Moment Forests

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

October 28, 2019

Abstract

We propose a new methodology, *moment forests*, that allows parameters in a broad class of moment-based models to vary across groups of observations on the basis of observable characteristics. Leveraging a generalization of classification trees, the estimator first assigns observations to disjoint subgroups sharing common parameters, and then estimates an empirical model within each group. We prove the uniform consistency of this estimator and show that standard rates of convergence apply to the second stage estimates under weak regularity conditions. We showcase our approach by estimating heterogeneous treatment effects in a regression discontinuity design in a development setting.

Keywords: Classification Trees; Heterogeneous Parameters; Model Selection; General Method of Moments

*Nekipelov: University of Virginia (denis.nekipelov@gmail.com); Novosad: Dartmouth College, (paul.novosad@dartmouth.edu); Ryan: Olin Business School, Washington University in St. Louis; NBER; and CESifo (stephen.p.ryan@wustl.edu). We thank numerous participants at seminars for helpful comments; all errors remain our own. Formerly titled "Classification Trees for Heterogeneous Moment-Based Models." First draft: Spring 2016.

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 units in the population. While theory is often helpful in addressing the first issue, and nonparametric methods can, in principle, address the latter two concerns in complete generality, data limitations often force the researcher to make decisions about the empirical specification. In practice, the 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 "robustness checks."

Two major issues arise from this process. First, the model resulting from the search may have different statistical properties from the original model, as the result of choosing the specification on the basis of the answers it produces. Second, the researcher often only considers a subset of the possible modeling choices, potentially introducing specification bias in the estimates. This paper proposes a method that addresses both of these issues, recovering the correct specification in a systematic fashion without introducing bias in the estimates due to the search process.

Stated informally, we are interested in estimating the parameters of the following conditional moment:

$$E[Y - G(X; \theta(Z))|X = x] = 0, \tag{1.1}$$

where G is a known data-generating process mapping the matrix of covariates X to the vector of outcomes Y.[1] Our key innovation over the standard method of moments estimator is that we allow the parameters to be governed by the unknown function $\theta(Z)$, where Z is a subvector of X. That function is the object of interest in this paper. This is a classic problem in the semiparametric conditional moment inference in Chamberlain (1986), Robinson (1988), Powell, Stock, and Stoker (1989), Bickel, Klaassen, Bickel, Ritov, Klaassen, Wellner, and Ritov (1993), Newey (1994), Powell (1994), and Ai and Chen (2003) among others. Our goal in this paper is to introduce a class of methods that replace the often computationally-burdensome and data-intensive process of estimation of the infinite-dimensional parameter $\theta(\cdot)$ with a greedy search that has superior small-sample properties.

We introduce *moment forests* to estimate $\theta(Z)$. Each moment forest is an ensemble of

[1] We present formal definitions in Section 3.

moment trees, which are generalizations of classification trees that group observations in a sample together on the basis of moment functions. In each moment tree, we recursively partition the data into binary splits such that the fit of the underlying moment functions are maximized on each subset, subject to appropriate stopping criteria. The outcome of this process is a different parameter vector in each disjoint subset of the sample. To leverage the well-known result that ensemble estimators have lower error than their individual components, the moment forest then reports the average parameter vector across all its constituent trees as its estimate. As opposed to a standard mixture model—e.g., a random coefficients logit, where individuals are assigned a type from some distribution but are all assumed to follow the same model—our method assigns a model with certainty to a group of observations. Our method builds on the split sample method of *honest trees* (Cappelli, Mola, and Siciliano (2002), Wager and Athey (2015), Athey and Imbens (2015)). Using one subset of the data, we first estimate the structure of $\theta(Z)$, which governs how parameters should be assigned to observations. In a second step, we then take the structure as given and estimate parameters on the remaining data.

We make two general contributions to econometric theory in this context. First, we show the uniform convergence of conditional moment-based semiparametric models that use (ensembles of) classification trees, like our moment forests, to control for unobserved heterogeneity. This allows us to prove the consistency of our approach for recovering the structure of $\theta(Z)$. Second, we show that if the complexity of the unobserved heterogeneity is relatively low (i.e. the number of components of the unobserved heterogeneity grows sublinearly with the sample size), then the first step of the estimation does not affect the uniform convergence of the semiparametric conditional moment function over the values of the finite-dimensional parameter. This step builds on the observation that classification and regression trees are local estimators that aggregate information from the data in shrinking neighborhoods of the parameter space. This results allows the researcher to use the usual standard errors in the second step of the procedure, greatly simplifying the application of our method. We also show that a simple version of the bootstrap yields valid confidence sets for estimated parameters.

Our setting is one with a long academic literature. Medical researchers and applied microeconomists have long struggled with the issue of subgroup analysis.[2] The basic issue is that researchers, through statistical ignorance (or more nefarious motivations), may search

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

across subgroups until they find one with a treatment effect that is statistically significantly different from the baseline. Emphasizing this finding and ignoring groups for which the effect is zero leads to substantial reporting bias and can provide misleading policy implications. However, understanding the true structure of treatment heterogeneity can be essential in understanding program impacts. The problem of subgroup analysis has become so severe that it is becoming common to pre-announce testing hypotheses in public before engaging in an experiment via a "pre-analysis plan."[3]

Further, it is nearly universal in the economics literature to find several specifications reported in the same paper. Researchers commonly estimate a "baseline" or "preferred" statistical model, either on the basis of theory, intuition, or fit. The next step, especially in reduced-form settings where the 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. One approach is to include various levels of fixed effects that operate at some aggregated level. A second approach is to allow for some form of interactions between demographic characteristics and outcomes.[4] While econometric theory exists for various specification tests for growing or pruning models, this step is rarely guided by formal econometric intuition. Instead, researchers most often consider a (small) finite number of specifications to run and report those results as "robustness checks." Robustness checks are pervasive across all fields in economics; see e.g., Chetty, Hendren, and Katz (2016) in education, Banerjee, Barnhardt, and Duflo (2018) 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.[5]

While the desire to know whether estimates are sensitive to the particular modeling choices made in forming those estimates is clearly laudable, there are two important limitations to this approach. The first, as mentioned above, is that 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

[3]The Hypothesis Registry at J-PAL (`https://www.povertyactionlab.org/Hypothesis-Registry`) is an early example of a pre-analysis plan registry, now subsumed by the the AEA RCT Registry (`https://www.socialscienceregistry.org/`).

