition may differ in different places of the paper. Assume that for each $P \in \mathcal{P}$, $\mathrm{E}_P|\tilde{Y}|^{4+\delta} < C$ and let $0 < \underline{w} \leqslant w(Z) \leqslant \overline{w} < \infty$ denote the generic weights, and that $\{(\tilde{Y}_i, Z_i, D_i)\}_{i=1}^N$ are i.i.d. copies of $(\tilde{Y}, Z, D)$. Let $\{\mathrm{Data}_A \in \mathcal{A}_N\}$ be the event such that the ML algorithm, operating only on Data_A, produces a vector $X_A = X(Z, D; Data_A)$ that obeys, for $\epsilon_A = \tilde{Y} - X'\beta_A$ defined by: $\mathbb{E}_P[\epsilon_A w(Z) X_A \mid \mathrm{Data}_A] = 0$, the following inequalities, uniformly in $P \in \mathcal{P}$*

$$\mathbb{E}_P[\|X_A\|^{4+\delta} \mid \mathrm{Data}_A] \leqslant C, \;\; \textit{mineig}\; \mathbb{E}_P[X_A X_A' \mid \mathrm{Data}_A] > c, \;\; \textit{mineig}\; \mathbb{E}_P[\epsilon_A^2 X_A X_A' \mid \mathrm{Data}_A] > c.$$

Suppose that $\mathbb{P}_P\{\mathrm{Data}_A \in \mathcal{A}_N\} \geqslant 1 - \gamma \to 1$ uniformly in $P \in \mathcal{P}$, as $N \to \infty$. Let $\widehat{\beta}_A$ be defined by:

$$\mathbb{E}_{N,M}[w(Z) X_A \widehat{\epsilon}_A] = 0, \quad \widehat{\epsilon}_A = Y_A - X'\widehat{\beta}_A.$$

Let $\widehat{V}_{N,A} := (\mathbb{E}_{N,M} X_A X_A')^{-1} \mathbb{E}_{N,M} \widehat{\epsilon}_A^2 X_A X_A' (\mathbb{E}_{N,M} X_A X_A')^{-1}$ be an estimator of

$$V_{N,A} = (\mathbb{E}_P[X_A X_A' \mid \mathrm{Data}_A])^{-1} \mathbb{E}_P[\epsilon_A^2 X_A X_A' \mid \mathrm{Data}_A] (\mathbb{E}_P[X_A X_A' \mid \mathrm{Data}_A])^{-1}.$$

Then for any convex set R in $\mathbb{R}^d$, we have that uniformly in $P \in \mathcal{P}$:

$$\mathbb{P}_P[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathrm{Data}_A] \to_P \mathbb{P}(N(0, I) \in R),$$

$$\mathbb{P}_P[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \{\mathrm{Data}_A \in \mathcal{A}_N\}] \to \mathbb{P}(N(0, I) \in R),$$

and the same results hold with $\widehat{V}_{N,A}$ replaced by $V_{N,A}$.

Proof. It suffices to demonstrate the argument for an arbitrary sequence $\{P_n\}$ in $\mathcal{P}$. Let $z \mapsto \tilde{X}_{A,N}(z)$ be a deterministic map such that the following inequalities hold, for $\tilde{e}_A$ defined by

$$\mathrm{E}_{P_n}[\tilde{e}_A w(Z) \tilde{X}_{A,N}(Z)] = 0$$

and $\tilde{X}_A = \tilde{X}_A(Z)$:

$$\mathrm{E}_{P_n}[\|\tilde{X}_{A,N}\|^4] < C, \;\; \text{mineig}\; \mathrm{E}_{P_n}[\tilde{X}_{A,N} \tilde{X}_{A,N}'] > c, \;\; \text{mineig}\; \mathrm{E}_{P_n}[\tilde{e}_A^2 \tilde{X}_{A,N} \tilde{X}_{A,N}'] > c.$$

