Moment Trees for Heterogeneous Moment-Based Models

Sam Asher, Denis Nekipelov, Paul Novosad, and Stephen P. Ryan*

November 28, 2016

Abstract

A basic problem in applied settings is that different parameters may apply to the same model in different populations. We address this problem by proposing a method for the consistent estimation of moment-based models with heterogeneous parameters using *moment trees*. Leveraging the basic intuition of a classification tree, our method partitions the covariate space into disjoint subsets and fits a set of moments within each subspace. We prove the consistency of this estimator. Monte Carlo evidence demonstrates the excellent small sample performance and faster-than-parametric convergence rates of the estimator. Finally, we showcase the usefulness of the approach by applying our approach to estimate heterogeneous treatment effects in a regression discontinuity design in a development setting.

Keywords: GMM; Classification Trees; Heterogeneous Treatment Effects; Model Selection

*Asher, World Bank (sasher@worldbank.org); Nekipelov, University of Virginia (denis.nekipelov@gmail.com); Novosad, Dartmouth College, (paul.novosad@dartmouth.edu); Ryan, Washington University in St. Louis and NBER (stephen.p.ryan@wustl.edu).

1 Introduction

Applied researchers are faced with a multitude of decisions when constructing statistical models, such as which variables to include in the model, how those variables are related to the outcome variable, and how that mapping may vary across the units in the population. It is rare that the question of interest can be answered in complete statistical generality, necessitating decisions about the empirical specification. This process of determining the statistical model is often ad hoc, with the researcher adding and removing variables and interactions in a non-systematic fashion, either as a result of intuitive exploration or in the process of producing so-called "robustness checks." Two major issues arise from this process: the resulting statistical model after the search may have different statistical properties than the original model, as the result of choosing the specification on the basis of the answers it produces. The second problem is that the researcher often only considers a subset of the possible modeling choices, potentially introducing bias in the estimates. The aim of this paper is to propose a method that addresses both of those issues, recovering the correct specification in a systematic fashion without introducing bias in the estimates due to the search process.

Our method builds on classification trees, a technique from the computer science and machine learning literature for grouping observations in a sample together on the basis of some criterion function. We leverage these methods to assign statistical models to disjoint sets of a sample. As opposed to standard mixture models, e.g., a random coefficients logit, where individuals are assigned a type from some distribution but are assumed to follow one model, our method assigns a model with certainty to a group of observations. Using recent results on growing honest trees (Cappelli, Mola, and Siciliano (2002), Wager and Athey (2015), Athey and Imbens (2015)), we randomly split our sample into two halves. In the first sample, we estimate how models should be assigned to observations. We then estimate the parameters of those models using the second sample. Building on recent results by Wager and Walther (2015), we prove the consistency of this approach and show that bootstrapping can produce proper standard errors.

Our setting is one with a long academic literature. The academic medical community has long struggled with this issue, where it is commonly referred to as subgroup analysis.[1] The basic issue is that researchers, through statistical ignorance (or more nefarious motivations), may search across subgroups given a treatment until they find one with a statistically sig-

[1] See, for example, Assmann, Pocock, Enos, and Kasten (2000).

nificant deviation from the baseline. Only emphasizing this finding, while typically ignoring those other groups for which the effect is zero, leads to substantial reporting bias and can provide misleading policy implications. This problem has become so severe in the medical literature that it is becoming common to pre-announce your testing hypotheses in public before engaging on a clinical trial via a "pre-analysis plan." This practice has also started to become more widespread in economics, particularly in development.[2]

The problem of determining which models apply to which groups of observations is pervasive in economics. A simple example provides clear motivation; suppose that the researcher is interested in estimating the relationship between some outcome, y_i, and a vector of observable characteristics, x_i. A simple linear regression encapsulating these relationships might be:

$$y_i = x_i\beta + \epsilon_i, \tag{1.1}$$

where ϵ_i is an additive error term. Ignoring complicating issues such as selection bias, omitted variables, and measurement error, the researcher faces a problem of determining the form of the relationship between x and y. In principle, one can run a completely nonparametric regression, but in practice this is rarely, if ever, done for reasons of computational burden, lack of data, and poor statistical properties. Instead, researchers often take the following ad hoc heuristic approach to estimation.

First, either on the basis of theory or intuition, they estimate a "baseline" statistical model that estimates a common parameter vector across all observations. This might be a simple specification where all covariates are additive and separable. While some papers stop there, a common next step, especially in modeling settings where the estimation's computational burden may not present significant barriers to repeated specification testing, is to estimate a sequence of models where the parameters are allowed to vary across observations in some observable fashion. These models often take the form of interactions between demographic characteristics and outcomes. For example, Card (1999), in an influential chapter in the *Handbook of Labor Economics*, has a section discussing observable heterogeneity with many citations to prominent papers using statistical models with interaction effects. While econometric theory exists for various specification tests for growing or pruning models, this step is rarely guided by formal econometric intuition. Instead, the researchers consider a (small)

[2]The Hypothesis Registry at J-PAL (https://www.povertyactionlab.org/Hypothesis-Registry) is an early example; it is now subsumed by the the AEA RCT Registry (https://www.socialscienceregistry.org/).

finite number of specifications to run and report those results as "robustness checks." Robustness checks are pervasive throughout applied economics at the very highest level across all fields in economics; see e.g., Chetty, Hendren, and Katz (2015) in education, Banerjee, Barnhardt, and Duflo (2016) in development, Collard-Wexler and De Loecker (2015) in industrial organization, Barreca, Clay, Deschênes, Greenstone, and Shapiro (2015) in environmental, Doyle, Graves, Gruber, and Kleiner (2015) in health, and Heckman, Pinto, and Savelyev (2013) in labor. As a signal of what is being emphasized in graduate schools, every single one of the 2016 Ph.D. job market candidates at a top university writing in an applied field had some variety of robustness checks in their job market paper.[3]

While the desire to have a sense that one's estimates are not sensitive to the particular modeling choices made in forming those estimates is clearly laudable, there are two important limitations to this approach. The first is that these checks are rarely exhaustive or guided by some econometrically sound search process. One may erroneously conclude that the estimates are robust simply due to the subset of specifications that were chosen. In models with discrete variables, it is generally unheard of to see results that estimate the model on all subsets of the data. For one reason, there are typically too many subsets to consider. This problem becomes infinitely-dimensional when continuous variables are introduced, as any and all sub-intervals of the continuous variable may be considered. The other reason brings us to our second concern: the statistical properties of models constructed after a researcher searches through the model space are not the same as those if the models were predefined. One must account for that search process in order to engage in proper inference. That is the exact motivation for our paper, and the rest of the paper is organized around discussing the estimator (Section 2), developing its statistical properties (Sections 3-5), showing its small-sample performance in a Monte Carlo (Section 6), and applying it to a regression discontinuity design in a development setting (Section 7). Section 8 concludes.

2 The Estimator

Decision trees are an example of a recursive binary partitioning algorithm. Trees start with an initial "stump," with all the data grouped together, and proceed to recursively split the data along one dimension of the data at a time according to some criterion. For continuous variables, the algorithm chooses a split point somewhere along their support. For discrete variables, it searches over all disjoint binary sets. The split generates two disjoint sets of the

[3] A majority have a section expressly labeled "Robustness Checks."

data, each known as a "leaf." The algorithm repeats this process on each leaf, cutting the data into smaller and smaller subsets until a stopping criterion is met. The literature has considered several stopping criteria, such as requiring the number of observations in each leaf to be above some minimum integer k, requiring the proportion of data in each leaf to be at least some α, or requiring the improvement in the criterion function after the split to be greater than some threshold. Several variants of trees fall into this taxonomy. Two of the most common are classification trees and regression trees. Classification trees vote for assignment for an observation into a group on the basis of the observable variables; the criterion function is typically "node impurity," a measure of the dissimilarity of observations in a given node. Regression trees fit the average value of the subsample's dependent variable; the criterion function is the mean squared error within the leaf.

Our approach uses a variant of a classification tree that we term a *moment tree*. Like classification trees, we seek to group together observations that have the same parameter vector conditional on observable X. However, our criterion function is a moment function, which models the dependent variable as some function of the observable variables, parameters, and unobservable shocks. We recursively partition the data on the basis of observables into K sets, $X = \{X_1, \ldots, X_K\}$, and assign a unique parameter vector, θ_k, for each X_k to solve a moment function in that subgroup:

$$E[m(Y; X_k, \theta_k)] = 0. \tag{2.2}$$

If the moment function cannot be satisfied in a given sample, the leaf is assigned a value of null.

The literature has considered many variants of the basic decision tree approach. One variation that we adopt here is the extension of our moment tree to a *moment forest*. Forests are formed by repeatedly resampling the data with replacement and then growing a tree on each resampled data set. A key difference from the standard tree is that only a random subset of p variables are considered for splitting at each node. The estimate of θ_k is then the arithmetic average of the θ_k across all trees in the forest. This approach has at least two benefits; first, it is possible to show that one can reduce mean-squared prediction error down to irreducible structural error using resampling; and second, it allows the method to scale with large dimension X datasets, as only a subset of X is searched over at each split. To see the first property, let $\phi(x)$ be a predictor of Y in a given sample, and let $\mu(x) = E_x(\phi(x))$

be its expectation. Then:

$$\begin{aligned}
E([Y_x - \phi(x)]^2) &= E([(Y_x - \mu(x)) + (\mu(x) - \phi(x))]^2) \\
&= E([Y_x - \mu(x)]^2) + 2E(Y_x - \mu(x))E(\mu(x) - \phi(x)) + E([\mu(x) - \phi(x)]^2) \\
&= E([Y_x - \mu(x)]^2) + E([\mu(x) - \phi(x)]^2) \\
&= E([Y_x - \mu(x)]^2) + Var(\phi(x)) \\
&\geq E([Y_x - \mu(x)]^2).
\end{aligned}$$

Our approach relies on the use of so-called "honest trees," which are implemented by splitting the data into two samples. In the first sample, one grows the classification tree. In the second sample, the tree is taken as given but the values of the θ at each leaf are estimated using the second sample. We show below formally that this guarantees that the tree is uniformly consistent.

At the stump, our model is exactly the same as a standard GMM-based model, which encompasses an extraordinarily large class of empirical problems. One solves for the θ using the entire sample and computes the value of the GMM criterion function. Our approach extends the GMM approach by considering an addition step, which is to then search over all split of the data along each X to find the split that most decreases the value of the GMM criterion function across the two subsets.

3 Econometric Theory

3.1 Classification forest for moment models

We consider a general model which is defined by moment function $\rho(\cdot;\cdot) : \mathcal{Y} \times \Theta \mapsto \mathcal{M}$, where $\mathcal{Y}$ is a subset of $\mathbb{R}^n$, Θ is a convex compact subset of $\mathbb{R}^p$ and $\mathcal{M}$ is a subset of $\mathbb{R}^m$. We assume that the data generating process is characterized by vector of random variables (Y, X) where random variable X takes the values in $\mathcal{X} \subset \mathbb{R}^q$.

The data generating process can characterized by the marginal distribution of random vector X. We assume that this distribution has an absolutely continuous density $f_X(\cdot)$. Our results will apply to the cases where some of the components of X are discrete. The DGP is also characterized by the mixture over continuous distributions $f^k_{Y|X}(\cdot \,|\, x)$. The mixture weights $\pi^k(\cdot)$ are functions of the vector Z which is a strict subset of X. The support of Z, $\mathcal{Z}$ is

an open convex susbset of $\mathbb{R}^r$ $(r < q)$. The number of mixture components is bounded by $\bar{K} < \infty$.

ASSUMPTION 1. *Suppose that K is the actual number of mixture components. Then for each $1 \leq k \leq K$ there exists a convex compact subset $\mathcal{Z}^k \subset \mathcal{Z}$ such that $\pi^k(z) = 1$ for all $z \in \mathcal{Z}^k$.*

We formulate the econometric problem as the problem of estimation of the collection $\{K,\ \{\theta_k\}_{k=1}^K\}$ that includes the number of mixture components K and a set of parameters θ_k such that

$$E^k[\rho(Y;\theta^k) \,|\, X = x] = 0, \tag{3.3}$$

where $E^k[\cdot|X = x]$ corresponds to the expectation taken with respect to the mixture component k. We assume that both order and rank conditions are satisfied for each θ^k. Moreover for a given fixed $\delta > 0$ and for any $k \neq p$ we have $\|\theta_k - \theta_p\| \geq \delta$.

We now develop tree-based algorithm to estimate $\{K,\ \{\theta_k\}_{k=1}^K\}$ in Model (3.3).

For our analysis we use the notion of the random forest that will be based on the application of classification trees. The classification tree partitions the set $\mathcal{Z}$ into non-overlapping rectangles. Then each rectangle is assigned the label k and parameter θ^k corresponding to the appropriate component of the distribution mixture if such assignment is possible. We reserve label 0 and $\emptyset$ instead of the estimated parameter for the case where a particular element of the partition cannot be classified.

In our further analysis we assume that continuous components of $\mathcal{Z}$ lie in the interior of the hypercube. This can be done without loss of generality since any open convex sets in $\mathbb{R}^r$ are homeomorphic, i.e. we can define a one-to-one mapping from $\mathcal{Z}$ to the interior of the hypercube in $\mathbb{R}^r$. Our further analysis will then apply once Z is mapped into the hypercube.

The partitioning is performed recursively such that the algorithm begins with considering the set $S^{(0)} = \mathcal{Z} \subset \mathbb{R}^r$ (parent node of the tree). For this set we select dimension $1 \leq d \leq r$ and the threshold c such that $S^{(0)}$ is split into two children $S^{(1,1)} = S^{(0)} \cap \{z \in S^{(0)} \,|\, z^d \leq c\}$ and $S^{(1,2)} = S^{(0)} \cap \{z \in S^{(0)} \,|\, z^d > c\}$. If the component d is discrete, then we choose a particular value c of z^d and split $S^{(0)}$ into two children $S^{(1,1)} = S^{(0)} \cap \{z \in S^{(0)} \,|\, z^d = c\}$ and $S^{(1,2)} = S^{(0)} \cap \{z \in S^{(0)} \,|\, z^d \neq c\}$.

Then at split k we choose one of $k+1$ sets $S^{(k,i)}$. Then we choose the dimension d and, assuming that it is continuous, we select the threshold c and construct two sets $S^{(k+1,i)} =$

$S^{(k,i)} \cap \{z \in S^{(k,i)} \,|\, z^d = c\}$ and $S^{(k+1,k+2)} = S^{k,i} \cap \{z \in S^{(k,i)} \,|\, z^d \neq c\}$. Then we re-index the remaining sets $S^{(k,j)}$ as $S^{(k+1,j)}$.

The sequence of k splits induces the partition of $\mathcal{Z}$ which we denote $\mathcal{S}$. By L we denote a generic leaf of the partition. Also, let $L(z)$ be the element of $\mathcal{S}$ containing the point z. $L(z)$ will also be called the leaf of the classification tree containing z.

Following Wager and Walther (2015) we define $\{\alpha, k\}$-valid partition $\mathcal{S}$ as a partition generated by the recursive partitioning in which each node contains at least a fraction α of the data points in its parent node for some $0 < \alpha < \frac{1}{2}$ and each terminal node contains at least k observations. Use the notation $\Sigma_{\alpha,k}\left(\{z_i\}_{i=1}^n\right)$ to denote the set of all $\{\alpha, k\}$-valid partitions of the sample.