[4]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.

[5]At a top economics department in 2016, every single job market paper in an applied field contained some variety of robustness checks. A majority had a section expressly labeled "Robustness Checks."

in proper inference. The second issue is that robustness checks are almost never exhaustive, nor are they guided by some econometrically-sound search process that guarantees convergence to the true underlying data-generating process. One may erroneously conclude that the estimates are robust simply because non-robust specifications happen to have not been chosen. In models with discrete variables, it is almost unheard of to estimate the model on all subsets of the data, in part because there are typically too many subsets to consider. When continuous variables are introduced, the problem becomes infinite-dimensional because any sub-intervals of the continuous variables can be considered; researchers in this context usually choose arbitrary partitions based on quantiles or round numbers. By partitioning observations on the basis of an objective criterion and then generating standard errors that account for the partitioning process, our estimator jointly solves both problems.

The remainder of the paper is organized as follows: we start with a high-level overview of the related literature and the estimator (Section 2) before developing its statistical properties (Section 3). We then show its small-sample performance in a Monte Carlo (Section 4), and apply it to a regression discontinuity design in a development setting (Section 5). Section 6 concludes.

2 A High-Level Overview and Examples

Our goal is to assign parameters to groups of observations on the basis of observables. Our approach is to divide the data into subsets with constant parameters, while simultaneously avoiding overfitting and accounting for the partitioning process when generating standard errors.

Decision trees provide one approach to partitioning the data. We begin with an initial partition (known as the "root" or "stump") containing all of the data. The decision tree is then generated by recursively splitting the data into two sub-partitions along a single splitting variable at a time. For continuous splitting variables, the algorithm searches for an inequality threshold value that maximizes the criterion function; this results in a split into two partitions by $1(z \leq \bar{z})$ for splitting variable z and threshold $\bar{z}$. For discrete variables, such as fixed effects, the algorithm searches across (all) possible disjoint binary subsets. After finding the variable and split that maximizes the criterion function, the algorithm generates two disjoint sets of the data, each known as a "leaf." The algorithm is then repeated on each leaf, cutting the data into smaller and smaller partitions until a stopping criterion is met.[6]

[6]The literature has considered several different types of stopping criteria. We will require the number of

The resulting tree is a collection of rules which sorts observations together on the basis of a given criterion function.

The classification of observations into groups using tree-like reasoning has a long intellectual history. The use of trees in statistical settings can trace its more recent roots to Morgan and Sonquist (1963) (regression trees) and Messenger and Mandell (1972) (classification trees). Regression trees fit the average value of the subsample's dependent variable; the criterion function is the mean squared error within the leaf. Classification trees vote for assignment for an observation into a group on the basis of observable variables; the criterion function is typically "node impurity," a measure of the dissimilarity of observations in a given node. Academic and practical interest in the area grew rapidly after the publication of the seminal book by Leo, Friedman, Olshen, and Stone (1984). Breiman advanced the use randomization and ensembles of tree predictors with two very incredibly influential papers on bagging (Breiman, 1996) and random forests (Breiman, 2001). A more comprehensive overview of the history of trees and their many recent variants is given by Loh (2014).

We use 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 observables Z. Our criterion function is a standard generalized method of moments (GMM) objective function, which measures the fit between dependent variables and a parametric function of the observable variables, parameters, and unobservable shocks. At each stage of the partitioning process, we split the data in a way that optimizes the value of the moment function in each leaf. The partitioning continues until a set of stopping criteria are met, at which point the data are partitioned into K disjoint sets, $Z = \{Z_1, \ldots, Z_K\}$. We then assign a unique parameter vector, θ_k, that solves the moment function in each partition Z_k.[7] The process is summarized below in Algorithm 1.

Our approach results in a generalization of the standard GMM estimator. At the root node, our model is exactly the same as a standard GMM-based model, which encompasses an extraordinarily large class of empirical problems. In the conventional GMM approach, one solves for θ using the entire sample by optimizing the value of the GMM criterion function. Our approach extends and generalizes that approach by partitioning the data and allowing the parameters to vary across subsets of the data.

observations in each leaf to be above some minimum integer $\bar{k}$, the proportion of data in each leaf to be at least some $\bar{\alpha}$, and the improvement in the criterion function after the split to be greater than some threshold. The convergence criteria are set using using cross-validation on a holdout sample.

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

Let $\mathcal{S} \in \{\mathcal{S}_0, \ldots, \mathcal{S}_k\}$ represent a set of disjoint partitions of the data. Informally, denote the solution to a set of moment functions on a partition $\mathcal{S}_j$ of the data by $\theta(\mathcal{S}_j)$. Denote the set of partitions labeled terminal by $\mathcal{T}$.

Data: A set of observations on outcomes $Y \in \mathbb{R}^m$ and covariates $X \in \mathbb{R}^q$, of which Z is a subvector; let $W = (Y, X)$.