Let $d := \dim \beta_A$. Then we have that (abusing notation):

$$B_N := \sup_{\tilde{X}_{A,N}} \sup_{h \in \mathrm{BL}_1(\mathbb{R}^d)} |\mathbb{E}_{P_n} h(\tilde{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \mid \tilde{X}_{A,N}) - \mathbb{E} h(N(0, I))| \to 0,$$

by the standard argument for asymptotic normality of the least squares estimator, which utilizes the Lindeberg-Feller Central Limit Theorem. Here

$$\tilde{V}_{N,A} := (\mathrm{E}\tilde{X}_A\tilde{X}_A')^{-1}\mathrm{E}\tilde{\epsilon}_A^2\tilde{X}_A\tilde{X}_A'(\mathrm{E}\tilde{X}_A\tilde{X}_A')^{-1},$$

and $\mathrm{BL}_1(\mathbb{R}^d)$ denotes the set of Lipschitz maps $h : \mathbb{R}^d \to [0,1]$ with the Lipschitz coefficient bounded by 1.

Then, for the stochastic sequence $X_{A,N} = X_{A,N}(\mathrm{Data}_A)$, we have

$$\sup_{h\in \mathrm{BL}_1(\mathbb{R}^d)} |\mathbb{E}_{P_n}[h(V_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}] - \mathbb{E}[h(N(0,I))]| \leqslant B_N + 2(1 - 1\{\mathrm{Data}_A \in \mathcal{A}_N\}) \to_{P_n} 0.$$

Since under the stated bounds on moments, $\widehat{V}_{N,A}^{1/2}V_{N,A}^{-1/2} \to_{P_n} I$ by the standard argument for consistency of the Eicker-Huber-White sandwich, we further notice that

$$\sup_{h\in \mathrm{BL}_1(\mathbb{R}^d)} |\mathbb{E}_{P_n}[h(\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}] - \mathbb{E}_{P_n}[h(V_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}]|$$

$$\leqslant \mathbb{E}_{P_n}[\|\widehat{V}_{N,A}^{-1/2}V_{N,A}^{1/2} - I\| \wedge 1 \cdot \|V_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)\| \wedge 1 \mid X_{A,N}] \to_{P_n} \mathbb{E}[0 \wedge 1 \cdot \|N(0,I)\| \wedge 1] = 0,$$

in order to conclude that

$$\sup_{h\in \mathrm{BL}_1(\mathbb{R}^d)} \mathbb{E}_{P_n}[h(\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}] - \mathbb{E}[h(N(0,I))] \to_P 0.$$

Moreover, since $\mathbb{E}_{P_n}[h(\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}] = \mathbb{E}_{P_n}[h(\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid \mathrm{Data}_A]$, the first conclusion follows: $\mathbb{P}_{P_n}[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathrm{Data}_A] \to_{P_n} \mathbb{P}(N(0,I) \in R)$, by the conventional smoothing argument (where we approximate the indicator of a convex region by a smooth map with finite Lipschitz coefficient). The second conclusion

$$\mathbb{P}_{P_n}[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathrm{Data}_A \in \mathcal{A}_N] \to \mathbb{P}(N(0,I) \in R)$$

follows from the first by

$$\mathbb{P}_{P_n}[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathrm{Data}_A \in \mathcal{A}_N] =$$

$$= \mathbb{E}_{P_n}[\mathbb{P}_{P_n}[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathrm{Data}_A]1(\{\mathrm{Data}_A \in \mathcal{A}_N\})/\mathbb{P}_{P_n}\{\mathrm{Data}_A \in \mathcal{A}_N\}]$$

$$\to \mathrm{E}[\mathbb{P}(N(0,I) \in R) \cdot 1],$$

using the definition of the weak convergence, implied by the convergence to the constants in probability. ∎

References

[1] Alberto Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

[2] Alberto Abadie, Matthew M Chingos, and Martin R West. Endogenous stratification in randomized experiments. Technical report, National Bureau of Economic Research, 2017.