The idea behind the construction of the classification tree is the following. Suppose that L is a leaf of the classification tree. If $L \subseteq \mathcal{Z}^k$ for some k, then the moment condition

$$E[\rho(Y,\theta) \,|\, X = x] = \sum_{l=1}^{K} E^l[\rho(Y,\theta) \,|\, X = x]\pi^l(z) = E^k[\rho(Y,\theta) \,|\, X = x] = 0$$

has a solution θ^k for each point of L. However, if $L \not\subset \mathcal{Z}^k$ for any k, then the moment condition above does not have solutions.

We associate the unknown conditional expectation $E^k[\cdot \,|\, X = x]$ with an infinite-dimensional parameter which we denote $\eta \in \mathcal{H}$. Then we consider an estimator for the moment function $m(x;\theta,\eta) = E[\rho(Y,\theta) \,|\, X = x]$, denote it $\widehat{m}(x;\theta)$. We take weighting function $w(\cdot) : \mathcal{X} \mapsto \mathbb{R}^p$ such that $E[w(X)w(X)'] < \infty$ and $E[w(X)\frac{\partial m(X;\theta,\eta)}{\partial \theta'}]$ has full rank for each $\eta \in \mathcal{H}$ and for all θ in some fixed neighborhood of θ^k. In that case the finite-dimensional parameter of interest θ_k is identified from any leaf $L \subseteq \mathcal{Z}_k$ that generates function

$$M_L(\theta,\eta) = E\left[w(X)m(X;\theta,\eta)\,\mathbf{1}\{Z \in L\}\right]$$

Then we estimate the conditional expectation that yields $m(x;\theta,\widehat{\eta})$. Thus corresponding sample analog for $M(\cdot,\cdot)$ can be constructed as

$$\widehat{M}_L(\theta,\widehat{\eta}) = \frac{\sum\limits_{i\,:\,z_i \in L} w(x_i)m(x_i;\theta,\widehat{\eta})}{\#\{i \,:\, z_i \in L\}}$$

The classification will be based on the norm $\|\cdot\|$ and the threshold $\underline{M}_n > 0$. For the valid

partition we define the classification tree such that for each element of partition

$$T_{\mathcal{S}} : \mathcal{S} \mapsto \Theta \cup \emptyset,$$

and

$$T_{\mathcal{S}}(L) = \begin{cases} \arg\inf_\theta \|\widehat{M}_L(\theta, \widehat{\eta})\|, \text{ if } \inf_\theta \|\widehat{M}_L(\theta, \eta)\| \leq \underline{M}_n, \\ \emptyset, \quad \text{otherwise.} \end{cases}$$

In other words, the classification tree returns the parameter that solves the empirical moment condition if the minimum of the moment function is below the pre-set threshold. If the minimum is above the threshold (meaning that the solution that equates the moment function to zero cannot be found), then the tree returns null. Inside the leaves where the minimum is below the threshold we can replace the procedure with solving equation

$$\widehat{M}_L(\theta, \widehat{\eta}) = o(1)$$

which corresponds to the standard Z-estimator. The leaves of the tree are then assigned integer labels based on the inferred parameters. For a given $\delta_n > 0$ we assign two leaves L and L' the same integer label if $\|T_{\mathcal{S}}(L) - T_{\mathcal{S}}(L')\| \leq \delta_n$. The family of $\{\alpha, k\}$-valid trees is denoted $\mathcal{T}_{\alpha,k}(\{z_i\}_{i=1}^n)$.

Then the partition-optimal tree can be defined using the moment function

$$M_L(\theta, \eta) = E[w(X)m(X; \theta, \eta) \,|\, Z \in L]$$

and the norm $\|\cdot\|$ such that

$$T^*_{\mathcal{S}} : \mathcal{S} \mapsto \Theta \cup \emptyset$$

with

$$T^*_{\mathcal{S}}(L) = \begin{cases} \arg\inf_\theta \|M_L(\theta, \eta)\|, \text{ if } \inf_\theta \|M_L(\theta, \eta)\| = 0, \\ \emptyset, \quad \text{otherwise.} \end{cases}$$

Further, using the notation of Wager and Walther (2015), we define partition-optimal forests by considering a bootstrap sample of size B and the collection of $\{\alpha, k\}$-valid classification trees $T_{\mathcal{S}^{(1)}}, \ldots, T_{\mathcal{S}^{(B)}} \in \mathcal{T}_{\alpha,k}(\{z_i\}_{i=1}^n)$ and define $\{\alpha, k\}$-valid random forest.

To do that let

$$\mathcal{K}^k = \{(L, b) \in \Lambda \times \{1, \ldots, B\} : \forall (L, b), (L', b'), \|T_{\mathcal{S}^{(b)}}(L) - T_{\mathcal{S}^{(b')}}(L')\| < \delta, d_H(L, L') < \Delta\},$$

where $d_H(\cdot,\cdot)$ is the Hausdorff distance. Let $\bar{\mathcal{K}} = \cup_k \mathcal{K}$ and $\bar{K} = \#\bar{\mathcal{K}}$. Then the random forest is defined as a mapping

$$H_{\{\mathcal{S}^{(b)}\}_{b=1}^{B}} : \{1,\dots,\bar{K}\} \mapsto \Theta \cup \emptyset,$$

such that

$$H_{\{\mathcal{S}^{(b)}\}_{b=1}^{B}}(k) = \frac{1}{\#\mathcal{K}^k} \sum_{(b,L)\in\mathcal{K}^k} T_{\mathcal{S}^{(b)}}(L).$$

The set of $\{\alpha, k\}$-valid random forests is denoted $\mathcal{H}_{\alpha,k}\left(\{z_i\}_{i=1}^n\right)$. The partition-optimal forest is defined using the notion of the partition-optimal tree with

$$H^*_{\{\mathcal{S}^{(b)}\}_{b=1}^{B}}(k) = \theta_k,\ k = 1,\dots,K$$

for all partitions $\mathcal{S}$ such that for each $\mathcal{Z}^k$, $k = 1,\dots,K$ there exists a leaf $L \in \mathcal{S}$ with $L \subseteq \mathcal{Z}^k$.

3.2 Implementation of honest splitting rules

Wager and Athey (2015) propose to use an application of a cross-validation procedure to evaluate the tree splits. We adapt this idea to the evaluation of moment classification trees. We split the sample into two subsamples, where one subsample is used to estimate the moment functions $\widehat{m}(\theta; x)$ and the other one is used to split $\mathcal{Z}$ into rectangles.

To implement the procedure we take the sample $\{y_i, x_i, z_i\}_{i=1}^n$. First, we draw a subsample of size s from this sample without replacement and split it into two non-overlapping subsets $\mathcal{D}_t$ and $\mathcal{D}_v$.

Second, using the subset $\mathcal{D}_t$ we grow the tree.

Third, once the splits are made, we compute parameters and assign labels based on the minimization of the empirical moment function $\widehat{M}_L(\theta, \widehat{\eta})$ for each leaf using sample $\mathcal{D}_v$.

We adhere to a specific methodology for growing the tree, since unlike standard regression trees, the classification tree can assign a null label to elements of partition. The goal of the recursive splitting is to ensure that estimated moment function well approximates the true moment function defined by (3.3). Then we consider the weighted norm $\|\cdot\|$ with the

positive definition weighting matrix Ω such that

$$\|a\|^2 = a' \, \Omega \, a$$

and compute the overall prediction error for a given L as

$$\sum_{i \in \mathcal{D}_v} \sum_{L \in \mathcal{S}} \|w(x_i)\rho(y_i; \theta_L^*)\mathbf{1}\{z_i \in L\} - M_L(\widehat{\theta}, \widehat{\eta})\mathbf{1}\{z_i \in L\}\|^2,$$

where

$$\theta_L^* = \arg \inf_{\theta} \|E\left[w(X)m(X; \theta, \eta)\, \mathbf{1}\{Z \in L\}\right]\|$$

and

$$\widehat{\theta}_L = \arg \inf_{\theta} \|M_L(\theta, \widehat{\eta})\| \, .$$

The prediction error can be further re-written as

$$\sum_{i \in \mathcal{D}_v} \sum_{L \in \mathcal{S}} \Big(\|w(x_i)\rho(y_i; \theta_L^*)\|^2 + \|M_L(\widehat{\theta}_L, \widehat{\eta})\|^2 - 2\, \rho(y_i; \theta_L^*)' w(x_i)' \, \Omega M(\widehat{\theta}_L, \widehat{\eta})) \Big) \mathbf{1}\{z_i \in L\}.$$

Provided that $M_L(\cdot, \cdot)$ is fixed within the leaf L and there is a single minimizer $\widehat{\theta}_L$ of $M_L(\cdot . \widehat{\eta})$ for all $L \in \mathcal{S}$, then we can re-write

$$\sum_{i \in \mathcal{D}_v} \rho(y_i; \theta_L^*)' w(x_i)' \, \Omega \, M_L(\widehat{\theta}_L, \widehat{\eta})) = \sum_{L \in \mathcal{S}} \sum_{i \, : \, z_i \in L} \rho(y_i; \theta_L^*)' w(x_i)' \, \Omega \, M_L(\widehat{\theta}_L, \widehat{\eta})).$$

Under technical conditions that we discuss further, we can show that

$$\frac{1}{\#\{i \, : \, z_i \in L\}} \sum_{i \, : \, z_i \in L} w(x_i)\rho(y_i; \theta_L^*) = M(\widehat{\theta}_L, \widehat{\eta})) + o_p(1).$$

This means that

$$\begin{aligned} \sum_{i \in \mathcal{D}_v} \rho(y_i; \theta_L^*)' w(x_i)' \, \Omega \, M(\widehat{\theta}_L, \widehat{\eta})) &= \sum_{L \in \mathcal{S}} \#\{i \, : \, z_i \in L\} M(\widehat{\theta}_L, \widehat{\eta}))' \, \Omega \, M(\widehat{\theta}_L, \widehat{\eta})) + o_p(1) \\ &= \sum_{i \in \mathcal{D}_v} \|M(\widehat{\theta}_L, \widehat{\eta}))\|^2 + o_p(1). \end{aligned}$$

and the prediction error can be re-written as

$$\sum_{i\in\mathcal{D}_v} \|w(x_i)\rho(y_i;\theta_L^*)\|^2 - \sum_{i\in\mathcal{D}_v} \|M(\widehat{\theta}_L,\widehat{\eta}))\|^2 + o_p(1).$$

In other words, the optimal partition has to maximize the variance of the moment function. This result extends the observation in Athey and Imbens (2015) made for standard regression trees.

Now based on this idea we can construct an actual mechanism for producing new splits. Consider step k of the recursive splitting algorithm that paritions $\mathcal{Z}$ into subsets $S^{(k,1)},\dots,S^{(k,k+1)}$. Next, for each $i=1,\dots,k+1$ and each dimension d we consider threshold c that generates the new partition $S^{(k+1,1)}(i,c,d),\dots,S^{(k+1,k+2)}(i,c,d)$ according to the algorithm that we outlined previously. In each subset $S^{(k+1,j)}(i,c,d)$ we estimate the moment function $m(\theta;x)$ and define function

$$\widehat{M}_{i,c,d}^{(k+1,j)}(\theta,\widehat{\eta}) = \frac{\sum\limits_{i\,:\,z_i S^{(k+1,j)}(i,c,d)} w(x_i)m(x_i;\theta,\widehat{\eta})}{\#\{i\,:\,z_i\in S^{(k+1,j)}(i,c,d)\}}.$$

Then we find the set of minimizers

$$\widehat{\theta}_{i,c,d}^{(k+1,j)} = \arg\min_{\theta} \|\widehat{M}_{i,c,d}^{(k+1,j)}(\theta,\widehat{\eta})\|.$$

We note that we need to compute this only in the newly created elements of partition, while functions $\widehat{M}$ and their minimizers on the remaining elements of partition stay the same. Then we choose the triple (i,c,d) by maximizing the variance of the moment function

$$\max_{i,c,d} \sum_{j=1}^{k+2} \left\| \widehat{M}_{i,c,d}^{(k+1,j)}\left(\widehat{\theta}_{i,c,d}^{(k+1,j)},\widehat{\eta}\right)\right\|^2.$$

Step k, therefore, requires us to solve $2(k+1)$ minimization problems.

4 Consistency for Classification Trees and Forests

In this section, we develop a consistency result for a single tree and then generalize it to the random forest. Our first goal will be to show that the splits of the honest tree provide a uniform approximation to function $M_L(\theta,\eta)$ both over the parameter space and over the leafs

contained in $\mathcal{Z}^k$. Our first assumption establishes the properties of the moment function considered in estimation. In particular, with a known infinite dimensional parameter, we require that the moment functions have low complexity in the finite dimensional parameter.

ASSUMPTION 2. *The class of functions* $\{m(\cdot, \theta, \eta_0),\ \theta \in \Theta\}$ *has envelope* $\tilde{m}(\cdot)$ *such that* $\tilde{m}(\cdot) \leq \bar{m} < \infty$ *and* $P\,\tilde{m}(\cdot)^2 < \infty$ *and it admits polynomial discrimination.*

Our results rely on the existence of a "high quality" estimator for the ancillary parameter η. The corresponding condition is formulated as follows

ASSUMPTION 3. *Suppose that there exists a uniformly consistent estimator* $\widehat{\eta}$ *and deterministic sequence* $r_n \to \infty$ *such that* $r_n\, n^{1/4} \to \infty$ *such that*

$$\sup_{x \in \mathcal{X}, \theta \in \Theta} |r_n(m(x; \theta, \widehat{\eta}) - m(x; \theta, \eta_0))| = o_p(1) \quad (A.1)$$

and

$$\|\widehat{\eta} - \eta_0\| = o_p(n^{-1/4}).$$

For the rate defined in Assumption 3 we can define a weighted empirical process $\mathbb{G}_{r_n} = r_n(eP_n - P)$ and denote its norm with respect to a function class $\mathcal{F}_n$ as $\|\mathbb{G}_{r_n}\|_{\mathcal{F}_n} = \sup_{f \in \mathcal{F}_n} |\mathbb{G}_{r_n} f|$. Consider sequence $\delta_n \to 0$, and define the class of functions

$$\mathcal{M}_n = \left\{ \left(m(\cdot, \theta^k, \eta) - m(\cdot, \theta^k, \eta_0) \right) \mathbf{1}\{\cdot \in L\},\ \|\eta - \eta_0\| \leq \delta_n,\ L \subset \mathcal{Z}^k,\ P\,\mathbf{1}\{\cdot \in L\} \geq V/2 \right\}$$

and the shrinking neighborhood

$$U_n = \{(\theta, \eta)\ :\ \|\theta - \theta^k\| \leq \delta_n,\ \|\eta - \eta_0\| \leq \delta_n\}.$$

The following condition imposes stochastic equicontinuity on the empirical process associated with the moment function.

ASSUMPTION 4. *For any* $\delta_n \to 0$

$$\|\mathbb{G}_{r_n}\|_{\mathcal{M}_n} = O_p(1) \quad (A.2)$$

and

$$\mathbb{G}_{r_n} \left(m(\cdot, \theta, \eta) - m(\cdot, \theta^k, \eta) \right) \mathbf{1}\{\cdot \in L\} = O_p(\|\theta - \theta^k\|),\ \ (\theta, \eta) \in U_n. \quad (A.3)$$

Next we require the population moment function to be sufficiently smooth when estimated on the leaves that are subsets of $\mathcal{Z}^k$,

ASSUMPTION 5. *There exists a neighborhood of* (θ^k, η_0) *such that for all* $L \subset \mathcal{Z}^k$ *with* $P\,\mathbf{1}\{\cdot \in L\} \geq V/2$

$$P\left(m(\cdot, \theta, \eta) - m(\cdot, \theta^k, \eta_0)\right)\mathbf{1}\{\cdot \in L\} = A_L(\theta - \theta^k) + O(\|\theta - \theta^k\|^2 + \|\eta - \eta_0\|^2) \quad (A.4)$$

and $A_L \geq \underline{A}$ *for all such* L.