Initialize: Initialize $\mathcal{S}$ with a single partition $\mathcal{S}_0$ containing all of the data.

while *all* $\mathcal{S}_j \notin \mathcal{T}$ **do**

  foreach $\mathcal{S}_j \notin \mathcal{T}$ **do**

    1. Solve for $\theta(\mathcal{S}_j)$.
    2. Randomly choose $p \leq q$ components $Z_p \subseteq Z_L \in \mathbb{R}^q$ to use as splitting variables.
    3. For each $Z \in Z_p$, solve for the optimal splitting point (continuous Z) or binary partition (discrete Z) according to the splitting criterion (see the text for a discussion).
    4. Set the optimal splitting variable, Z^*, by choosing $Z \in Z_p$ that maximizes the splitting criteria *and* that satisfies regularity conditions (see the text for conditions).

    if *there exists an optimal split* **then**
      Replace S_j with two partitions split along Z^*;
    else
      Add S_j to $\mathcal{T}$.
    end
  end
end

Algorithm 1: Growing a Moment Tree

Any such search process requires a method to avoid overfitting of the data and to generate standard errors that take the search process into account. We use a two-step approach to generate "honest trees" with desirable theoretical properties. We split the data into two samples. In a first step, using one of the samples, we run the algorithm described above to generate a moment tree. In a second step, we then treat the tree's structure as a known function and re-estimate the values of the θ_k in each leaf using the second sample. Below, we formally show that this approach guarantees two important theoretical results: first, that the tree will uniformly converge to the true data-generating process; and second, that, under certain regularity conditions, the rate of convergence of the first step is faster than parametric. This second result allows the the researcher to ignore the influence of statistical

for *b=1,...,B* **do**

1. Resample the data with replacement.
2. Split the data into two disjoint partitions, the *tree structure sample* and the *estimation sample*.
3. Estimate the structure of the moment tree on the tree structure sample (see Algorithm 1).
4. Using the estimated tree structure, replace estimates of $\theta_b(Z)$ using the estimation sample.

end

return $\theta(Z) = \frac{1}{B}\sum_{b=1}^{B} w_b(Z)\theta_b(Z)$, where $w_b(Z)$ is a tree- and Z-specific weight, such as the inverse estimated standard error.

Algorithm 2: Growing a Moment Forest

error from the first step when calculating standard errors in the second step.

We extend this approach to random ensembles, as in Breiman (1996, 2001), to produce a *moment forest*. As with a random forest, we repeatedly resample the data with replacement and run the two-stage estimation outlined above on each resampled data set. However, a substantial difference is that only a random subset of p variables are considered for splitting at each node. The final 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 Z datasets, as only a subset of variables is searched over at each split. The process is summarized in Algorithm 2.

For a given forest, we compute θ using an inverse-variance weighted average to increase the accuracy of the estimated parameter. Intuitively, estimates from trees with high standard errors are downweighted relative to more precise estimates from other trees. We calculate standard errors using the bootstrap applied to the construction of the entire random forest. The final output of the random forest is a parameter estimate and standard error for every observation in the original dataset; this parameter estimate is effectively a weighted mean of the parameter estimate in each leaf of a tree where that particular observation appeared.

Our work is most closely related to Athey, Tibshirani, Wager, et al. (2019) (ATW), which

leverages a generalization of random forests to produce weights for nearby observations in a nonparametric regression.[8] While both papers use similar tools, they differ in framing, the details of the algorithm, and the statistical results they derive.

ATW frame their approach as a form of adaptive nearest-neighbor estimation extended to local (linear) likelihood. Their framework requires the researcher to specify a scoring function that satisfies a number of assumptions which may be non-trivial; ATW provide score functions for IV regression and quantile regression. In contrast, our approach applies to all generalized method of moment estimators that satisfy a weak regularity condition controlling the rate of convergence of the underlying primitives, which covers virtually all linear and nonlinear contexts encountered in practice.[9] Finally, we emphasize a broad set of economic applications that have not previously been considered in a random forest context, including regression discontinuity, multiple treatments and multiple outcomes, and structural estimation.

A key distinction between these papers is that we are able to prove uniform consistency and the uniform convergence rate of our estimator. The ATW algorithm works by producing a weighting function for nearby observations which are then plugged into a local estimator. They then construct a pointwise consistent estimator in the spirit of a classic local polynomial regression. The resulting parameters are not formed by averaging estimates from different trees, so they cannot make use of the typical approaches to proving uniform consistency. In contrast, our semiparametric moment-based setting leads to a different algorithm and conceptually different econometric results. Our algorithm pools information based on the parameter values as opposed to the location in the state space. In turn, this allows us to prove the stronger statistical results of uniform convergence of our estimator to the underlying data-generating process, and, in turn, uniform consistency for the set of estimated parameters. Moreover, we derive the uniform convergence rate for our estimator making it suitable as an input into the broadly used "plug-in" estimators that typically require a lower bound on the uniform convergence rate of $o(n^{-1/4})$.

2.1 Examples

Our estimator has a broad set of applications. We illustrate several common settings where allowing the parameters to vary across an observable characteristic of the sample may be

[8] Both papers were written contemporaneously and independently of each other.

[9] A Java package on Github (`https://github.com/cactus911/momentForests`) provides an implementation that requires the researcher only to specify the moment function.

useful.

2.1.1 Linear Regression

The most commonly-applied statistical model in econometrics is the linear regression:

$$Y = X\theta + \epsilon, \tag{2.2}$$

where Y is a $N \times 1$ vector, X is a $N \times K$ matrix of covariates, and ϵ is a $N \times 1$ vector of unobservable errors. The parameter of interest in this model is (a subset of) θ. A common modeling approach is to saturate Equation 2.2 with many fixed effects at various levels of aggregation. The idea is to "control" for other confounding effects that may influence the outcome variables at those grouped levels. There are several issues with this approach: first, some of the fixed effects may be equal to each other. Grouping them together will improve statistical precision. Second, fixed effects only control for additive effects at exactly the group level, while critically still imposing a constant relationship between X and Y. Our approach generalizes Equation 2.2 to the following model:

$$Y = X\theta(Z) + \epsilon, \tag{2.3}$$

allowing θ to be a function of Z, a $N \times L$ matrix of variables that may or may not overlap with X. The parameter of interest, θ, may now depend on observable characteristics of each observation, such as the categorical variables used to generate fixed effects. This model is substantially more flexible than the baseline fixed effects model. One can replicate the model in Equation 2.2 while allowing the relationship between Z and θ to be arbitrarily nonlinear, to include arbitrary interactions between variables, or interactions between continuous variables and fixed effects.