[3] James Albrecht, Anders Bjorklund, and Susan Vroman. Is there a glass ceiling in sweden? *Journal of Labor Economics*, 21(1):145–177, 2003.

[4] Manuela Angelucci, Dean Karlan, and Jonathan Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by compartamos banco. *American Economic Journal: Applied Economics*, 7(1):151–182, 2015.

[5] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[6] Susan Athey and Guido W Imbens. The econometrics of randomized experiments. *Handbook of Economic Field Experiments*, 1:73–140, 2017.

[7] Orazio Attanasio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons, and Heike Harmgart. The impacts of microfinance: Evidence from joint-liability lending in mongolia. *American Economic Journal: Applied Economics*, 7(1):90–122, 2015.

[8] Britta Augsburg, Ralph De Haas, Heike Harmgart, and Costas Meghir. Microfinance, poverty and education. 2012.

[9] Abhijit Banerjee, Emily Breza, Esther Duflo, and Cynthia Kinnan. Do credit constraints limit entrepreneurship? heterogeneity in the returns to microfinance. *Evanston, USA: Department of Economics Northwestern University*, 2015.

[10] Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1):22–53, 2015.

[11] Abhijit Vinayak Banerjee. Microcredit under the microscope: what have we learned in the past two decades, and what do we need to know? *Annu. Rev. Econ.*, 5(1):487–519, 2013.

[12] G. Barnard. Discussion of "Cross-validatory choice and assessment of statistical predictions" by Stone. *Journal of the Royal Statistical Society. Series B (Methodological)*, page 133?135, 1974.

[13] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81:608–650, 2014.

[14] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post selection inference for lad regression models. *arXiv preprint arXiv:1304.0282*, 2013.

[15] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation of nonparametric functions via square-root lasso. *arXiv preprint arXiv:1105.1475*, 2011.

[16] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[17] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[18] Francine D. Blau and Lawrence M. Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865, September 2017.

[19] V. Chernozhukov, I. Fernández-Val, and A. Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009.

[20] V. Chernozhukov, I. Fernandez-Val, and Y. Luo. The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *ArXiv e-prints*, December 2015.

[21] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

[22] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818, 2014.

[23] DR Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

[24] Bruno Crépon, Florencia Devoto, Esther Duflo, and William Parienté. Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in morocco. *American Economic Journal: Applied Economics*, 7(1):123–150, 2015.

[25] Jonathan Davis and Sara B Heller. Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. Technical report, National Bureau of Economic Research, 2017.

[26] Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *arXiv preprint arXiv:1606.03940*, 2016.

[27] Esther Duflo, Rachel Glennerster, and Michael Kremer. Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962, 2007.

[28] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[29] Christopher Genovese and Larry Wasserman. Adaptive confidence bands. *The Annals of Statistics*, pages 875–905, 2008.

[30] Evarist Giné and Richard Nickl. Confidence bands in density estimation. *The Annals of Statistics*, 38(2):1122–1170, 2010.

[31] Christian Hansen, Damian Kozbur, and Sanjog Misra. Targeted undersmoothing. *arXiv preprint arXiv:1706.07328*, 2017.

[32] John A Hartigan. Using subsample values as typical values. *Journal of the American Statistical Association*, 64(328):1303–1317, 1969.

[33] Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[34] Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

[35] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[36] Dean Karlan and Jonathan Zinman. Expanding credit access: Using randomized supply decisions to estimate the impacts. *The Review of Financial Studies*, 23(1):433–464, 2009.

[37] Dean Karlan and Jonathan Zinman. Microcredit in theory and practice: Using randomized credit scoring for impact evaluation. *Science*, 332(6035):1278–1284, 2011.

[38] Leslie Kish and Martin Richard Frankel. Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–37, 1974.

[39] Max Kuhn. Caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[40] Guillaume Lecué and Charles Mitchell. Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics*, 6:1803–1837, 2012.

[41] Mark G Low et al. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.