The set of imposed assumptions allows us to show consistency for each honest classification tree.

THEOREM 1. *Suppose that* $\widehat{L}$ *is the leaf of honest* $\{\alpha, k\}$*-valid classification tree that returns label* k *and*

$$\widehat{M}_{\widehat{L}}(\widehat{\theta}, \widehat{\eta}) = o(r_n^{-1}),$$

where $\widehat{M}_{\widehat{L}}(\cdot, \cdot)$ *is estimated from the subsample that was not used for splitting. Then whenever* $k = O(n/V)$ *under Assumptions 1-5*

$$d(\widehat{\theta}, \theta^k) = o_p(1).$$

Proof:
Consider the following decomposition

$$\widehat{M}_{\widehat{L}}(\theta, \widehat{\eta}) - M_L(\theta, \eta_0) = \widehat{M}_{\widehat{L}}(\theta, \widehat{\eta}) - \widehat{M}_{\widehat{L}}(\theta, \eta_0) + \widehat{M}_{\widehat{L}}(\theta, \eta_0) - \widehat{M}_L(\theta, \eta_0) + \widehat{M}_L(\theta, \eta_0) - M_L(\theta, \eta_0)$$

Assumption 3 guarantees that

$$\sup_\theta \left|\widehat{M}_{\widehat{L}}(\theta, \widehat{\eta}) - \widehat{M}_{\widehat{L}}(\theta, \eta_0)\right| = o_p(1).$$

Next consider the class of functions

$$\mathcal{F}_n(L, \gamma) = \left\{m(\cdot, \theta, \eta_0)\mathbf{1}\{\cdot \in L'\} - m(\cdot, \theta, \eta_0)\mathbf{1}\{\cdot \in L\},\, L' \subseteq L,\, e^{-\gamma}P\mathbf{1}\{\cdot \in L'\} \leq P\mathbf{1}\{\cdot \in L\}\right\}.$$

Wager and Walther (2015) establish the result that for $P\mathbf{1}\{\cdot \in L\} \geq V/2$, the cardinality of

the set of leaves L' that lead to class $\mathcal{F}_n(L,\gamma)$ is bounded by

$$\frac{2}{V}\left(\frac{8r^2}{\gamma^2}\left(1-\log_2 \text{round}(2/V)\right)\right)^r (1+O(\gamma)) = O(\gamma^{-2r}).$$

Let σ_i be the sequence of i.i.d. Radamacher random variables (i.e. $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = \frac{1}{2}$. For $f \in \mathcal{F}_n(L,\gamma)$ we define the symmetrized empirical process

$$\mathbb{P}^o f = \frac{1}{n}\sum_{i=1}^n \sigma_i\, f(x_i).$$

Then due to the symmetrization lemma (Van Der Vaart and Wellner (1996))

$$P\left(\sup_{\mathcal{F}_n(L,\gamma)} |\mathbb{P}f - Pf| > \epsilon\right) \le 4P\left(\sup_{\mathcal{F}_n(L,\gamma)} |\mathbb{P}^o f| > \frac{\epsilon}{4}\right), \quad \text{for } T \ge 8\epsilon^{-2}$$

Given the sample choose $g_1,\ldots,g_M$ where M is the $\frac{\epsilon}{8}$ cover of $\mathcal{F}_n(L,\gamma)$ meaning that

$$\min_j \mathbb{P}|f-g_j| \le \frac{1}{8}\epsilon, \quad \text{for each } f \in \mathcal{F}_n(L,\gamma).$$

Let f^* be the argmin. For any function $g \in L_1(\mathbb{P})$:

$$|\mathbb{P}^o g| = \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right| \le \frac{1}{n}\sum_{i=1}^n |f(x_i)| \equiv \mathbb{P}|g|.$$

Now we focus on the uncertainty associated with the Radamacher sequence σ_i and compute the probabilities conditional on the sample. Choose $g = f - f^*$ leading to

$$\begin{aligned} P\left(\sup_{\mathcal{F}_n(L,\gamma)} |\mathbb{P}^o f| > \tfrac{\epsilon}{4} \,\middle|\, \{x_i\}_{i=1}^n\right) &\le P\left(\sup_{\mathcal{F}_n(L,\gamma)} (|\mathbb{P}^o f^*| + \mathbb{P}|f-f^*|) > \tfrac{\epsilon}{4} \,\middle|\, \{x_i\}_{i=1}^n\right) \\ &\le P\left(\max_j |\mathbb{P}^o g_j| > \tfrac{\epsilon}{8} \,\middle|\, \{x_i\}_{i=1}^n\right) \\ &\quad M \max_j P\left(|\mathbb{P}^o g_j| > \tfrac{\epsilon}{8} \,\middle|\, \{x_i\}_{i=1}^n\right). \end{aligned}$$

Now recall than g_j are bounded by $\bar{m}$, thus can use Hoeffding inequality

$$\begin{aligned} P\left(|\mathbb{P}^o g_j| > \tfrac{\epsilon}{8} \,\middle|\, \{x_i\}_{i=1}^n\right) &= P\left(|\sum_{i=1}^n \sigma_i g_j(x_i)| > \tfrac{n\epsilon}{8} \,\middle|\, \{x_i\}_{i=1}^n\right) \\ &\leq 2\exp\left(-2\left(\tfrac{n\epsilon}{8}\right)^2 \Big/ \sum_{i=1}^n (2g_j(x_i))^2\right) \\ &\leq \exp\left(-n\epsilon^2/(128\bar{m})\right). \end{aligned}$$

Next note that since $\{m(\cdot,\theta,\eta_0),\, \theta \in \Theta\}$ admits polynomial discrimination then there exist constants $a > 0$ and $b > 0$ such that the size of the cover of this class is at most $a\epsilon^{-b}$. Provided that the cardinality of approximating rectangles is at most $O(\gamma^{-2r})$ then $M \leq O(\epsilon^{-b}\gamma^{-2r})$. As a result

$$P\left(\sup_{\mathcal{F}_n(L,\gamma)} |\mathbb{P}^o f| > \frac{\epsilon}{4} \,\middle|\, \{x_i\}_{i=1}^n\right) \leq 2\exp\left(O\left(2r\log\frac{1}{\gamma} + b\log\frac{1}{\epsilon}\right) - n\epsilon^2/(128(\bar{m})^2)\right).$$

Since the right-hand side of the evaluation does not depend on the sample, this bound holds for the unconditional probability. We also notice that the second term in the exponent dominates the first term if

$$\frac{\epsilon n}{\log\, n} \gg O(1)$$

and $\gamma = O(\epsilon)$ In that case we can evaluate

$$P\left(\sup_{\mathcal{F}_n(L,\gamma)} |\mathbb{P}^o f| > \frac{\epsilon}{4}\right) \leq \exp\left(O(-n\,\epsilon^2)\right).$$

This shows that the choice $\frac{\epsilon n}{\log\, n} \gg O(1)$ ensures that

$$\sup_\theta \left|\widehat{M}_{\widehat{L}}(\theta,\widehat{\eta}) - \widehat{M}_L(\theta,\eta_0)\right| = o_p(1).$$

The last term is evaluated in an analogous fashion. Therefore, we just established that

$$\sup_\theta \left|\widehat{M}_{\widehat{L}}(\theta,\widehat{\eta}) - M_L(\theta,\eta_0)\right| = o_p(1).$$

Provided our assumption of continuity of $M_L(\theta,\eta_0)$, this leads to consistency of the estimator $\widehat{\theta}$. *Q.E.D.*

Having established consistency of our estimator, we can now evaluate its convergence rate. To do that we make an additional assumption regarding the differentiability of the map associated with $M_L(\cdot,\cdot)$.

ASSUMPTION 6. *Map $M_L(\cdot,\eta_0)$ is Frechet-differentiable at θ^k for each $L \subset \mathcal{Z}^k$ with $P\mathbf{1}\{\cdot \in L\} \geq V/2$ so that*

$$\|M_L(\theta,\eta_0) - M_L(\theta_0,\eta_0) - \dot{M}_L(\theta-\theta_0)\| = o(\|\theta-\theta_0\|).$$

THEOREM 2. *Suppose that conditions of Theorem 1 are satisfied. Then*

$$r_n d(\widehat{\theta},\theta^k) = O_p(1).$$

Proof:
We established that $d(\widehat{\theta},\theta^k) = o(1)$ and thus we can now focus on the shrinking neighborhood U_n. Consider the empirical process

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f,\ f \in \mathcal{F}_n(L,\gamma).$$

We have demonstrated that for the choice of $\gamma = \epsilon$ the class $\mathcal{F}_n(L,\gamma)$ admits polynomial discrimination with bound $O(\epsilon^{-2r-b})$. Let $N(\epsilon\|\tilde{m}\|_{Q,2}, \mathcal{F}_n(L,\gamma), L_2(Q))$ be the covering number for the class $\mathcal{F}_n(L,\gamma)$. Recall that the covering integral is defined as

$$J(\delta,\mathcal{F}_n(L,\gamma)) = \sup_Q \int_0^\delta \sqrt{1+\log\, N(\epsilon\|\tilde{m}\|_{Q,2}, \mathcal{F}_n(L,\gamma), L_2(Q))}\, d\epsilon.$$

Provided that we were able to bound $N(\epsilon\|\tilde{m}\|_{Q,2}, \mathcal{F}_n(L,\gamma), L_2(Q))$ by $O(\epsilon^{-2r-b})$, then we can evaluate the corresponding covering integral as

$$O(\delta\sqrt{\log\frac{1}{\delta}}),$$

which is bounded from above by $\delta^*\sqrt{\log\frac{1}{\delta^*}}$ where $\delta^*\log\frac{1}{\delta^*} = \frac{1}{2}$. This allows us to evaluate

$$\|\mathbb{G}_n\|_{\mathcal{F}_n(L,\epsilon)} = O(P\,\tilde{m}(\cdot)^2)$$

using Theorem 2.14.1 in Van Der Vaart and Wellner (1996). This means that the classification error is negligible relative to the estimation error. Thus we can use the convergence

result for standard Z-estimators adopted to the empirical process $\mathbb{G}_{r_n}$ applying Theorem 3.3.1 in Van Der Vaart and Wellner (1996) which yields the statement of the theorem. *Q.E.D.*

5 Extensions of the Moment-based Model

Previously we constructed a semiparametric model whose complexity was restricted in two ways. First, we assumed that the moment function $\rho(\cdot;\cdot)$ is a function of a finite-dimensional parameter vector. Second, we assumed that the heterogeneity of the sample is characterized by a finite set of semiparametric models that reflect a finite set of applied policies or behavior models.

We now consider extensions of our framework in two important directions. First, while maintaining the assumption that the observed data is generated by a finite set of policies or behaviors, we allow the models themselves to become more complex. The moment vector $\rho(\cdot;\cdot)$ is also characterized by an infinite-dimensional parameter. That means that we model the data generating process using a finite set of semi-parametric or non-parametric models.

Second, we consider the same semiparametric structure with the moment function $\rho(\cdot;\cdot)$ characterized by a finite-dimensional parameter. At the same time, we allow the set of models to be large and potentially grow with the sample size. This extension is based on the observation in Wager and Athey (2015) who observe that the regression tree framework can be considered as adaptive smoothing, similar to the k-nearest neighbor estimator. In this case, $\rho(\cdot;\cdot)$ plays the role of the local model whose parameters depend on point in the support of conditioning variable Z.

5.1 Semiparametric model with the finite tree structure

Suppose that our Assumption 1 holds meaning that the tree structure is finite. Consider a general moment model, as before characterized by the moment function $\rho(\cdot,\cdot)$ which is now a function of the infinite-dimensional parameter $h(\cdot)$ that needs to be estimated a nuisance parameter η (reflecting the conditional expectation)

$$m(x, h(\cdot), \eta(\cdot)) = E[\rho(Y, h(\cdot)) | X = x]$$

The two infinite-dimensional components , $h(\cdot) \in \mathcal{H}_h$ and $\eta(\cdot) \in \mathcal{H}_\eta$ that are contained in the Banach spaces $\mathcal{H}_\eta$ and $\mathcal{H}_h$. The sieve approach, studied in a sequence of papers by

Chen and Shen (1998), Ai and Chen (2003) and Chen, Linton, and Van Keilegom (2003), approximates the class of infinite dimensional functions $\mathcal{H}$ using a parametric family of functions $\mathcal{H}_n$ whose dimension increases to infinity with the sample size n. Since in our setup both the parameter of interest $h(\cdot)$ and the nuisance parameter $\eta(\cdot)$ can be infinite-dimensional, there will be a non-trivial interaction between their asymptotic behaviors. As a result, we choose to explicitly model η considering two most commonly used methods for estimation of conditional expectations: orthogonal series and kernel smoothers.

The infinite-dimensional parameter of interest $h(\cdot)$ is assumed to be estimated using sieves. The series estimator used to recover the conditional moment function is based on the vector of basis functions $p^N(x) = (p_{1N}(x), \ldots, p_{NN}(x))'$,

$$m\left(x; h, \widehat{\eta}\right) = p^{N\prime}(x) \left(\frac{1}{n} \sum_{i=1}^{n} p^N(x_i) p^{N\prime}(x_i) \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} p^N(x_i) \rho\left(y_i,\, h\right). \tag{5.4}$$

The kernel estimator is defined using a multi-dimensional kernel function $K(\cdot)$ and a bandwidth sequence b_n as

$$m\left(x; h, \widehat{\eta}\right) = \left(\frac{1}{n b_n^{d_x}} \sum_{i=1}^{n} K\left(\frac{x_i - x}{b_n^{d_x}} \right) \right)^{-1} \frac{1}{n b_n^{d_x}} \sum_{i=1}^{n} K\left(\frac{x_i - x}{b_n} \right) \rho\left(y_i, h\right). \tag{5.5}$$

In either case, we will denote the resulting estimator by $m\left(x; h\widehat{\eta}\right)$. We then consider the same structure for the estimator as in the parametric case where we split the support of the conditioning covariate vector Z and estimate parameter h by setting the projected moment function $\widehat{M}_L(h, \widehat{\eta})$ in the leaf of the tree L equal to zero.

As in our analysis of parametric models, we focus on i.i.d data samples. We also impose standard assumptions on the basis functions as in Newey (1997). Well known conditions that satisfy Assumption 7 are available in, for example, the handbook chapter by Chen (2007).

ASSUMPTION 7. *For the basis functions $p^N(x)$ the following holds:*

(i) The smallest eigenvalue of $E\left[p^N(X)\, p^{N\prime}(X)\right]$ is bounded away from zero uniformly in N.[4]

(ii) For some $\zeta_0(N)$ such that $\zeta_0(N)^2 N/n \to 0$, $\sup_{x \in \mathcal{X}} \|p^N(x)\| \leq \zeta_0(N)$.

[4]We note that the considered series basis may not be orthogonal with respect to the semi-metric defined by the distribution of X.