The same setup can be directly applied to other models based on linear regression, such as the regression discontinuity design, which we describe briefly below and explore in an empirical example in Section 5.

Heterogeneous Treatment Effects, Multiple Treatment Arms, and Multiple Outcomes A special case of the linear regression model is the randomized controlled trial (RCT). Athey and Imbens (2015) introduce the following heterogeneous treatment effects

specification:

$$Y = X\theta + W\tau(Z) + \epsilon, \tag{2.4}$$

where W is an indicator for treatment status and $\tau(Z)$ is the treatment effect, which may depend on Z.

Equation 2.4 can be extended to both multiple treatment arms and multiple outcomes. For M treatments, the model is:

$$Y = X\theta + 1(Y \in M_1)W_1\tau_1(Z) + \cdots + 1(Y \in M_M)W_M\tau_M(Z) + \epsilon, \tag{2.5}$$

where $1(Y \in M_m)$ is an indicator for whether an observation belongs to treatment arm m, W_m is a treatment status indicator within each arm, and $\tau_m(Z)$ is the treatment effect in arm m. While this model allows for multiple treatment effects by treatment arm, it also allows the treatment effect to be grouped across different arms, e.g. if the treatment effect is zero across several arms. This can result in a significant improvement in precision. The Z may contain other variables, such as demographics, that further increase the flexibility of the model to allow for heterogeneous treatment effects within and across groups in different treatment arms. For example, it may be that the treatment effect is zero in some arms for all observations, takes different values in other arms for all observations, and has different values for different groups in the remaining arms.

In the case of multiple outcomes, suppose that the econometrician has J outcome variables Y_j, which is common in RCT settings. The model for assessing multiple treatment outcomes can be written as:

$$Y = X\theta + 1(Y \in J_1)W_1\tau_1(Z) + \cdots + 1(Y \in J_J)W_J\tau_M(Z) + \epsilon, \tag{2.6}$$

where $1(Y \in J_j)$ is an indicator function that identifies which $j \in J$ outcomes Y belongs to. Researchers typically estimate all the outcome variables separately. The joint approach made possible by this paper has the benefit of grouping together similar treatment effects, improving precision. It also allows researchers to study the correlation of treatment effects across different outcome variables, which is otherwise difficult. As above, it also allows treatment effects to vary across other observables Z, while generating correct standard errors.

Regression Discontinuity Design (RDD) In settings where treatments are assigned on the basis of an exogenously-given and non-manipulable cutoff, c, as measured by some variable X, a regression discontinuity design may be used to estimate the causal effects with the following equation:

$$Y = \alpha + D\tau + X\theta + \epsilon, \tag{2.7}$$

where $D = 1$ if $X \geq c$ and $D = 0$ otherwise. Our model generalizes Equation 2.7 to allow for multiple treatment effects by estimating the structure of $\tau(Z)$, where Z are additional observables. Our application below uses the "fuzzy" variant of the RDD method where the treatment probability changes discontinuously around the cutoff.

2.1.2 Two-Step Dynamic Estimators

The approach of Bajari, Benkard, and Levin (2007) estimates parameters of dynamic models using a two-step procedure. In the first step, the econometrician estimates a set of reduced form policy functions linking agent behavior with a set of observable state variables. In the second step, these reduced form policy functions are projected onto an underlying structural model, recovering estimates of the dynamic parameters. A key assumption of the estimator is that the first-step policy functions must be grouped in such a way that within a group they are all estimated on data generated by the same equilibrium, otherwise they may be biased (Otsu, Pesendorfer, and Takahashi, 2016). Our approach makes it possible to ensure consistent estimates of policy functions that vary across markets, for example, if firms are playing different equilibria in different markets. In this case, the relevant model is:

$$Y = f(X, Z, \epsilon), \tag{2.8}$$

where f is the reduced-form policy function, such as a probit used to estimate entry probabilities, X is a vector of state variables, and Z a vector of market-level indicator functions. Our approach assigns a policy function to each market; in the limiting case, it will find that there are no significant differences across markets and group all of the policy functions together, as is often done in practice without formal justification (see, for example, Ryan (2012)).

2.1.3 *Logit Models*

In Greenstone, Ryan, Yankovich, and Greenberg (2017), the authors use re-enlistment data to estimate the value of statistical life (VSL) for soldiers in the US Army. They model the probability that soldier i re-enlists after their first term as:

$$Pr(i \text{ re-enlists}) = \frac{\exp(\alpha b_i + \gamma h_i + X_i'\theta)}{1 + \exp(\alpha b_i + \gamma h_i + X'\theta)}, \tag{2.9}$$

where b_i is the bonus offered for re-enlistment, h_i is a measure of the mortality hazard faced by the soldier at the time of re-enlistment, and X is a vector of demographics. The two key parameters are the marginal utilities of the bonus offer, α, and the marginal disutility of the excess hazard, γ; the ratio of the two is the value of statistical life. A key question is how the VSL varies by observable group. Using the approach in this paper, the model with heterogeneous VSL is:

$$Pr(i \text{ re-enlists}) = \frac{\exp(\alpha(Z) b_i + \gamma(Z) h_i)}{1 + \exp(\alpha(Z) b_i + \gamma(Z) h_i)}. \tag{2.10}$$

The model now depends only on the bonus and excess hazard covariates, with all other observable demographics entering Z. The key distinction between Equations 2.9 and 2.10 is that the object of the interest, the value of statistical life, can now vary in an unrestricted way across all observable characteristics of the soldiers. This is a very complicated object, as there are literally billions of possible combinations of discrete variables, and an infinitely-large number of continuous cuts of the data, that could interact with b_i and h_i in the base specification. Moment trees guarantee that the true underlying model will be found asymptotically, and will uncover the most important features of that heterogeneity in finite samples.