[42] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[43] Frederick Mosteller and John Wilder Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

[44] Natalia Rigol, Reshmaan Hussam, and Benjamin Roth. Targeting high ability entrepreneurs using community information: Mechanism design in the field, 2016.

[45] Alessandro Rinaldo, Larry Wasserman, Max G'Sell, Jing Lei, and Ryan Tibshirani. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv preprint arXiv:1611.05401*, 2016.

[46] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[47] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982.

[48] Alessandro Tarozzi, Jaikishan Desai, and Kristin Johnson. The impacts of microcredit: Evidence from ethiopia. *American Economic Journal: Applied Economics*, 7(1):54–89, 2015.

[49] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

[50] Larry Wasserman. Machine learning overview. In *Becker-Friedman Institute, Conference on ML in Economics*, 2016.

[51] Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

[52] Marten Wegkamp et al. Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273, 2003.

[53] Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.

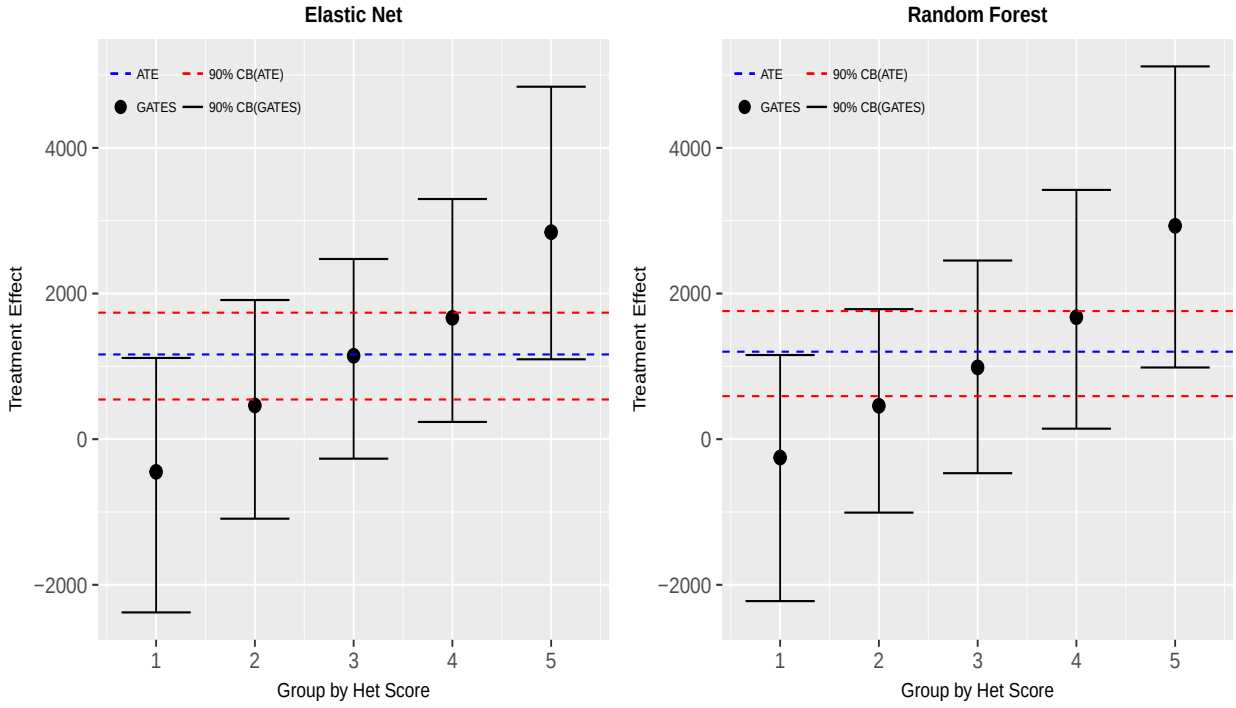

FIGURE 3. GATES of Microfinance Availability: Amount of Loans. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in half