(iii) The population conditional moment belongs to the completion of the sieve space and for some $\alpha > 0$,

$$\sup_{(h,\eta)\in\mathcal{H}_h\times\mathcal{H}_\eta}\sup_{x\in\mathcal{X}}\left\|m\left(x;h,\eta\right)-proj\left(m\left(x;h,\eta\right)\mid p^N(x)\right)\right\|=O\left(N^{-\alpha}\right).$$

Assumption 7[ii] is convenient because $\rho(\cdot)$ is uniformly bounded. It can potentially be relaxed to allow for a sequence of constants $\zeta_0(N)$ with $\sup_{x\in\mathcal{X}}\|p^N(x)\|\leq\zeta_0(N)$, where $\zeta_0(N)$ grows at appropriate rates as in Newey (1997) such as $\zeta_0(N)^2N/n\to 0$ as $n\to\infty$.

When all the basis functions are uniformly bounded, typically $\zeta_0(N)=\sqrt{N}$. In the above

$$proj\left(m\left(x;h,\eta\right)\mid p^N(x)\right)=p^N(x)'\left(Ep^N(X)p^N(X)'\right)^{-1}Ep^N(X)m\left(X;h,\eta\right).$$

The following assumption on the moment function $\rho(\cdot)$ does not require smoothness or continuity (see Shen and Wong, 1994; Zhang and Gijbels, 2003).

ASSUMPTION 8. *(i) The moment functions are uniformly bounded: $\sup_{h,y}\|\rho(y,h)\|\leq C$. The density of covariates X is uniformly bounded away from zero on its support.*

(ii) Suppose that $0\in\mathcal{H}_n$ and for some $C>0$,

$$\sup_{x\in\mathcal{X},h\in\mathcal{H}_n,|h|<C,}Var\left(\rho\left(Y,h\right)\mid X=x\right)=O\left(1\right),$$

(iii) For each n, the class of functions $\mathcal{F}_n=\{\rho(\cdot,h),\,h\in\mathcal{H}_n\}$ is Euclidean whose graphs form a polynomial class of sets and whose coefficients depend on the number of sieve terms. There exist constants A, and $0^+\leq r_0<\frac{1}{2}$ such that the covering number satisfies

$$\log N\left(\delta,\mathcal{F}_n,\mathbf{L}_1\right)\leq A\,n^{2r_0}\log\left(\frac{1}{\delta}\right),$$

and for $r_0=0^+$, n^{0+} is defined as $\log n$.

Denote $\pi_n h=\arg\inf_{h'\in\mathcal{H}_n}\|h'-h\|_\infty$. And let $d(\cdot)$ be the metric generated by the $\mathbf{L}^1$ norm. The following result extends Theorem 37 of Pollard (2012) to the case of sieve estimators. A related idea for unconditional sieve estimation has been used in Zhang and Gijbels (2003).

LEMMA 1. *Suppose that* $d(\pi_n h, h) = O\left(n^{-\phi}\right)$. *Under Assumptions 7 and 8 for series estimator* $\widehat{\eta}$

$$\sup_{d(h,h_0)=o(1),h\in\mathcal{H}_n} |m(x;h,\widehat{\eta}) - m(x;h,\eta)| = o_p(1)$$

uniformly in x *provided that* $N \to \infty$, *and* $\zeta_0(N)^2 N\, n^{2r_0-1} \log n \to 0$.

Proof of Lemma 1

It follows directly from Assumption 7.[iii] that for $(h,\eta) \in \mathcal{H}_h \times \mathcal{H}_\eta$

$$\left|m(x;h,\eta) - proj\left(m(x;h,\eta)\,|p^N(x)\right)\right| = O\left(\frac{1}{N^\alpha} + \frac{1}{n^\phi}\right),$$

that will converge to zero if $N \to \infty$ as $n \to \infty$.

Therefore it suffices to prove Lemma 1 for

$$(*) = \left|m(x;h,\widehat{\eta}) - proj\left(m(x;h,\eta)\,|p^N(x)\right)\right|.$$

As demonstrated in Newey (1997), for $P = \left(p^N(x_1),\ldots,p^N(x_n)\right)'$ and $\hat{Q} = P'P/n$

$$\|\hat{Q} - Q\| = O_p\left(\sqrt{\frac{N}{n}}\zeta_0(N)\right),$$

and Q is non-singular by Assumption 7.[i] with the smallest eigenvalue bounded from below by some constant $\underline{\lambda} > 0$. Hence the smallest eigenvalue of $\hat{Q}$ will converge to $\underline{\lambda} > 0$. Following Newey (1997) we use the indicator 1_n to indicate the cases where the smallest eigenvalue of $\hat{Q}$ is above $\frac{1}{2}$ to avoid singularities.

We consider conditional expectation $E[\rho(Y_i,h)\,|\,X_i = x]$ as a function of x (given h). We can project this function of x on N basis vectors of the sieve space. Let β be the vector of coefficients of this projection. Denote $\Gamma(h) = (\rho(Y_i,h))_{i=1}^n$ Also define $G(h) = (E[\rho(Y_i,h)\,|\,X_i])_{i=1}^n$. Then $(*)$ equals to a linear combination of $1_n|p^{N\prime}(x)\left(\hat{\beta} - \beta\right)|$. Note that

$$p^{N\prime}(x)\left(\hat{\beta} - \beta\right) = p^{N\prime}(x)\left(\hat{Q}^{-1}P'(\Gamma - G)/n + \hat{Q}^{-1}P'(G - P\beta)/n\right). \tag{5.6}$$

For the first term in (5.6), we can use the result that smallest eigenvalue of $\hat{Q}$ is converging

to $\underline{\lambda} > 0$. Then application of the Cauchy-Schwartz inequality leads to

$$\left| p^{N\prime}(x)\hat{Q}^{-1}P'\left(\Gamma - G\right) \right| \leq \left\| \hat{Q}^{-1}p^{N}(x) \right\| \left\| P'\left(\Gamma - G\right) \right\|.$$

Then $\left\| \hat{Q}^{-1}p^{N}(x) \right\| \leq \frac{\zeta_0(N)}{\underline{\lambda}}$,and

$$\begin{aligned} \left\| P'\left(\Gamma - G\right) \right\| &= \sqrt{\sum_{k=1}^{N}\left(\sum_{i=1}^{n} p_{Nk}(x_i)\left(\Gamma_i(h) - G_i(h)\right)\right)^2} \\ &\leq \sqrt{N}\max_k \left| \sum_{i=1}^{n} p_{Nk}(z_i)\left(\Gamma_i(h) - G_i(h)\right) \right| \end{aligned}$$

Thus,

$$\left| p^{N\prime}(x)\hat{Q}^{-1}P'\left(\Gamma - G\right) \right| \leq \frac{\zeta_0\left(N\right)\sqrt{N}}{\underline{\lambda}} \max_k \left| \sum_{i=1}^{n} p_{Nk}(x_i)\left(\Gamma_i(h) - G_i(h)\right) \right|.$$

Denote $\mu_n = \mu\, N^{-1}$. Next we adapt the arguments for proving Theorem 37 in Pollard (2012) to provide the bound for $P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\|p^{N\prime}(z)\hat{Q}^{-1}P'\left(\Gamma - G\right)\| > N\mu_n\right)$. For N non-negative random variables Y_i we note that

$$P\left(\max_i Y_i > c\right) \leq \sum_{i=1}^{N} P\left(Y_i > c\right).$$

Using this observation, we can find that

$$P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\|p^{N\prime}(z)\hat{Q}^{-1}P'\left(\Delta - G\right)\| > N\mu_n\right) \leq \sum_{k=1}^{N} P\left(\sup_{\mathcal{F}_n}\left\| \frac{1}{n}\sum_{i=1}^{n} p_{Nk}(x_i)\left(\Gamma_i - G_i\right)\right\| > \frac{\sqrt{N}}{\zeta_0\left(N\right)}\mu_n\right)$$

This inequality allows us to substitute the tail bound for the class of functions $m\left(x; h, \eta\right)$ that is indexed by h, η and x by a tail bound for a much simpler class

$$\mathcal{P}_n = \{p_{Nk}(\cdot)\left(\Gamma(h) - G(h)\right) \ : \ d(h, h_0) = o(1),\, h \in \mathcal{H}_n\}.$$

We note that, according to Lemma 2.6.18 in Van Der Vaart and Wellner (1996), provided that each $p_{Nk}(\cdot)$ is a fixed function, the covering number for $\mathcal{P}_n$ has the same order as

the covering number for $\mathcal{F}_n$. Then we pick A to be the largest constant for the covering numbers $A_k n^{2r_0} \log\left(\frac{1}{\delta}\right)$ over classes $\mathcal{P}_n$. By Assumption 7.[i] and 8.[i] any $f \in \mathcal{P}_n$ is bounded $|f| < C < \infty$. Next we note that $\text{Var}(f) = O(1)$ for $f \in \mathcal{P}_n$ by Assumption 8.[ii]. The symmetrization inequality (30) in Pollard (2012) holds if $1/(16n\,\mu_n^2) \leq \frac{1}{2}$. This will occur if $\frac{n}{N^2} \to \infty$. Provided that the symmetrization inequality holds, we can follow the steps of Theorem 37 in Pollard (2012) to establish the tail bound on the sample sum via a combination of the Hoeffding inequality and the covering number for the class $\mathcal{P}_n$. As a result, we obtain that

$$
\begin{aligned}
&P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\left\|\sum_{i=1}^{n} p_{Nk}(x_i)\left(\Gamma_i - G_i\right)\right\| > 8\frac{\sqrt{N}}{\zeta_0(N)}\mu_n\right) \\
&\leq 2\exp\left(An^{2r_0}\log\frac{\zeta_0(N)}{\sqrt{N}\mu_n}\right)\exp\left(-\frac{1}{128}\frac{nN\mu_n^2}{\zeta_0(N)^2}\right) + P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\left\|\sum_{i=1}^{n} p_{Nk}(z_i)\left(\Delta_i - G_i\right)\right\|^2 > 64\right).
\end{aligned}
$$

The second term can be evaluated with the aid of Lemma 33 in Pollard (2012):

$$
P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\left\|\sum_{i=1}^{n} p_{Nk}(z_i)\left(\Delta_i - G_i\right)\right\|^2 > 64\right) \leq 4\exp\left(An^{2r_0}\right)\exp\left(-n\right).
$$

As a result, we find that

$$
\begin{aligned}
P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\|p^{N\prime}(x)\hat{Q}^{-1}P'\left(\Gamma - G\right)\| > N\mu_n\right) &\leq 2N\exp\left(An^{2r_0}\log\frac{\zeta_0(N)}{\sqrt{N}\mu_n} - \frac{1}{128}\frac{nN\mu_n^2}{\zeta_0(N)^2}\right) \\
&\quad + 4N\exp\left(An^{2r_0} - n\right)
\end{aligned}
$$

We start the analysis with the first term. Consider the case with and $r_0 > 0$. Then the log of the first term takes the form

$$
\begin{aligned}
&An^{2r_0}\log\left(\zeta_0(N)\sqrt{N}/\left(\mu\right)\right) - \frac{1}{128}\frac{n}{\zeta_0(N)N}\mu^2 + \log\, N \\
&= An^{2r_0}\log\left(\frac{N\zeta_0(N)\sqrt{N}n^{2r_0}}{\mu n}\right) - \frac{1}{128}\frac{\mu^2\epsilon_n n}{\zeta_0(N)^2 N} - An^{2r_0}\log\frac{Nn^{2r_0}}{\mu\, n} + \log\, N.
\end{aligned}
$$

If $N\log n/n \to 0$, then one needs that $\frac{n}{\zeta_0(N)^2 N\, n^{2r_0}\log n} \to \infty$ if $r_0 > 0$ and $\frac{n}{\zeta_0(N)^2 N\, \log^2 n} \to \infty$ if $r_0 = 0^+$. Hence the first term is of $o(1)$. This condition is also sufficient for the exponent in the second term become infinitesimal. □

Next we provide a similar result for the case where the conditional moment function is estimated via a kernel estimator. We begin with formulating the requirement on the kernel.

ASSUMPTION 9. *The kernel function $K(\cdot)$ is differentiable single-peaked function. Moreover, it integrates to 1, is bounded and of q-th order, and is square-integrable.*

We formulate the following lemma replicating the result of Lemma 1 for the case of the kernel estimator. For uniformity we rely on Assumption 8(i) that requires the density of covariates to be uniformly bounded away from zero.

LEMMA 2. *Under Assumptions 8 and 9*

$$\sup_{d(h,h_0)=o(1),h\in\mathcal{H}_n} |m(x;h,\widehat{\eta}) - m(x;h,\eta)| = o_p(1)$$

uniformly in x provided that $b_n \to 0$ and $b_n^{-d_z} n^{2r_0-1} \log n \to 0$.

Using Lemmas 1 and 2 we can formulate the consistency result for the directional derivative.

Proof of Lemma 2 Recall the definition of the kernel estimator

$$m(x;h,\widehat{\eta}) = \left(\frac{1}{nb_n^{d_z}}\sum_{i=1}^n K\left(\frac{x-x_i}{b_n}\right)\right)^{-1} \frac{1}{nb_n^{d_z}}\sum_{i=1}^n \rho\left(y_i,h\right) K\left(\frac{x-x_i}{h_n}\right)$$

For the expression of interest, we can consider

$$\frac{\widehat{m}\left(\theta,\eta+\epsilon_n w,z\right)-\widehat{m}\left(\theta,\eta-\epsilon_n w,z\right)}{\epsilon_n} = \left(\frac{1}{nb_n^{d_z}}\sum_{i=1}^n K\left(\frac{z-z_i}{b_n}\right)\right)^{-1}$$
$$\times \frac{1}{nb_n^{d_z}\epsilon_n}\sum_{i=1}^n \left[\rho\left(\theta,\eta+\epsilon_n w,y_i\right)-\rho\left(\theta,\eta-\epsilon_n w,y_i\right)\right] K\left(\frac{z-z_i}{b_n}\right).$$

Then we can consider a class of functions

$$\mathcal{G}_n = \{\rho\left(\cdot,h\right) K\left(\frac{x-\cdot}{b_n}\right),\, h\in\mathcal{H}_n, x\in\mathcal{X}\}.$$

Consider the class $\mathcal{G}_n$. We can represent it as

$$\mathcal{G}_n = \{g = f\kappa \;:\; f\in\mathcal{F}_n,\, \kappa\in\mathcal{F}\}.$$

$N_1(\cdot)$ and $N_2(\cdot)$ correspond to the L_1 and L_2 covering numbers. Consider the covering numbers for classes $\mathcal{F}_n$ and $\mathcal{F}$. We select $\epsilon > 0$, then there exist $m_1 < N_1(\epsilon, \mathcal{F}_n, L_1(Q))$ and $m_2 < N_1(\epsilon, \mathcal{F}, L_1(Q))$ and covers $\{f_j\}_{j=1}^{m_1}$ and $\{\kappa_i\}_{i=1}^{m_2}$ such that for $f \in \mathcal{F}_n$ and $\kappa \in \mathcal{F}$ $\min_j Q|f - f_j| < \epsilon$ and $\min_i Q|\kappa - \kappa_i| < \epsilon$. We note that $|f| \leq C$ and $|g| \leq C$. Consider the cover $\{f_j\kappa_i\}_{j,i=1}^{j=m_1,i=m_2}$ noting that $f_j\kappa_i - f\kappa = (f_j - f)(\kappa_i - \kappa) + f(\kappa_i - \kappa) + \kappa(f_j - f)$. Then, in combination with Cauchy-Schwartz we have that

$$\min_{i,j} Q|\kappa_i f_j - \kappa f| \leq \min_j \left(Q|f_j - f|^2\right)^{1/2} \min_i \left(Q|\kappa_i - \kappa|^2\right)^{1/2} + C \min_j Q|f_j - f| + C \min_i Q|\kappa_i - \kappa|$$

Given the relationship between L_1 and L_2 covering numbers covers $\{f_j\}_{j=1}^{m_1}$ and $\{\kappa_i\}_{i=1}^{m_2}$ are sufficient to achieve $\min_j (Q|f_j - f|^2)^{1/2} < \epsilon$ and $\min_i (Q|\kappa_i - \kappa|^2)^{1/2} < \epsilon$. This means that $\min_{i,j} Q|\kappa_i f_j - \kappa f| < 3C\epsilon$. Thus, the L_1 covering number for $\mathcal{G}_n$ is bounded by a product of L_2 covering numbers for $\mathcal{F}$ and $\mathcal{F}_n$ (which corresponds to the number of elements in the cover $\{f_j\kappa_i\}_{j,i=1}^{j=m_1,i=m_2}$.