3 Econometric Theory

In this section we formalize our analysis by placing it into the classic framework of estimation of *semiparametric conditional moment models*. In these models, the conditional moment function is determined by the target finite dimensional parameter of interest as well as, possibly, an infinite-dimensional nuisance parameter. We focus on the case where the infinite dimensional parameter corresponds to the conditional distribution of the data given covariates. We allow this distribution to be heterogeneous subject to an *a priori* unknown

exclusion restriction. We do so by assuming that there exists a linearly independent subvector of the vector of covariates which can fully explain the structural breaks in the conditional distribution of the data. The goal of the moment tree algorithm is to provide a flexible way of estimating the target finite-dimensional parameter while accounting for the heterogeneity of the data distribution.

Our theoretical analysis consists of five steps. We first characterize the general structure of the data-generating process as consisting of conditional moment functions with parameters that vary as a function of observables. We specify this distribution of heterogeneity to be finite-dimensional, but we also allow the dimension of the target parameter to grow as the sample size increases. Second, we propose an algorithm to jointly recover the parameters of those conditional moment functions and their distribution as a function of observables in a two-step process using a generalization of classification trees. Third, we describe the basic statistical properties of our estimator. Our method leads to a repeated computation of the minimizer of the empirical GMM objective function. We characterize the estimator for the target parameter as a classic Z-estimator. To show its consistency, we demonstrate uniform convergence of the empirical moment function that arises as an outcome of our algorithm to its population counterpart. This result requires us to have a bound on the complexity of a special class of empirical processes that randomly downsample a given proportion of the data. This also allows us to provide a lower bound guarantee for the uniform convergence rate for both the target and the nuisance parameters. Fourth, under additional assumptions on the data generating process, we establish asymptotic normality of the target finite-dimensional parameter. Fifth, we conclude with an extension of our primary results from moment trees to moment forests.

3.1 Setup: conditional moment model with heterogeneous data distribution

We begin our analysis by characterizing a heterogeneous data generating process and defining the target parameter as a zero of a conditional moment.

We consider the model defined by the moment function $\rho(\cdot;\cdot) : \mathcal{W} \times \Theta \mapsto \mathcal{M}$, where $\mathcal{W}$ is a subset of $\mathbb{R}^n$, Θ is a convex compact subset of $\mathbb{R}^p$ and $\mathcal{M}$ is a subset of $\mathbb{R}^p$ (where we set the dimension of the moment vector $\rho(\cdot;\cdot)$ to be equal to the dimension of the parameter vector θ without loss of generality). We assume that the data generating process is characterized by the distribution of a random vector $W = (Y, X)$ where the random variable X takes the values in $\mathcal{X} \subset \mathbb{R}^q$.

We assume that this distribution has an absolutely continuous density $f_X(\cdot)$. Our results will also apply to the cases where some of the components of X are discrete. Our model has a special structure in the sense that we assume that there is a subvector of X that we denote Z determining data heterogeneity, *i.e.* conditional on Z, there is fixed single distribution of the remaining components of X. We assume that Z does not belong to any proper linear subspace spanned by the remaining component of X. In other words, Z forms a valid exclusion restriction that characterizes the heterogeneity of the data generating process. We do not assume, however, that components of X that form Z are a priori known.

The support of Z, $\mathcal{Z}$ is an open convex subset of $\mathbb{R}^r$ $(r < q)$. The special structure of the model that we consider relies on the existence of the system of subsets of $\mathcal{Z}$ such that in each subset the data generating process corresponds to single homogeneous conditional moment model. Then the object of interest is the conditional moment

$$E[\rho(W;\theta) \,|\, X = x] = 0, \tag{3.11}$$

and the value of the parameter θ when z belongs to specific subsets of $\mathcal{Z}$.

Formally, we characterize the structure of heterogeneity in the following assumption.

ASSUMPTION 1. *For $K = 1, 2, \ldots$ there exists a system of K subsets $\mathcal{Z}^{kK}$, $k = 1, \ldots, K$ such that $\mathcal{Z}^{iK} \cap \mathcal{Z}^{jK} = \emptyset$ and*

1. *For $P_Z(\cdot)$, the distribution of random vector Z, and each k, $P_Z\left(\mathcal{Z}^{kK}\right) = O(1/K)$*
2. *For numeric sequences a_K and b_K decreasing in K, for each $\mathcal{Z}^{kK}$ there exists θ^{kK} such that*

$$\sup_{x \in \mathcal{X},\, z \in \mathcal{Z}^{kK}} \left\| E[\rho(W;\theta^{kK}) \,|\, X = x] \right\| = O(a_K), \tag{3.12}$$

   with $\min\limits_{i \neq j = 1, \ldots, K} |\theta^i - \theta^j| \geq b_K$, and

$$\inf_{\theta \in \Theta} \sup_{x \in \mathcal{X},\, z \in \mathcal{Z} \setminus \cup_{k=1}^{K} \mathcal{Z}^{kK}} \left\| E[\rho(W;\theta^{kK}) \,|\, X = x] \right\| \gg a_K$$

3. *For each $\mathcal{Z}^{kK}$ and θ^{kK} in (3.12) there exists a matrix function $A(\cdot) : \mathcal{X} \mapsto \mathbb{R}^{p \times p}$ such that eigenvalues of $A(\cdot)$ are uniformly bounded by some constant λ and*

$$E\left[A(X)\frac{\partial \rho(W;\theta^{kK})}{\partial \theta}\right]$$

is strictly positive definite.