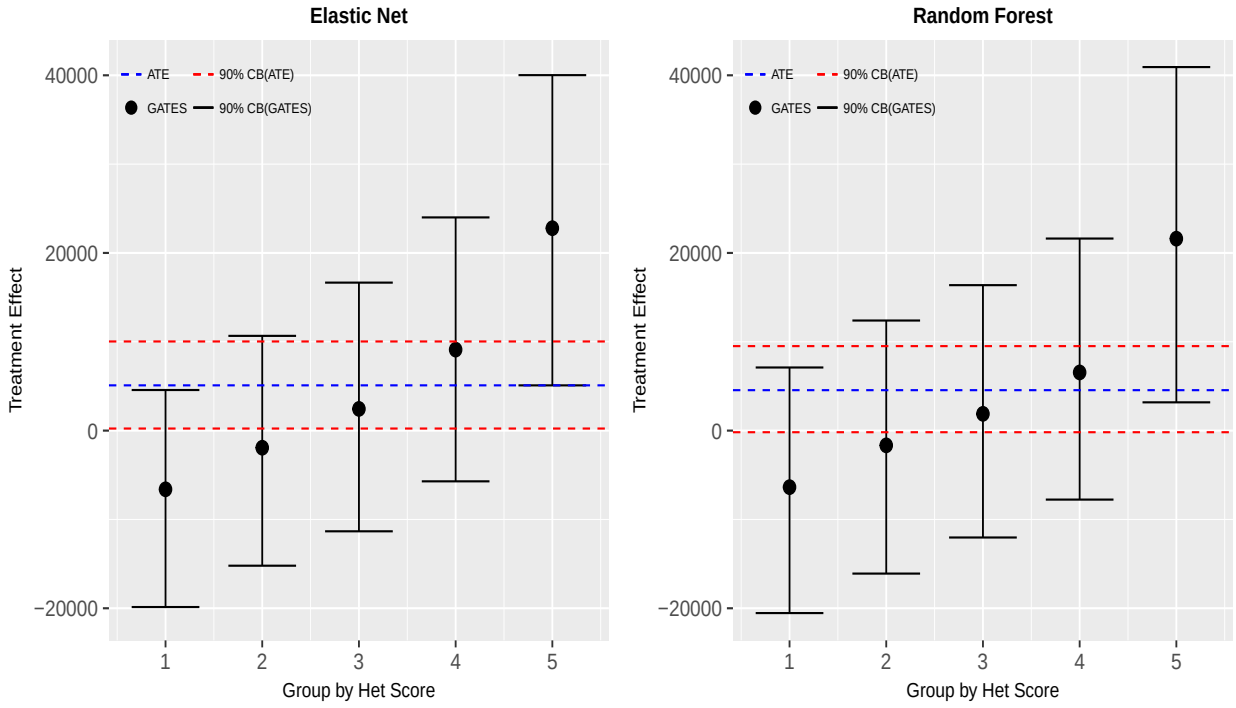

FIGURE 4. GATES of Microfinance Availability: Output. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in half