Provided that classes $\mathcal{F}_n$ and $\mathcal{F}$ satisfy Euclidean property, we can apply Lemma 2.6.20 from**?**. This means that the class $\mathcal{G}_n$ is Euclidean with parameters depending on n. Provided that $\text{Var}(g) = O(b_n)$ for $g \in \mathcal{G}_n$, we can use a similar logic as in the proof of Theorem 37 in Pollard (2012) with the results similar to those in the proof of Lemma 1. This leads to condition $\frac{nb_n^{dz}}{n^{2r_0}\log n} \to \infty$. We note that the bias due to kernel smoothing $E[m(X_i; h, \widehat{\eta})|X_i = x] = O(b_n^m)$, where m is the order of the kernel, and the bias due to the sieve approximation is $n^{-\phi}$. Then

$$\left\| L_{1,p}^{\epsilon_n,w} E\left[\widehat{m}(\theta, \eta, Z_i) | Z_i = z\right] - L_{1,p}^{\epsilon_n,w} m(\theta, \eta, z) \right\| = O\left(b_n^m + n^{\phi}\right),$$

which converges to zero if $b_n^{-m} \to \infty$ and $n^{\phi} \to \infty$. □

Provided the unform consistency result, we can replicate our result for the finite-dimensional parameter that is estimated using partitioning via honest trees.

THEOREM 3. *Suppose that $\widehat{L}$ is the leaf of honest $\{\alpha, k\}$-valid classi*

cation tree that returns label k and $\widehat{M}_{\widehat{L}}(h, \widehat{\eta}) = o(1)$ where $\widehat{M}_{\widehat{L}}(\cdot, \cdot)$ is estimated from the subsample that was not used for splitting. Then whenever $k = O(n/V)$ under Assumptions 1-5 and conditions of either Lemma 1 or Lemma 2

$$d(\widehat{h}, h^k) = o_p(1).$$

5.2 Semiparametric model with growing tree structure

Our next extension is based on considering the model that we analyzed before but now allow the tree to partition the support of covariates such that the number of elements in the partition increases as the sample size grows.

The idea of this extension in quite straightforward. Given that Assumption 4 holds for any decreasing sample size sequence $n \to \infty$, it would also work for any of its subsequence. The expected number of points that falls in the leaf of volume V in the hypercube $[0,1]^{d_z}$ can be minorized as $n\,V\,\inf_z f_Z(z)$. As a result, given that k splits produce volumes of at least 2^{-k}, then the sufficient condition that yields the uniform consistency is that $n\,2^{-k} \to \infty$. That is guaranteed whenever $k \ll \log n$.

6 Monte Carlo Evidence

To showcase the performance of our estimator, we conduct several Monte Carlo experiments. We first demonstrate the ability of the estimator to successfully identify heterogeneous treatment effects in an experimental setting. We then consider the case of a regression discontinuity design (RDD). We anticipate that these two settings will be fruitful applications of our framework, and our Monte Carlo is designed to highlight the strengths of our approach while also illustrating potential tradeoffs that a practitioner faces in real settings.

6.1 Monte Carlo: RCT

We consider the following data-generating process which mimics a typical randomized controlled trial (RCT) design. Let the outcome variable be defined as:

$$Y = \tau(X) \cdot W + f(X, \beta) + \epsilon, \tag{6.7}$$

where W is an indicator for treatment, X is a vector of observable covariates, and ϵ is an idiosyncratic, normally-distributed shock with mean zero and unit variance. The object of interest is $\tau(X)$, the true treatment effect, which may be a function of the observables, X. We initially draw two discrete X variables that are uniformly distributed over the integers from 1 to 8; this generates 64 distinct subgroups. We consider several specifications for $\tau(X)$ in increasing complexity. In the simplest RCT setting, W is randomly assigned independent of X. We draw a uniform random variable and set W to one when the draw is greater than one-half and zero otherwise.

Table 1: Monte Carlo: Uniform RCT

		Tree		OLS	
numObs / $(\alpha, k, \overline{MSE})$	Num Leafs	Dim(τ)	MSPE	Dim(τ)	MSPE
50	1.000	64.000	0.275	0.000	NaN
(0.01, 13, 1e-07)	(0.000)	(0.000)	(0.221)	(0.000)	(NaN)
100	1.000	64.000	0.158	0.250	NaN
(0.01, 28, 1e-07)	(0.000)	(0.000)	(0.152)	(0.433)	(NaN)
200	1.000	64.000	0.113	3.250	0.413
(0.01, 55, 1e-07)	(0.000)	(0.000)	(0.050)	(1.090)	(0.181)
400	1.000	64.000	0.074	22.875	0.556
(0.01, 100, 1e-07)	(0.000)	(0.000)	(0.055)	(3.370)	(0.074)
800	1.000	64.000	0.049	57.625	0.415
(0.01, 211, 1e-07)	(0.000)	(0.000)	(0.037)	(1.932)	(0.029)
1600	1.000	64.000	0.053	64.000	0.287
(0.01, 411, 1e-07)	(0.000)	(0.000)	(0.031)	(0.000)	(0.017)
3200	1.000	64.000	0.031	64.000	0.212
(0.01, 811, 1e-07)	(0.000)	(0.000)	(0.030)	(0.000)	(0.015)
6400	1.000	64.000	0.028	64.000	0.152
(0.01, 1611, 1e-07)	(0.000)	(0.000)	(0.012)	(0.000)	(0.012)

For sake of comparison, we start with the simplest possible case: the treatment effect is equal to ten for all treated units, and zero otherwise, generating a single treatment effect. We highlight two features of the results, shown in Table 1. First, the estimator assigns a single treatment effect at all sample sizes for all Monte Carlo runs, consistently recovering the true underlying model. Column 3 reports the count of all subgroups that are assigned a statistically significant treatment effect at the five percent level; here the tree finds significant effects for all 64 subgroups. We have assigned a k, α, and $\overline{MSE}$ through cross-validation against a holdout sample.[5] Of these parameters, α is always set at the corner solution of $\alpha = 0.01$, while k and $\overline{MSE}$ become increasing stringent. Second, the calculation of the root mean squared prediction error (MSPE) is a useful baseline to compare following (more complex) models against. Here, the MSPE reflects only the statistical sampling error, whereas the more complex models we consider next have a convolution of statistical sampling and model misspecification.

We also report the performance of OLS estimates run on each subgroup separately in the

[5] k is the minimum number of observations in each leaf. α is the minimum proportion of data in each leaf. $\overline{MSE}$ is the minimum improvement in MSE after each split.

Table 2: Monte Carlo: Group RCT

		Tree		OLS	
numObs / $(\alpha, k, \overline{MSE})$	Num Leafs	Dim(τ)	MSPE	Dim(τ)	MSPE
50	12.875	6.625	0.971	0.000	NaN
(0.01, 1, 1e-01)	(1.536)	(2.595)	(0.276)	(0.000)	(NaN)
100	8.750	6.250	0.476	0.125	NaN
(0.01, 1, 1e-01)	(6.036)	(2.107)	(0.258)	(0.331)	(NaN)
200	2.125	8.000	0.178	0.625	0.413
(0.01, 7, 1e-01)	(0.331)	(0.000)	(0.072)	(0.696)	(0.181)
400	2.125	8.000	0.129	7.500	0.556
(0.01, 12, 1e-01)	(0.331)	(0.000)	(0.051)	(2.000)	(0.074)
800	2.000	8.000	0.080	15.125	0.415
(0.01, 1, 1e-01)	(0.000)	(0.000)	(0.041)	(2.619)	(0.029)
1600	2.000	8.000	0.069	16.625	0.287
(0.01, 1, 1e-01)	(0.000)	(0.000)	(0.028)	(2.997)	(0.017)
3200	2.000	8.000	0.042	18.000	0.212
(0.01, 163, 1e-02)	(0.000)	(0.000)	(0.027)	(2.828)	(0.015)
6400	2.000	8.000	0.036	18.125	0.152
(0.01, 323, 1e-02)	(0.000)	(0.000)	(0.009)	(2.803)	(0.012)
12800	2.000	8.000	0.019	17.250	0.099
(0.01, 1, 1e-02)	(0.000)	(0.000)	(0.011)	(2.487)	(0.007)

last two columns. The estimator is badly biased at smaller samples, failing to find even a single statistically significant treatment effect. The difference between the two estimators is driven by the fact that the tree can group together observations from different X, while the OLS estimator is forced to estimate separately on each subgroup. At higher sample sizes, the OLS estimator is able to recover the true number of effects, but has a large precision penalty as it is unable to group together similar observations to improve the standard error. This highlights one benefit of using the tree method even when the true model is a single treatment effect.

To assess the performance of the estimator when we introduce observable heterogeneity, we set the treatment effect to ten if the observation has $x_1 = 1$, and zero otherwise, generating two treatment effects. Table 2 reports the results. Initially, the tree estimates too many splits, producing too few statistically significant treatment effects (there are 8 groups for which the treatment effect is non-zero). At the highest sample sizes, the estimator reconciles this error, finding the true model and estimating the treatment effects precisely. The decrease in the MSPE reflects this, as the rate returns to a parametric rate once the true model has

Table 3: Monte Carlo: Sparse RCT

		Tree		OLS	
numObs / $(\alpha, k, \overline{MSE})$	Num Leafs	Dim(τ)	MSPE	Dim(τ)	MSPE
50	1.000	0.750	1.320	0.000	NaN
(0.01, 13, 1e-07)	(0.000)	(0.433)	(0.075)	(0.000)	(NaN)
100	5.750	0.625	1.119	0.125	NaN
(0.01, 1, 1e-01)	(5.379)	(0.484)	(0.275)	(0.331)	(NaN)
200	2.375	0.625	1.035	0.250	0.413
(0.01, 1, 1e-01)	(1.798)	(0.484)	(0.469)	(0.433)	(0.181)
400	51.500	1.000	0.784	5.125	0.556
(0.01, 1, 1e-02)	(4.822)	(0.000)	(0.048)	(1.364)	(0.074)
800	32.250	1.000	0.435	10.000	0.415
(0.01, 1, 1e-02)	(9.588)	(0.000)	(0.089)	(2.693)	(0.029)
1600	6.375	1.000	0.146	10.750	0.287
(0.01, 1, 1e-02)	(0.992)	(0.000)	(0.034)	(3.382)	(0.017)
3200	5.250	1.000	0.078	12.500	0.212
(0.01, 1, 1e-02)	(1.392)	(0.000)	(0.017)	(2.693)	(0.015)
6400	3.875	1.000	0.051	12.250	0.152
(0.01, 1, 1e-02)	(0.927)	(0.000)	(0.020)	(3.192)	(0.012)
12800	3.000	1.000	0.022	11.625	0.099
(0.01, 1, 1e-02)	(0.000)	(0.000)	(0.011)	(2.643)	(0.007)
25600	3.000	1.000	0.025	10.500	0.070
(0.01, 1, 1e-02)	(0.000)	(0.000)	(0.008)	(2.915)	(0.003)

been recovered. Compared to the baseline case of a treatment effect without any observable heterogeneity, the standard errors are approximately twice as large. OLS is biased upward and has much large standard errors, particularly at larger sample sizes.

We next consider a case of a sparse treatment effect, where $\tau(X) = 10$ if and only if $x_1 = 1$ and $x_2 = 1$. Otherwise, $\tau(X) = 0$. This is a challenging specification for the estimator, as there are 63 null treatment effects which may appear to be true effects due to within-group statistical errors. Table 3 reports the results for 500 replications.

There are several notable features. First, the classification tree has a downward bias on the estimated number of treatment effects for the smallest sample sizes. This results from the optimal tradeoff of variance (too few observations in each leaf) versus bias (not enough partitions of the data to capture the true number of effects). At sample sizes of $n = 400$ and above, the tree grows more complex and converges to finding one statistically significant treatment effect. The faster-than-parametric decrease in the MSPE at that threshold reflects

Table 4: Monte Carlo: RCT with Saturated Sub-Group Heterogeneity

		Tree		OLS	
numObs / $(\alpha, k, \overline{MSE})$	Num Leafs	Dim(τ)	MSPE	Dim(τ)	MSPE
50	14.160	45.816	12.218	0.006	NaN
(0.01, 1, 1e-01)	(1.749)	(7.166)	(2.798)	(0.077)	(NaN)
100	25.468	47.866	7.198	0.114	NaN
(0.01, 1, 1e-01)	(2.273)	(4.976)	(2.398)	(0.336)	(NaN)
200	42.696	51.388	3.228	2.732	NaN
(0.01, 1, 1e-07)	(2.528)	(3.151)	(1.161)	(1.501)	(NaN)
400	58.564	58.590	1.348	22.960	0.718
(0.01, 1, 1e-03)	(1.964)	(2.073)	(0.295)	(2.688)	(0.107)
800	63.752	63.124	0.648	56.918	0.591
(0.01, 1, 1e-07)	(0.496)	(0.738)	(0.091)	(2.287)	(0.058)
1600	64.000	63.700	0.415	63.646	0.415
(0.01, 1, 1e-07)	(0.000)	(0.458)	(0.036)	(0.522)	(0.036)
3200	64.000	63.932	0.288	63.932	0.288
(0.01, 1, 1e-07)	(0.000)	(0.252)	(0.026)	(0.252)	(0.026)
6400	64.000	63.994	0.201	63.994	0.201
(0.01, 1, 1e-07)	(0.000)	(0.077)	(0.018)	(0.077)	(0.018)
12800	64.000	64.000	0.142	64.000	0.142
(0.01, 1, 1e-07)	(0.000)	(0.000)	(0.012)	(0.000)	(0.012)
25600	64.000	64.000	0.100	64.000	0.100
(0.01, 1, 1e-07)	(0.000)	(0.000)	(0.009)	(0.000)	(0.009)

the decline in specification error. It is instructive to contrast the performance of the tree against the naive OLS estimates. At the smallest sample sizes, the OLS estimator does not find any statistically significant treatment effects. As the sample size grows, the OLS finds an increasing number of statistically significant treatment effects. Even after the tree has converged to the true model, the OLS estimator continues to overestimate the number of treatment effects. Aside from its bias of estimating ten times as many treatment effects as the truth, the OLS estimator interestingly has lower prediction error for sample sizes of 200, 400, and 800. At higher sample sizes, OLS continues to be biased and is also dominated on prediction error by the classification tree. The relatively poor MSPE of OLS reflects the fact that the estimator cannot group together observations to improve the precision of the estimated treatment effect.

To consider a more complex case, we modify the data-generating process for $\tau(X)$ to be:

$$\tau(X) = x_1(1 + (x_2 - 1)). \tag{6.8}$$

This results in 64 treatment effects as a combination of x_1 and x_2. Table 4 reports the results. We note that the optimal $k = 1$, which was set as the lower bound in the cross validation search, for all sample sizes. This is driven by two factors. First, setting k higher makes it mechanically impossible to cut the data enough times to reproduce the number of true treatment effects. For example, when $k = 25$, the sample size must be at least $n = 25 \cdot 64 = 1600$ before the tree could even potentially match the true set of underlying treatment effects. Second, all possible interactions of the two dummy variables have true treatment effects, so this design will not experience an over-fitting problem. The optimal size of the tree is controlled here by the acceptance criterion, which becomes more lax as the sample size grows.