4. *For each fixed k and K there exists a sequence of sets $\mathcal{Z}^{kN}$ for $N = K+1, K+2, \ldots$ such that for each $N > K+1$ $\mathcal{Z}^{k(N+1)} \subseteq \mathcal{Z}^{kN} \subseteq \mathcal{Z}^{kK}$ and the corresponding sequence of parameters θ^{kN} converges to a limit θ^{k*}.*

The general premise of Assumption 1 is the requirement of the existence of constant-parameter approximation of the data generating process by the conditional moment models that are valid within the finite volume subspaces of the support of the exclusion variable Z. The quality of approximation is controlled by the approximation bias a_K that vanishes as the number of elements in the partition of Z increases. Assumption 1 requires the existence of the system of subsets of $\mathcal{Z}$ with non-vanishing volumes where a conditional moment model is valid uniformly in each subset (up to an error of a_K) and it is not valid outside of this system of subsets. Moreover, the parameter values that solve the conditional moment within each subset are separated by b_K. Examples of settings where Assumption 1 holds include the case where $\mathcal{Z}$ has a finite partition where, in each element of the partition, the data can be characterized by a single conditional moment model. This could be the case where $\mathcal{Z}$ corresponds to geography and different models apply to different separated geographical locations (such as counties or states). Another simple example is the case where there is a unique subset of $\mathcal{Z}$ where the conditional model is valid.

We formulate the econometric problem as a problem of estimation of a set of parameters θ^{kK} by recovering subsets of sets in the set system $\{\mathcal{Z}^{kK}\}_{k=1}^{K}$ for a given K. There is no guarantee that we can recover sets $\mathcal{Z}^{kK}$. However, given that the conditional moment is valid inside each $\mathcal{Z}^{kK}$ it will be sufficient to recover some strict subsets of $\mathcal{Z}^{kK}$. We now develop a tree-based algorithm that estimates parameters θ^{kK} by searching for rectangular strict subsets of sets $\mathcal{Z}^{kK}$.

3.2 Estimation of a conditional moment model with data heterogeneity with classification trees

Our next step will be to describe the algorithm for estimation of the target finite-dimensional parameter subject to the heterogeneity of the data generating process.

For our analysis we use the notion of classification trees. In theory, we can try to find the sets $\mathcal{Z}^{kK}$ by fully triangulating the set $\mathcal{Z}$ (e.g. by defining the grid in this set). However, such a procedure will face a severe curse of dimensionality in high-dimensional spaces. The

classification tree replaces the brute-force grid search with a dimension-wise search and splits $\mathcal{Z}$ into non-overlapping rectangles. Each rectangle is then assigned the label k and parameter θ^{kK} if such assignment is possible (i.e. the corresponding rectangle is a strict subset of $\mathcal{Z}^{kK}$). We reserve the labels 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 $\mathbb{R}^r$.

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 $S^{(k+1,k+2)} = S^{k,i} \cap \{z \in S^{(k,i)} \,|\, z^d \neq c\}$. We then 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 $\mathcal{S}$. This partition consists of non-overlapping rectangular regions L which we call leaves of the classification tree. Let $L(z)$ be the element of $\mathcal{S}$ containing the point z. $L(z)$ is the leaf of the classification tree containing z.

The idea behind the construction of the classification tree is the following. Suppose that L is a leaf of the classification tree. If $L \subseteq \mathcal{Z}^{kK}$ for some k, then if $z \in L$ there exists the parameter value θ^{kK} such that the moment function $E[\rho(W, \theta) \,|\, X = x]$ is bounded by a_K uniformly over components x that are not in z. However, by Assumption 1 if $L \not\subset \mathcal{Z}^{kK}$ for any k, then the norm of the moment function will not be close to zero.

We associate the unknown conditional expectation $E[\cdot \,|\, X = x]$ with an infinite-dimensional parameter which we denote $\eta \in \mathcal{H}$. Then we consider an estimator for the moment function $m(x; \theta, \eta) = E[\rho(W, \theta) \,|\, X = x]$, denoting it $\widehat{m}(x; \theta)$. We take weighting function $A(\cdot) : \mathcal{X} \mapsto \mathbb{R}^p$ such that $E[A(X)A(X)'] < \infty$ and $E[A(X)\frac{\partial m(X;\theta,\eta)}{\partial \theta'}]$ has full rank for each $\eta \in \mathcal{H}$ and

for all θ in some fixed neighborhood of θ^{kK}. In that case the finite-dimensional parameter of interest θ_k is identified from any leaf $L \subseteq \mathcal{Z}_k$ that generates function

$$M_L(\theta, \eta) = E\left[A(X)m(X;\theta,\eta)\,\mathbf{1}\{Z \in L\}\right] \tag{3.13}$$

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

$$\widehat{M}_L(\theta,\widehat{\eta}) = \frac{1}{n\,P_Z(L)} \sum_{i\,:\,z_i \in L} w(x_i)m(x_i;\theta,\widehat{\eta}) \tag{3.14}$$

The classification will be based on the norm $\|\cdot\|$ and the threshold $\underline{M}_n > 0$. The threshold sequence $\underline{M}_n$ is calibrated based on the property of the moment function. In the next section we provide an explicit expression for this threshold. For partition $\mathcal{S}$ we define the classification tree such that for each element of partition

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

and

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

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 $\|\theta_{\mathcal{S}}(L) - \theta_{\mathcal{S}}(L')\| \leq \delta_n$. Parameter δ_n is sample-dependent and is chosen based on the properties of the estimator. In many cases encountered in practice the choice $\delta_n = O(K/n)$ will be sufficient.

Wager and Athey (2015) propose to use an application of a cross-validation procedure to

evaluate the tree splits. The main idea of that procedure is that partitioning and estimation uses different independent subsamples. We adapt this idea to the evaluation of the moment classification trees. In this case, split the sample into two subsamples where one subsample is used to estimate the moment functions $m(x;\theta,\widehat{\eta})$ and the other one is used to split $\mathcal{Z}$ into rectangles. The estimation procedure implemented this way does not induce dependence between the observations that are used to construct the moment function within each leaf of the tree.