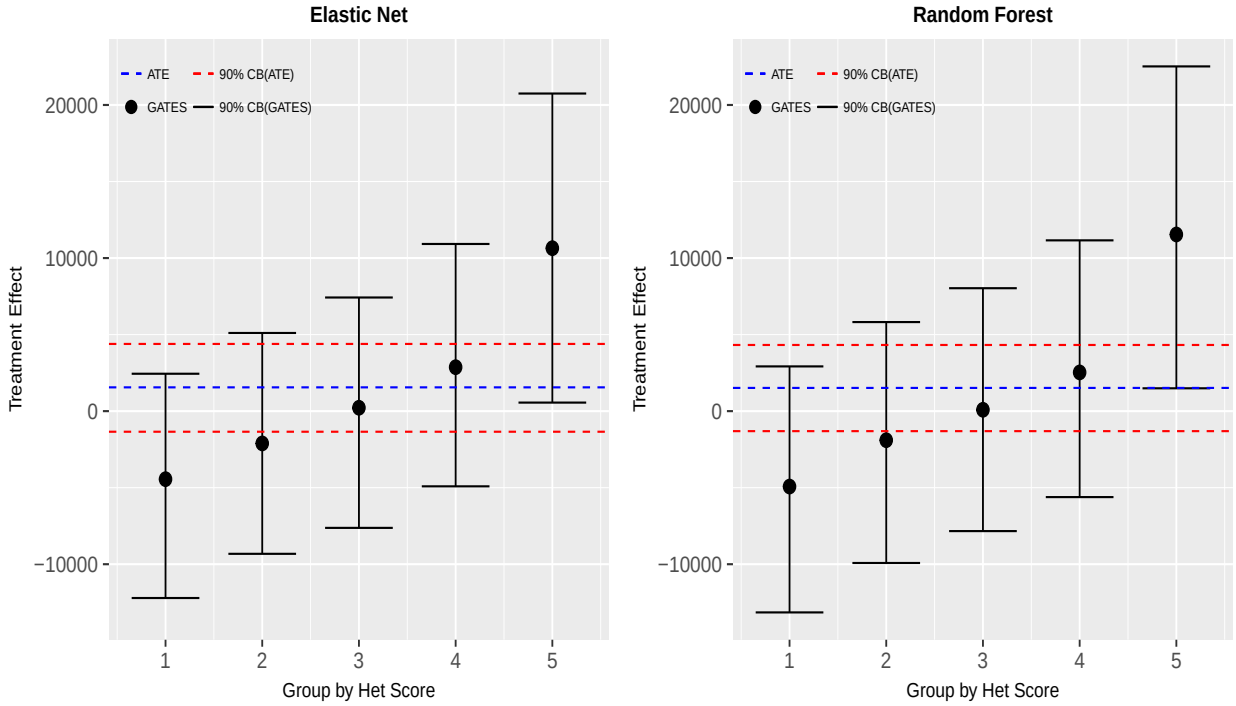

FIGURE 5. GATES of Microfinance Availability: Profit. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in half