The inability of the tree to grow complex enough for smaller sample sizes is reflected in the mean squared prediction error. The MSPE exhibits monotonic but highly nonlinear convergence. Once the threshold of $n = 1600$ is reached, the tree recovers the true underlying structure and MSPE drops discontinuously. Parametric error rates obtain after that point, reflecting the independence of the tree estimation and the estimation of treatment effects, as desired.

The OLS estimator in this case has a performance as good or better than the tree approach for all sample sizes for which it reports prediction error. This is expected, as the OLS estimator in this case is the true model. However, once the tree has found the true model, prediction errors are (mechanically) identical, which highlights the independence of the honest tree's predictive performance from the model selection step.

6.2 Monte Carlo: RDD

Our second set of Monte Carlo experiments uses a regression discontinuity design (RDD). RDD works by leveraging some known threshold, c, on a so-called *running variable* which functions as an assignment mechanism: to the left of the threshold, units do not receive a treatment, while those to the right of the threshold do. Assuming that units cannot manipulate their running variable, the discontinuous treatment on either side of the threshold can be used to estimate the causal effect of a treatment on outcomes, as sorting into the

control or treatment groups is as "good as random" under the maintained assumption. Examples of RDD settings include the assignment of educational treatment on the basis of test scores, and means-tested assignments of welfare, unemployment insurance, and disability programs on labor supply.

While the RDD setting has broad empirical appeal as a method for obtaining "credible" estimates of causal effects, the researcher still has to make a number of important assumptions. Among those assumptions are classifying units into different groups where the researcher may think that treatment effects vary. For example, the treatment effects of magnet schools on student achievements may vary in size depending on the income of the student's parents. For low-income students, the effects may be much larger than for high-income students. The researcher may split the sample into two groups and estimate separate RDD regressions on each group, producing two treatment effects. In general, this search of the model specification process will fail for the reasons discussed above. Our Monte Carlo illustrates how the present estimator can circumvent this problem by constructing a set of splits of the data without intervention of the researcher. A second holdout sample is then used to produce consistent estimates of the treatment effect within each sub-group.

We modify the above data-generating process by augmenting the experimental treatment to be a function of a running variable:

$$Y = \tau(X) \cdot W(R) + f(X, \beta) + \epsilon, \tag{6.9}$$

where $W(R)$ is now an indicator function that is equal to zero to left of a cutoff value, $\bar{R}$ and one to the right:

$$W = \begin{cases} 0 & \text{if } R < \bar{R}, \\ 1 & \text{else.} \end{cases} \tag{6.10}$$

This generates a sharp RDD, as opposed to a fuzzy RDD where the probability of treatment is positive everywhere but jumps discontinuously at $\bar{R}$. We draw R from uniform $U[0, 1]$. The object of interest is $\tau(X)$, the treatment effect as a function of the vector of covariates. We allow the treatment effect to depend on three covariates as follows:

$$\tau(X) = \begin{cases} 5 & \text{if } X_2 < 0.67, \\ -2 & \text{else.} \end{cases} \tag{6.11}$$

We augment the treatment effect by subtracting 2 if $X_3 = 1$ and adding 5 if $X_3 = 2$. This

generates six total treatment effects across the covariate space.

The problem facing the econometrician is deciding where to assign different treatment effects. It is possible that one could guess the data-generating process above, but it is both unlikely and statistically undesirable for the reasons outlined above. Our estimator circumvents this process by estimating the partitioning of the X space in a first stage. In a second stage, we estimate treatment effects using the standard RDD approach outlined in Imbens and Lemieux (2008) and Lee and Lemieux (2010), using a local-linear regression around the threshold. We control the window width around the threshold using cross-validation, and we report results for various choices of that window width below.

We generate 500 draws of each sample size. We report an out-of-sample mean squared prediction error for each sample size, which we constructed by generating data using the true (known) generating process and using the estimated tree and associated RDD models to compute predicted treatment effects. We then sum the squared difference from the true value, divide by sample size, and take the square root. Table 5 shows the main results for the model above when using all the data in sample on either side of the window ($h = 0.5$) and the threshold for improvement in the tree is set to 0.1.

Search over optimal K: show that optimal K chosen via cross-validation grows in the sample size. This is the tradeoff between variance and bias. In small samples, need to balance against missing features of the DGP versus overfitting in larger samples. Increase K does that.

First, we note that the model obtains consistent estimates of the number of treatment effects (true value: 6), the number of discrete splits (true value: 3), the number of continuous splits (true value: 2), and the level at which the second covariate, X_2, splits the sample (true value: 0.670). This convergence to the true model is rapid—at the sample size of 16,000 there is no appreciable variation across Monte Carlo experiments in the structure of the estimated tree. At that point, the estimator essentially recovers the true tree without error, as would be expected given the faster-than-parametric rate of convergence of the first stage of our estimation. The column labeled RMSPE reports the root mean squared prediction error on out-of-sample data. The RMSPE is the composition of two sources of error: errors in the specification of the classification tree, and errors arising from sampling error within each leaf. For smaller sample sizes, the rate of decline in the RMSPE is driven by both errors. As the tree converges at faster-than-parametric rates, so does the prediction error. Beyond $n = 8000$, when the tree is recovered with negligible error, the RMSPE is almost

Table 5: Monte Carlo: RDD

numObs	Dim(τ)	Count Discrete	Count Continuous	Mean X2	MSPE	Pr(Null)
500	6.587	3.207	2.380	0.539	0.918	0.000
	(0.785)	(0.480)	(0.772)	(0.252)	(0.393)	(0.000)
1000	6.107	3.113	1.993	0.671	0.669	0.000
	(0.449)	(0.317)	(0.337)	(0.027)	(0.303)	(0.000)
2000	6.133	3.087	2.047	0.675	0.497	0.000
	(0.499)	(0.281)	(0.291)	(0.026)	(0.273)	(0.000)
4000	6.040	3.020	2.020	0.671	0.321	0.000
	(0.280)	(0.140)	(0.140)	(0.014)	(0.187)	(0.000)
8000	6.027	3.013	2.013	0.670	0.246	0.000
	(0.229)	(0.115)	(0.115)	(0.013)	(0.213)	(0.000)
16000	6.000	3.000	2.000	0.669	0.198	0.000
	(0.000)	(0.000)	(0.000)	(0.006)	(0.174)	(0.000)

completely due to classical sampling error. At that point, the rate of convergence reverts to the parametric $\sqrt{n}$ rate. Finally, in the last column we report out-of-sample observations for which the tree is unable to produce estimates due to the starvation of a given leaf in the second stage of our estimation process. At small samples, there are some observations which cannot be predicted, but this is a vanishingly small problem that disappears completely at sample sizes beyond $n = 1000$.

The smaller sample sizes show some regularities, particularly with respect to bias. For one, our procedure tends to estimate too many treatment effects at smaller sample sizes, primarily of the continuous variety. This also introduced bias in the estimate of where X_2 is split. These biases disappear rapidly as the sample size grows.

6.3 Continuous Treatment Effects

An extension of our econometric results above considers the case where K is infinite. To demonstrate the small sample performance of our estimator in such a setting, we perform a Monte Carlo experiment in a univariate RDD setup with the following function for the treatment effect:

$$\tau(x_i) = \sin(4\pi x_i), \tag{6.12}$$

where x_i is a unidimensional covariate distributed uniformly on the unit interval. As before, we generate a $U[0, 1]$ running variable and assign the treatment if the running variable is

Table 6: Continuous Treatment Effects

	Without Error				With Error			
	Uniform x		Normal x		Uniform x		Normal x	
n	Dim(τ)	RMSPE	Dim(τ)	RMSPE	Dim(τ)	RMSPE	Dim(τ)	RMSPE
2000	15.548	0.137	19.560	0.148	11.942	0.373	7.282	0.360
	(0.976)	(0.006)	(0.697)	(0.010)	(0.755)	(0.052)	(0.599)	(0.046)
4000	24.520	0.090	37.156	0.081	12.222	0.298	14.190	0.292
	(1.044)	(0.004)	(1.254)	(0.004)	(0.667)	(0.033)	(0.806)	(0.038)
8000	30.332	0.071	64.148	0.046	16.880	0.361	14.306	0.235
	(0.823)	(0.002)	(1.353)	(0.001)	(0.840)	(0.033)	(0.770)	(0.024)
16000	40.244	0.057	113.440	0.028	24.764	0.193	22.666	0.197
	(1.339)	(0.002)	(1.905)	(0.002)	(1.168)	(0.019)	(1.157)	(0.019)

above one half. We estimate by splitting the sample in half, first fitting the tree on the first sample, and then fitting the estimates within each leaf on the second sample. We impose that $\alpha = 0.1$ and choose the minimum number of observations in each node via cross-validation. This guards against the possibility of growing the number of splits faster than the number of observations, which by extension ensures that each leaf will have an infinite number of observations in the limit, while also balancing finite-sample bias and variance.

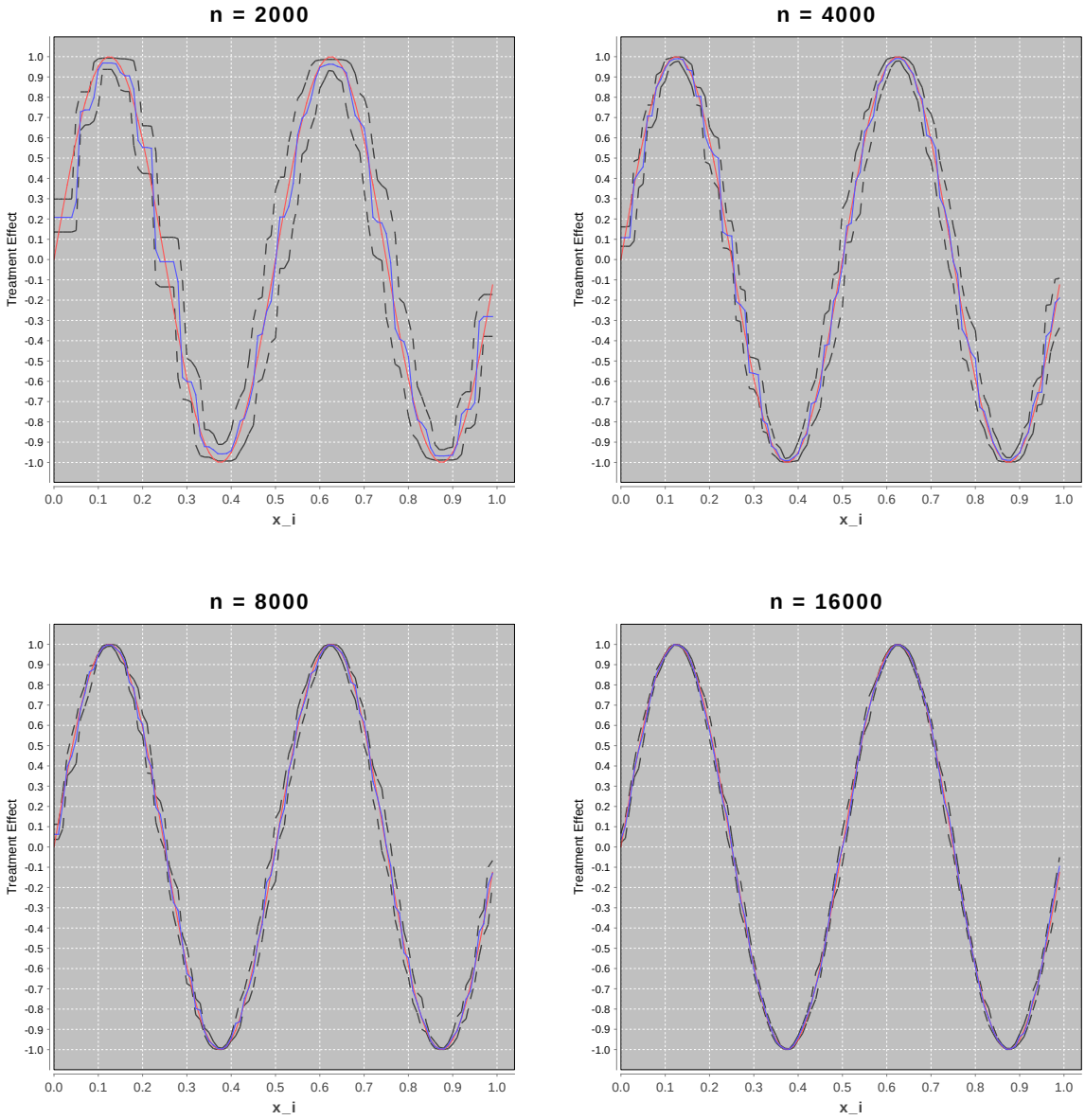
6.4 No Error Term

We begin by running our Monte Carlos with the variance of the idiosyncratic term set to zero. This captures the effect of pure approximation error.

Table 6 reports the results from this experiment. In all cases, the model did not estimate any nulls, so we omit that column. As the dimension of τ shows, the model fits an increasingly complex model to the data. This also results in a substantial decrease in mean squared prediction error.

Figure 1 shows the fit of the moment tree in this case. First, the general fit is excellent across the entire range of the function. There is a small bias evident at the peaks and troughs of the sine function, where the derivative is near zero. In smaller samples, the estimator fits a constant to these neighborhoods, which leads to some minor underfitting. This bias disappears in the large samples. By $n = 16000$, the underlying function is recovered uniformly and with nearly no variance.

Figure 1: Estimated and True Treatment Effect Function, Without Error



6.5 With Measurement Error

We now allow the error term to be drawn from a standard normal. The right panel of the table shows the results. The trees are simpler in this case, as the estimator has to balance variance against bias. The RMSPE is substantially larger, although it rapidly shrinks at higher sample sizes. Figure 2 shows the resulting estimated function across the domain of X.

6.6 Normally-Distributed Data

In this section, we show that the method works well even when data is not distributed uniformly across the domain of interest. We draw x from a normal distribution with mean one-half and standard deviation equal to 0.25, and truncate at zero and one, we can observe the effect of having non-uniformly distributed data across the interval. Figure 3 plots the estimated functions and the 95 percent confidence bands generated over 500 Monte Carlo iterations. It is immediately apparent that the estimator is best at capturing the variation in the underlying treatment effect function where the data is most frequent. The two tails have more constant approximations, which hones in on the true function rapidly as the sample size increases. This result gives confidence that the method is still able to consistently and accurately recover the true function in reasonably-sized data sets, even when the data density is unevenly distributed.