To implement the procedure we take the sample $\{y_i, x_i, z_i\}_{i=1}^n$. First, we then 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 the 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 the estimated moment function well approximates the true moment function defined by (3.11). Then we consider a standard Euclidean norm $\|\cdot\|$ and compute the overall discrepancy between the true and empirical moment for a given L within the validation sample (which we call the prediction error) as

$$\sum_{i\in\mathcal{D}_v}\sum_{L\in\mathcal{S}} \|A(x_i)m(x_i;\theta_{0L},\eta_0)\mathbf{1}\{z_i\in L\} - M_L(\widehat{\theta},\widehat{\eta})\mathbf{1}\{z_i\in L\}\|^2,$$

where η_0 is the true value of the infinite-dimensional parameter and

$$\theta_{0L} = \arg\inf_{\theta} \|E\left[A(X)m(X;\theta,\eta_0)\,\mathbf{1}\{Z\in L\}\right]\|$$

and

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

The prediction error can be further re-written as

$$\sum_{i\in\mathcal{D}_v}\sum_{L\in\mathcal{S}} \Big(\|A(x_i)m(x_i;\theta_{0L},\eta_0)\|^2 + \|M_L(\widehat{\theta}_L,\widehat{\eta})\|^2 - 2\,m(x_i;\theta_{0L},\eta_0)'A(x_i)'\,M(\widehat{\theta}_L,\widehat{\eta}))\Big)\mathbf{1}\{z_i\in L\}.$$

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 each $L \in \mathcal{S}$, then we can re-write

$$\sum_{i\in\mathcal{D}_v} m(x_i;\theta_{0L},\eta_0)'A(x_i)'\,M_L(\widehat{\theta}_L,\widehat{\eta})) = \sum_{L\in\mathcal{S}}\sum_{i\,:\,z_i\in L} m(x_i;\theta_{0L},\eta_0)'A(x_i)'\,M_L(\widehat{\theta}_L,\widehat{\eta})).$$

As we show further, whenever $L \subseteq \mathcal{Z}^{kK}$ then

$$\frac{1}{nP_Z(L)}\sum_{i\,:\,z_i\in L} A(x_i)m(x_i;\theta_{0L},\eta_0) = M(\widehat{\theta}_L,\widehat{\eta})) + o_p(1).$$

This means that

$$\begin{aligned}\sum_{i\in\mathcal{D}_v} m(x_i;\theta_{0L},\eta_0)'A(x_i)'\,M(\widehat{\theta}_L,\widehat{\eta})) &= \sum_{L\in\mathcal{S}} \#\{i\,:\,z_i\in L\}M(\widehat{\theta}_L,\widehat{\eta}))'\;M(\widehat{\theta}_L,\widehat{\eta})) + o_p(1)\\ &= \sum_{i\in\mathcal{D}_v} \|M(\widehat{\theta}_L,\widehat{\eta}))\|^2 + o_p(1)\end{aligned}$$

and the prediction error whenever $L \subseteq \mathcal{Z}^{kK}$ can be re-written as

$$\sum_{i\in\mathcal{D}_v} \|A(x_i)m(x_i;\theta_{0L},\eta_0)\|^2 - \sum_{i\in\mathcal{D}_v} \|M(\widehat{\theta}_L,\widehat{\eta}))\|^2 + o_p(1).$$

In other words, the partition that minimizes the squared deviation of the estimated moment function for each leaf within $\mathcal{Z}^{kK}$ from the true moment function has to maximize the variance of the estimated moment function. This result extends the observation made in Athey and Imbens (2015) 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 partitions $\mathcal{Z}$ into subsets $S^{(k,1)},\ldots,S^{(k,k+1)}$. Next, for each $i = 1,\ldots,k+1$ and each dimension d we consider threshold c that generates the new partition $S^{(k+1,1)}(i,c,d),\ldots,S^{(k+1,k+2)}(i,c,d)$ 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

$$\widehat{M}^{(k+1,j)}_{i,c,d}(\theta,\widehat{\eta}) = \frac{1}{nP_Z\left(S^{(k+1,j)}(i,c,d)\right)}\sum_{i\,:\,z_i S^{(k+1,j)}(i,c,d)} A(x_i)m(x_i;\theta,\widehat{\eta}).$$

Then we find the set of minimizers

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

Note that we need to compute this only in the newly created elements of the partition, while functions $\widehat{M}$ and their minimizers on the remaining elements of the 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.[10]

3.3 Uniform consistency of the empirical moment function with heterogeneous data generating process

Our next step will be to prove consistency of the estimator for the target parameter that is the outcome of sequential minimization of the GMM criterion subject to the data partitioning induced by classification trees.

As for any semiparametric model, this analysis requires bounding the complexity of the class of functions that form the empirical moment and the empirical objective function. A distinctive feature of such an analysis in the presense of data partitioning from the classification trees, is that only a fraction of the sample is used to estimate the moment function at a given part of the support of covariates. That requires us to use a modified notion of complexity of empirical functions that accounts for such a downsampling of the data. Our analysis allows us to directly link that complexity to the lower bound for the uniform convergence rate of the empirical moment function as well as the convergence rate of the target parameter.

The classification tree induces a structure on the empirical moment where it is a product of the weighted empirical moment function $A(\cdot)m(\cdot; \theta, \widehat{\eta})$ and the containment indicator $\mathbf{1}\{\cdot \in L\}$ indicating that a given observation is contained in leaf L of the classification tree. We can interpret this as an effective reduction of the sample using the containment

[10] While our discussion of the honest tree presumes an infinite amount of data, we note that, in finite samples, the estimates of some leaves in the second tree may not have enough variation to produce valid moments. For example, the first tree may split on combinations of levels of Z that do not appear in the second data set. If this is the case, we will "prune" back that leaf to its parent node recursively until a valid parameter vector is found.

indicator. In order to work with this type of moment function and establish consistency of the corresponding Z-estimator we need to evaluate the complexity of this class.