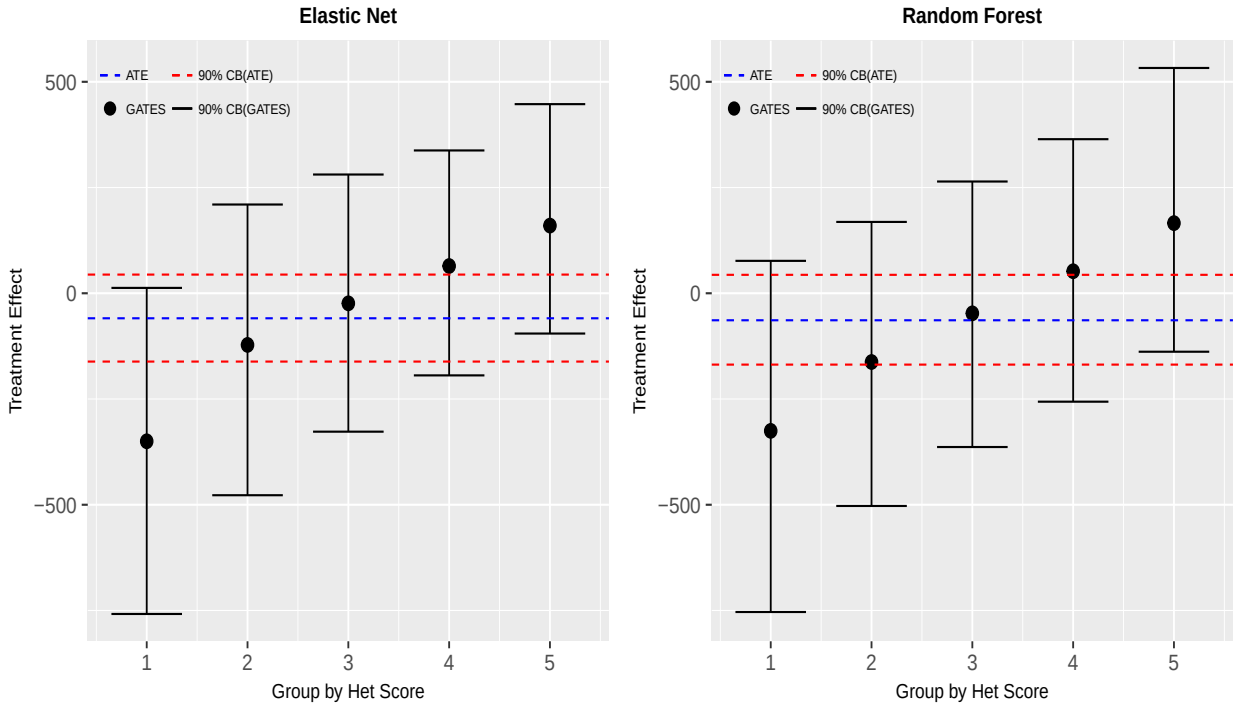

FIGURE 6. GATES of Microfinance Availability: Consumption. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in half

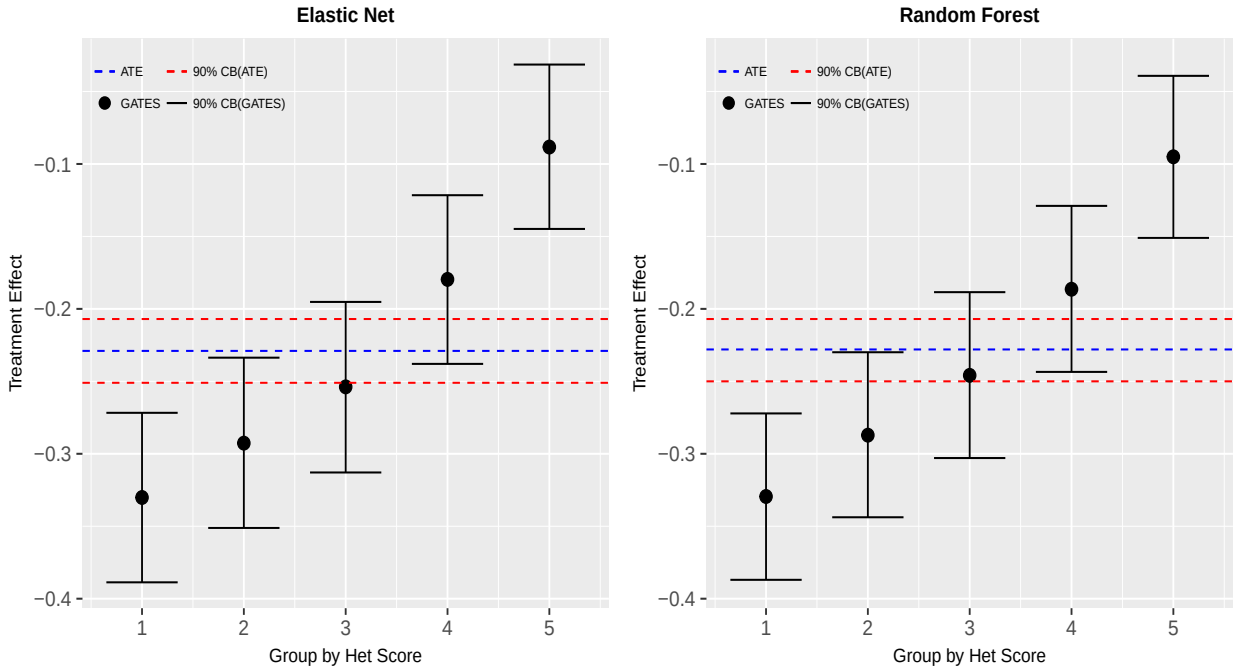

FIGURE 7. GATES of gender wage gap for $K = 5$ groups defined by the quintiles of $S(Z)$. Point estimates and 90% adjusted confidence intervals uniform across groups based on 1,000 random splits in half