7 Empirical Application

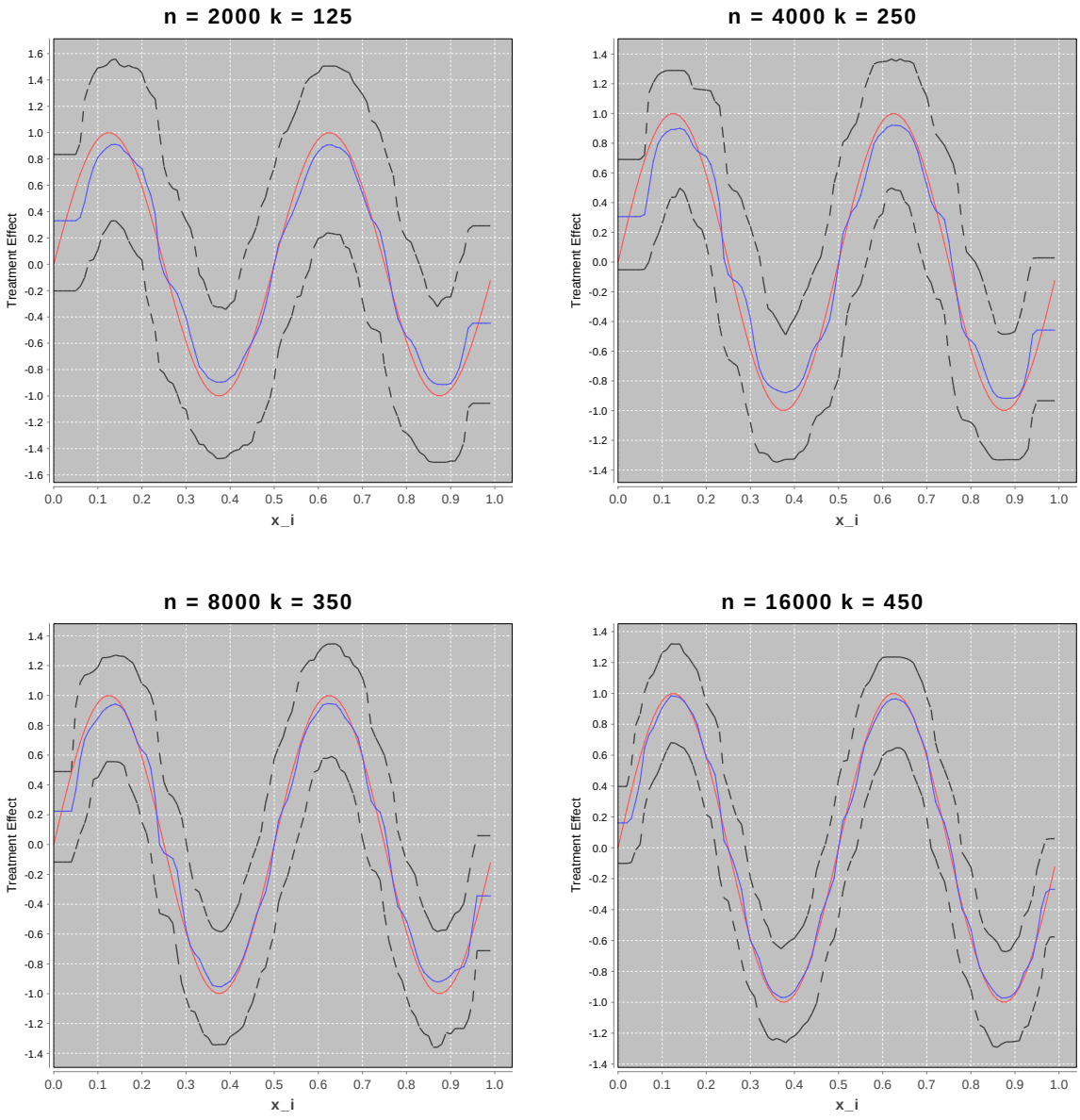
The Pradhan Mantri Gram Sadak Yojana (PMGSY)—the Prime Minister's Village Road Program—was launched in 2000 with the goal of providing all-weather access to unconnected habitations across India.[6] The focus was on the provision of new feeder roads to localities that did not have paved roads. By 2015, the government had built over 100,000 roads to over 185,000 villages at a cost of nearly $40 billion.[7]

National guidelines determine prioritization of road construction under the PMGSY. Most importantly for this empirical exercise, road construction is supposed to occur first in large localities, as defined by the 2001 Population Census. Program rules dictate that villages of

[6]Habitations are defined as clusters of population whose location does not change over time. They are distinct from, but form parts of, villages as defined by the Population Census. In this paper, we aggregate all data to the level of the census village.

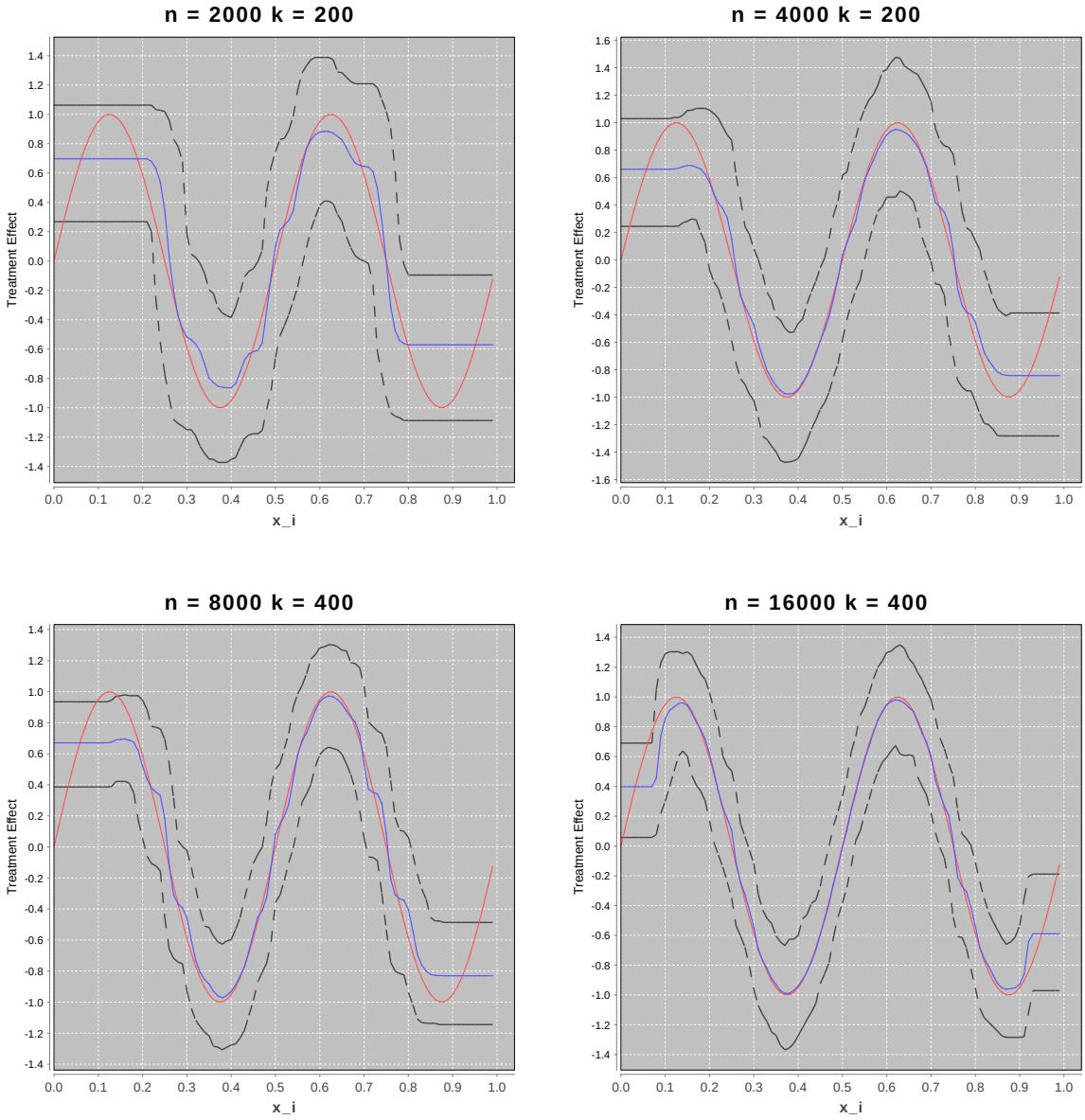
[7]For a more comprehensive description of the program, see Asher and Novosad (2016).

Figure 2: Estimated and True Treatment Effect Function, With Error and Optimal k



Notes: Each figure plots the mean estimated and true treatment effect function, $\tau(x_i)$, for various sample sizes. The minimum number of observations in each leaf, k, was chosen via cross-validation. The data-generating process is a regression discontinuity design with uniformly-distributed x_i. The dashed lines represent the 95 percent confidence interval. Results computed using 500 Monte Carlo experiments.

Figure 3: Estimated and True Treatment Effect Function, Normally-Distributed Data



Notes: Each figure plots the mean estimated and true treatment effect function, $\tau(x_i)$, for various sample sizes. The minimum number of observations in each leaf, k, was chosen via cross-validation. The data-generating process is a regression discontinuity design with truncated normal-distributed x_i with mean 0.5 and standard deviation 0.25. The dashed lines represent the 95 percent confidence interval. Results computed using 500 Monte Carlo experiments.

1000+ population were to be prioritized to villages in the population range of 500-999, which were in turn to be prioritized over smaller villages. These rules create discontinuities in the probability of road construction by 2011, the year for which we have outcome data from the 2011 Population Census. Villages with baseline (2001) populations to the right of the population cutoffs (500 and 1000) are approximately fifteen percentage points more likely to have received a road by 2011. We exploit these discontinuous jumps in the probability of road construction to estimate the impact of the PMGSY.

For this application, we focus on the question of the impact of road construction on the provision of public transportation. Specifically, we estimate the impact of PMGSY road construction on the likelihood that a village will be served by a regular bus route. Recent research has suggested that rural demand may not be sufficient to support the provision of bus services Raballand, Thornton, Yang, Goldberg, Keleher, and Müller (2011). This finding raises questions of whether road construction alone will be sufficient to expand economic opportunities in poor and low density rural areas. Bryan, Chowdhury, and Mobarak (2014) find large returns to subsidizing bus travel to nearby cities, so much so that their intervention is being scaled up into a major anti-poverty program in Bangladesh. Given the high cost of road construction, a better understanding of the conditions under which road provision leads to an expansion of actual transportation options would help policymakers maximize the impact of infrastructure investments. Our analysis builds upon Asher and Novosad (2016), who find evidence that road construction does lead to expansion of transportation services, but only in areas close to cities.

7.1 Data

To utilize village-level variation in road construction, we construct a high spatial resolution dataset that combines household and firm microdata with village aggregates describing amenities, infrastructure and demographic information.

Data on road construction (the treatment) come from the official online administrative records of the PMGSY. [8] For the purposes of our analysis, all variables are aggregated to the level of the census village, the geographic unit at which we measure outcomes. We consider a village to be treated by the PMGSY if at least one habitation in the village received a completed PMGSY road by 2010, the year before the most recent round of the

[8] All data are publicly available at http://omms.nic.in. The variables used in this paper were assembled from data scraped in January 2015.

Population Census.

The primary outcome of interest, the presence of regular bus service to a village, comes from the 2011 Population Census. Baseline village characteristics come from the 2001 Population Census, which was collected in the year that the first PMGSY roads were being constructed. These censuses contains village-level information on both demographic characteristics (population, social groups, literacy rates, etc) and local amenities (schools, medical centers, electrification, etc). We use these variables as controls, sources of heterogeneity and, in the case of village population, the running variable in our regression discontinuity design (described below). We also make use of GIS data: village and town latitude and longitude from ML Infomap allows us to generate straight line distances from villages to cities as a measure of urban market access.

7.2 Empirical Strategy

The endogeneity of road placement makes it difficult to assess the impacts of rural roads. We overcome this challenge by taking advantage of program guidelines that generate discontinuities in the probability of road construction by 2011 at two village population thresholds (500 and 1000). We exploit these population thresholds to estimate the economic impact of rural roads using a fuzzy regression discontinuity design, effectively comparing similar villages on either side of these population thresholds.

Under the assumption of continuity at the treatment threshold, the fuzzy RD estimator Imbens and Lemieux (2008) estimates the local average treatment effect (LATE) of receiving a new road, for a village with population equal to the threshold:

$$\tau = \frac{\lim_{pop \to T^+} \mathbb{E}[Y_v | pop_v = T] - \lim_{pop \to T^-} \mathbb{E}[Y_v | pop_v = T]}{\lim_{pop \to T^+} \mathbb{E}[newroad_v | pop_v = T] - \lim_{pop \to T^-} \mathbb{E}[newroad_v | pop_v = T]}, \tag{7.13}$$

where pop_v is the baseline village population, T is the threshold population, and $newroad_v$ is an indicator variable for whether village v received a new road in the sample period. The treatment effect can be interpreted as the discontinuous change in the outcome variable at the population threshold (the numerator) divided by the discontinuous change in the probability of treatment (the denominator). The local average treatment effect (LATE) estimated by our empirical design is specific to the complier set, namely those villages whose treatment status would be zero with population below the threshold and one with population above.

Our estimation follows the recommendations of Imbens and Lemieux (2008), Imbens and

Wooldridge (2009) and Gelman and Imbens (2014). Our preferred specification uses local linear regression to control for the running variable (village population) on either side of the threshold. We restrict our sample to those villages whose population is within a certain bandwidth around the threshold, formally $pop_v \in [T-h; T+h]$, where h is the value of the bandwidth around threshold T. We calculate an optimal bandwidth of 54 following Imbens and Wooldridge (2009) and use a triangular kernel that places the most weight on observations close to the cutoff, as in Dell (2015). Controls and fixed effects are not necessary for identification, but their inclusion increases the efficiency of the estimator.

We begin by estimating the following reduced form fuzzy RDD specification:

$$Y_{v,j} = \beta_0 + \beta_1 1\{pop_{v,j} \geq T\} + \beta_2 pop_{v,j} + \beta_3 pop_{v,j} * 1\{pop_{v,j} \geq T\} + \zeta X_{v,j} + \eta_j + \epsilon_{v,j}, \quad (7.14)$$

where $Y_{v,j}$ is the outcome of interest, T is the population threshold, $pop_{v,j}$ is baseline village population, $X_{v,j}$ is a vector of village controls measured at baseline, and η_j is a group fixed effect. Village controls and fixed effects are not necessary for identification but improve the efficiency of the estimation. The change in outcome $Y_{v,j}$ for a village at the population threshold T is captured by $\beta_1 + \beta_3 * T$. For ease of exposition, we subtract the threshold value T from the population variable, such that $T = 0$, and β_1 fully describes the change in outcome $Y_{v,j}$ at the treatment threshold.

We make the following choices when estimating this model. In the first stage regression, in which we estimate the change in the probability of treatment, $Y_{v,j}$ is a dummy variable that takes on the value one if the village has received a PMGSY road before 2011, the year of our primary outcome data.[9] For regressions in which we estimate the reduced form effect of road prioritization (i.e. being to the right of the population threshold) on economic outcomes, we discuss the definition of outcome variables as we present the results in Section 7.3. The vector of village controls, $X_{v,j}$, and fixed effects, η_j, are discussed below.

We understand the reduced form effect of road priority to be treatment effect of a new road times the change in the probability of road treatment at the population threshold. To estimate the treatment effect directly, we use the following fuzzy RDD specification in which

[9]This is the year that most data was collected for the SECC. When estimating outcomes measured in a different year, such as in the Population Census, we use the appropriate year of measurement for that particular set of regressions.

we instrument for treatment ($newroad_{v,j}$) with our road priority dummy $1\{pop_{v,j} \geq T\}$.

$$Y_{v,j} = \gamma_0 + \gamma_1 newroad_{v,j} + \gamma_2 pop_{v,j} + \gamma_3 pop_{v,j} * 1\{pop_{v,j} \geq T\} + \zeta X_{v,j} + \eta_j + \upsilon_{v,j}. \quad (7.15)$$

We estimate this equation using two stage least squares, where the first stage comes from Equation 7.14, with $newroad_{v,j}$ as the dependent variable.

As the objective of this paper is to estimate the impact of receiving a paved road for the first time, we restrict our sample to villages that did not have a paved road at the start of the program.[10] The PMGSY used multiple population thresholds to determine road prioritization: 1000, 500 and 250. Very few villages around the 250 population threshold received roads by 2012, so we limit our sample to villages with populations close to 500 and 1000. Further, only certain states followed the population threshold prioritization rules as given by the national guidelines of the PMGSY. We worked closely with the National Rural Roads Development Agency to identify the state-specific thresholds that were followed and define our sample accordingly. Our sample is comprised of villages from the following states, with the population thresholds used in parentheses: Chhattisgarh (500, 1000), Madhya Pradesh (500, 1000), Orissa (500, 1000), and Rajasthan (500). To maximize power, we pool our samples, using the same optimal bandwidth (54) for villages close to the 500 and 1000 thresholds.

7.3 Results

We first present our baseline RDD estimates before turning to heterogeneous treatment effects using our two-step classification tree method above. As the RDD estimator can be interpreted as a Wald estimator, we report the first stage in Table 7.

There is a very strong relationship, as expected, between the instrument (being above the pre-defined population threshold) and having a road. The second stage is reported in Table 8. Our baseline estimates suggest that the causal effect of a newly-built road to a rural village is to increase bus availability by 17 percentage points. The estimated effect is nominally significant at the 10 percent level. That may be due to either sampling error, a lot of heterogeneity in the correctly specified model, or specification error. To assess if we can improve on this estimates by accounting for unobserved heterogeneity, we next run our estimator on the same data.

[10] We define our sample of unconnected villages to be those that were recorded as lacking a paved road in the PMGSY administrative data.

Table 7: First Stage RDD

	(1) Road by 2011
Road priority	0.196***
	(0.0185)
2001 Pop * 1(Pop < Cutoff)	0.0158
	(0.0422)
2001 Pop * 1(Pop ≥ Cutoff)	0.0797
	(0.0422)
Constant	0.262***
	(0.0132)
Observations	10086
F	188.9

Standard errors in parentheses

Table 8: Baseline RDD Estimates

	(1) Bus service (2011)
Road	0.171*
	(0.0862)
2001 Pop * 1(Pop < Cutoff)	−0.0172
	(0.0394)
2001 Pop * 1(Pop ≥ Cutoff)	−0.00335
	(0.0430)
Constant	0.183***
	(0.0324)
Observations	10086

Standard errors in parentheses

We include several covariates as possible splitting variables: state fixed effects; distance to the nearest city with 10,000 people, 100,000 people, and 500,000 people, respectively; and level of bus usage in 2001, prior to the road-building program. The sample is split in half randomly 500 times, where we grow the tree on the first half of the data and then estimate the RDD effect within each leaf of that tree with the second half.

To illustrate what is happening in our estimation, it is useful to consider a single tree model. In the first stage, the estimator splits the data in half and estimates the structure of the tree. For comparison, a baseline RDD on the entire sample finds a treatment effect of 0.288 with a standard error of 0.116.[11] Output may look like the following:

```
x1 < 24.71; x0 in { 8 }; x2 < 108.99; x4 in { 0 },0.113 (0.093)  [537]
x1 > 24.71; x0 in { 8 }; x2 < 108.99; x4 in { 0 },0.467 (0.712)  [122]
x0 in { 22 }; x2 < 108.99; x4 in { 0 },0.591 (0.369) * [558]
x2 > 108.99; x0 in { 8 22 }; x4 in { 0 },0.070 (1.689)  [268]
x0 in { 21 23 }; x4 in { 0 },0.382 (0.208) ** [3014]
x0 in { 8 22 }; x4 in { 1 },0.197 (0.375)  [234]
x0 in { 21 23 }; x4 in { 1 },0.162 (0.374)  [353]
```

Each line is a rule for determining if the observation falls into that leaf. This is followed by the estimated treatment effect in that leaf, its standard error, star notation indicating level of significance at the 0.10 (*), 0.05 (**), or 0.01 (***) level, and the number of observations falling into that leaf in brackets. The estimator found two statistically significant effects. the first is for states 21 and 23 that previously did not have a road; the estimated treatment effect is a 38.2 percent point increase in the probability of receiving a bus route after building a road. The second effect is a more complex splitting of the data: for state 22, when the nearest city of 100,000 people is less than 108 kilometers away, and a bus route did not previously exist, the treatment effect is estimated to be 59.1 percent points with a standard error of 39.6 percent, indicating this effect is marginally significant at the 10 percent level. Importantly, presence of a prior bus route ($x4 = 1$) is associated with low and statistically insignificant treatment effects, which follows intuition given that it is not possible for the outcome variable to grow in this subsample.

[11] This is different than the grouped RDD effect on all the data because the data here is randomly split into two samples.

Table 9: Main Results: RDD, Discrete-Only Splits

Classification	Mean	SE
$x1 \in \{21\}$	0.538	0.182
$x1 \in \{23\}$	0.532	0.181

This first tree is only used for its structure; we use the second sample to estimate the treatment effects. This second tree is:

```
x2 < 108.99; x0 in { 8 22 }; x4 in { 0 },0.268 (0.110) *** [1217]
x2 > 108.99; x0 in { 8 22 }; x4 in { 0 },-0.085 (0.490)  [106]
x0 in { 21 23 }; x4 in { 0 },0.487 (0.075) *** [3515]
x4 in { 1 },0.217 (0.287)  [587]
```

The second tree has fewer nodes than the first tree. This is due to ex-post pruning which removes leaves that have null estimates. This can happen since the data is randomly sorted into the two samples. There may be insufficient variation in the second sample to estimate treatment effects in the tree estimated on the first sample. Interestingly, the second tree finds an even stronger effect for states 21 and 23 without prior bus routes. The two leaves with prior bus routes are now combined into one; there is still no significant effect. The first and third leafs are also combined, resulting in a more precise estimate for states 8 and 22 (versus just 22) with cities of more than 100,000 people less than 108 km away.

Two interesting outcomes of this process are that, first, the overall treatment effect of 0.288 is a composite estimate mixing together several different estimates. It misses the much stronger (and precise) effect in states 21 and 23. Second, the estimated model is more complex than anyone would ever stumble upon a priori. The first leaf's rule splits on a continuous variable and interacts with two other discrete variables.

We repeat this process 500 times to produce our overall estimator. This procedure produces estimated treatment effects within each estimated leaf, and those leaves may vary across trees.

When we restrict the forest to only cut on discrete variables, we obtain two statistically significant treatment effects. Table 9 reports; they are both similar, and the treatment only occurs in states 21 and 23.

Table 10 shows the results for the full model allowing for continuous variables to enter the

Table 10: Main Results: RDD, All Splits

x0	x1	x2	x3	x4	RDD	SE
21	11.16376	0	177.93822	0	0.482	0.293
23	11.16376	0	177.93822	0	0.482	0.286
23	22.32752	0	0	0	0.484	0.286
23	22.32752	0	29.65637	0	0.484	0.286
21	22.32752	0	0	0	0.485	0.286
21	22.32752	0	29.65637	0	0.485	0.286
23	22.32752	0	59.31274	0	0.485	0.286
21	0	0	177.93822	0	0.488	0.282
23	0	29.65637	148.28185	0	0.488	0.29
21	11.16376	0	148.28185	0	0.489	0.284
21	22.32752	0	59.31274	0	0.489	0.285
23	0	0	177.93822	0	0.489	0.276
23	11.16376	0	148.28185	0	0.489	0.278
21	11.16376	29.65637	88.96911	0	0.49	0.294
23	11.16376	29.65637	88.96911	0	0.491	0.288
21	11.16376	29.65637	0	0	0.493	0.277
21	11.16376	29.65637	29.65637	0	0.493	0.277
21	0	29.65637	118.62548	0	0.495	0.288
23	0	29.65637	118.62548	0	0.495	0.282
23	11.16376	29.65637	0	0	0.495	0.269
23	11.16376	29.65637	29.65637	0	0.495	0.269
21	0	0	148.28185	0	0.497	0.269
21	11.16376	0	118.62548	0	0.497	0.273
21	11.16376	29.65637	59.31274	0	0.497	0.274
23	0	0	148.28185	0	0.497	0.262
23	11.16376	0	118.62548	0	0.497	0.267
23	11.16376	29.65637	59.31274	0	0.497	0.267
21	0	29.65637	88.96911	0	0.499	0.278
21	0	29.65637	0	0	0.5	0.267
21	0	29.65637	29.65637	0	0.5	0.267
23	0	29.65637	88.96911	0	0.5	0.272
21	11.16376	0	88.96911	0	0.501	0.262
23	0	29.65637	0	0	0.502	0.259
23	0	29.65637	29.65637	0	0.502	0.259
23	11.16376	0	88.96911	0	0.502	0.255
21	0	29.65637	59.31274	0	0.503	0.265
23	0	29.65637	59.31274	0	0.504	0.258
21	11.16376	0	0	0	0.505	0.241
21	11.16376	0	29.65637	0	0.505	0.241
21	0	0	118.62548	0	0.506	0.256
23	0	0	118.62548	0	0.506	0.249
23	11.16376	0	0	0	0.507	0.232
23	11.16376	0	29.65637	0	0.507	0.232
21	11.16376	0	59.31274	0	0.508	0.239
23	11.16376	0	59.31274	0	0.509	0.231
21	0	0	88.96911	0	0.511	0.244
23	0	0	88.96911	0	0.511	0.236
21	0	0	0	0	0.512	0.229
21	0	0	29.65637	0	0.512	0.229
23	0	0	0	0	0.514	0.22
23	0	0	29.65637	0	0.514	0.22
21	0	0	59.31274	0	0.515	0.228
23	0	0	59.31274	0	0.515	0.221

tree. There are 43 distinct statistically significant treatment effects, ranging in magnitude from 0.482 to 0.515. This is startlingly uniform; for comparison the statistically insignificant effects (not reported) vary over a much greater range. To produce this table, we sampled over 10 intervals for the continuous variables and over all combinations of the discrete variables. For each x_i that had a significant RDD estimate, we added it to the table. A simple summary of the results is that there is a treatment effect of about 0.5 in villages located in states 21 and 23 (echoing the discrete-cut-only case above), where the closest village of 10,000 or 100,000 people is less than 23 or 30 kilometers away, respectively, and where there was no pre-existing bus route in 2001. Elsewhere, there is no statistically significant treatment effect.

The importance of these findings are two-fold: first, the average treatment effect in these subsamples is right around 0.5, which is three times larger than the treatment effect found on the grouped data. Second, the treatment effect only occurs in a subset of the villages. This is a highly complex nonlinear partitioning of the original data, highlighting the fact that it is highly unlikely anyone would ever be able to guess that this was the true model. The alternative of saturating the specification with every combination of all discrete variables runs into the problem of how to cut the continuous variables. There are literally an infinite number of cuts, and as such this is not a fruitful approach. Fully nonparametric approaches are fully general, but converge so slowly that it is unlikely to be a productive path for practitioners with finite data sets. Our estimator provides a middle path that allows for arbitrary structure while retaining the efficiency properties of pre-specified models.

8 Conclusion

We have presented a two-stage estimator for the problem of assigning statistical models to disjoint subsets of a sample. Leveraging recent results on the estimation of honest trees, we split the sample into two random halves. The first half is used to estimate the classification tree assigning observations to models. The second half is used to estimating parameters of those models within each assignment. Splitting the data in this fashion allows us to derive econometric results that the tree is consistently estimated, converges to the truth at a faster-than-parametric rate, and therefore can be ignored when constructing standard errors for the estimates in the second stage. Our method applies to all empirical settings where the researcher has reason to believe that the estimated model may vary across units of the sample in some observable fashion.

In future work, we hope to extend our results to the case of resampling and to extend the

estimation procedure to use random forests, which should improve efficiency. We also plan to bring in a micro-level data set at the household level to match with the village-building program. This will let us to test for observable heterogeneity at a much finer level than our current data allows for.

References

Ai, C., and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

Asher, S., and P. Novosad (2016): "Market Access and Structural Transformation: Evidence from Rural Roads in India," .

Assmann, S. F., S. J. Pocock, L. E. Enos, and L. E. Kasten (2000): "Subgroup analysis and other (mis) uses of baseline data in clinical trials," *The Lancet*, 355(9209), 1064–1069.

Athey, S., and G. Imbens (2015): "Machine learning methods for estimating heterogeneous causal effects," *arXiv preprint arXiv:1504.01132*.

Banerjee, A., S. Barnhardt, and E. Duflo (2016): "Can Iron-Fortified Salt Control Anemia? Evidence from Two Experiments in Rural Bihar," Discussion paper, National Bureau of Economic Research.

Barreca, A., K. Clay, O. Deschênes, M. Greenstone, and J. S. Shapiro (2015): "Convergence in Adaptation to Climate Change: Evidence from High Temperatures and Mortality, 1900–2004," *The American Economic Review*, 105(5), 247–251.

Bryan, G., S. Chowdhury, and A. M. Mobarak (2014): "Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh," *Econometrica*, 82(5), 1671–1748.

Cappelli, C., F. Mola, and R. Siciliano (2002): "A statistical approach to growing a reliable honest tree," *Computational statistics & data analysis*, 38(3), 285–299.

Card, D. (1999): "The causal effect of education on earnings," *Handbook of labor economics*, 3, 1801–1863.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71(5), 1591–1608.

Chen, X., and X. Shen (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, pp. 289–314.

Chetty, R., N. Hendren, and L. F. Katz (2015): "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment," Discussion paper, National Bureau of Economic Research.

Collard-Wexler, A., and J. De Loecker (2015): "Reallocation and Technology: Evidence from the US Steel Industry," *THE AMERICAN ECONOMIC REVIEW*, 105(1), 131–171.

Dell, M. (2015): "Trafficking networks and the Mexican drug war," *The American Economic Review*, 105(6), 1738–1779.

Doyle, J., J. Graves, J. Gruber, and S. Kleiner (2015): "Measuring returns to hospital care: Evidence from ambulance referral patterns," *The journal of political economy*, 123(1), 170.

Gelman, A., and G. Imbens (2014): "Why high-order polynomials should not be used in regression discontinuity designs," Discussion paper, National Bureau of Economic Research.

Heckman, J., R. Pinto, and P. Savelyev (2013): "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *THE AMERICAN ECONOMIC REVIEW*, 103(6), 1–35.

Imbens, G. W., and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of econometrics*, 142(2), 615–635.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

Lee, D. S., and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79(1), 147–168.

Pollard, D. (2012): *Convergence of stochastic processes*. Springer Science & Business Media.

Raballand, G., R. L. Thornton, D. Yang, J. Goldberg, N. C. Keleher, and A. Müller (2011): "Are rural road investments alone sufficient to generate transport flows? Lessons from a randomized experiment in rural Malawi and policy implications," *Lessons from a Randomized Experiment in Rural Malawi and Policy Implications (January 1, 2011). World Bank Policy Research Working Paper*, (5535).

Shen, X., and W. H. Wong (1994): "Convergence rate of sieve estimates," *The Annals of Statistics*, pp. 580–615.

Van Der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence*. Springer.

Wager, S., and S. Athey (2015): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *arXiv preprint arXiv:1510.04342*.

Wager, S., and G. Walther (2015): "Uniform Convergence of Random Forests via Adaptive Concentration," *arXiv preprint arXiv:1503.06388*.

Zhang, J., and I. Gijbels (2003): "Sieve empirical likelihood and extensions of the generalized least squares," *Scandinavian Journal of Statistics*, 30(1), 1–24.