We recall the following definition from empirical process theory which uses the concept of Radamacher random variables, which take values ± 1 with equal probability $\frac{1}{2}$.

DEFINITION 1. *For the class of functions $\mathcal{F} = \{f(\cdot; h),\ h \in H\}$ indexed by a subset H of the Banach space, i.i.d. random variables x_i and i.i.d. Radamacher random variables σ_i the empirical Radamacher complexity of $\mathcal{F}$ is*

$$\widehat{\mathcal{R}}(\mathcal{F}) = E\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\, f(x_i) \,\middle|\, x_1, \ldots, x_n\right].$$

and the Radamacher complexity of $\mathcal{F}$ is

$$\mathcal{R}(\mathcal{F}) = E\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\, f(x_i)\right].$$

The notion of the Radamacher complexity is central in establishing the uniform behavior of the empirical sum of functions of i.i.d. random variables within a given functional class. The rate of decay of the Radamacher complexity determines the rate of uniform convergence of the empirical sum to the corresponding expectation.

The presence of the containment indicator in the definition of the moment function induced by the classification tree changes the structure of the empirical sum. With this indicator in place, only a fraction of the sample will impact that empirical sum. Consequently, in the presence of this indicator we expect the uniform convergence of the empirical sum to change. Gu, Komarova, and Nekipelov (2017) find it useful to modify the notion of the Radamacher complexity to account for the presence of containment indicators. The resulting notion is referred to as Bernoulli-downsampled Radamacher complexity.

DEFINITION 2. *For the class of functions $\mathcal{F} = \{f(\cdot; h),\ h \in H\}$, i.i.d. random variables x_i, i.i.d. Radamacher random variables σ_i, and i.i.d. Bernoulli random variables ξ_i with parameter π, the Bernoulli-downsampled empirical Radamacher complexity of class $\mathcal{F}$ is*

$$\widehat{\mathcal{B}}_\pi(\mathcal{F}) = E\left[\sup_{f\in\mathcal{F}} \frac{1}{n\pi}\sum_{i=1}^{n} \sigma_i\xi_i\, f(x_i) \,\middle|\, x_1, \ldots, x_n\right]$$

and the Bernoulli-downsampled Radamacher complexity of class $\mathcal{F}$ *is*

$$\mathcal{B}_\pi(\mathcal{F}) = E\left[\widehat{\mathcal{B}}_\pi(\mathcal{F})\right].$$

The notion of Bernoulli-downsampled Radamacher complexity allows us to impose a condition directly on the class of moment functions that then ensures the validity of the uniform law of large numbers and guarantees that the Z-estimator that results from the classification tree converges.

ASSUMPTION 2. *Suppose that* $d_\Theta(\cdot,\cdot)$ *and* $d_\mathcal{H}(\cdot,\cdot)$ *are metrics in the spaces* Θ *and* $\mathcal{H}$ *respectively and moment function* $m(\cdot;\theta,\eta)$ *is defined in the product space* $\Theta \times \mathcal{H}$ *endowed with the corresponding product metric. Consider the class of functions*

$$\mathcal{F}_{\theta,\delta} = \{f(\cdot) = A(\cdot)\left(m(\cdot;\theta',\eta) - m(\cdot;\theta,\eta)\right),\ \theta' \in \Theta,\ d_\Theta(\theta',\theta) < \delta,\ \eta \in \mathcal{H}\}.$$

For each bounded converging sequence π_n *such that* $n\pi_n \to \infty$ *let* $\mathcal{B}_{\pi_n}(\mathcal{F}_{\theta,\delta})$ *be the Bernoulli-downsampled Radamacher complexity of class* $\mathcal{F}_{\theta,\delta}$*. We assume that there exists a function* $\gamma(\delta;\pi_n,n) > 0$ *such that* $\mathcal{B}_{\pi_n}(\mathcal{F}_{\theta,\delta}) \leq \gamma(\delta;\pi_n,n)$ *and* $\lim_{n\to\infty} \gamma(\delta;\pi_n,n) = 0$ *for each* δ*.*

Assumption 2 imposes a high-level condition that establishes the rate of uniform convergence for the empirical moment that is weighted by a stochastic binary sequence. We can provide the following lower level result that explicitly expresses the bound on the Bernoulli-downsampled Radamacher complexity for classes of moment functions with low entropy. As expected, this bound yields a guarantee of uniform convergence with the rate $O(\sqrt{n\pi_n})$ for classes of bounded functions with low entropy.

LEMMA 1. *Suppose that* $\mathcal{F}$ *is a class of measurable functions with envelope* F *(i.e., for each* $f(\cdot) \in F$*,* $|f| \leq F$*) with covering number* $N(\epsilon,\mathcal{F},L_2(Q))$ *and uniform covering integral*

$$J(\theta,\mathcal{F}) = \sup_Q \int_0^\theta \sqrt{1 + \log\ N(\epsilon\|F\|_Q,\mathcal{F},L_2(Q))}\, d\epsilon,$$

increasing in θ*. Then Bernoulli downsampled Radamacher complexity of class* $\mathcal{F}$ *for bounded converging sequence* π_n *can be bounded as*

$$\mathcal{B}_\pi(\mathcal{F}) \leq K\frac{J(1,\mathcal{F})\, R[F(Z)^2]^{1/2}}{\sqrt{n\,\pi_n}},$$

for a universal constant K.

Proof:
For a sequence of i.i.d. Bernoulli random variables ξ_i with parameter π and a sequence of i.i.d. Radamacher random variables σ_i we consider the process

$$(f) = \frac{1}{\sqrt{n\,\pi_n}} \sum_{i=1}^{n} \xi_i\,\sigma_i\,f(z_i)$$

By Hoeffding's inequality this process is sub-Gaussian with the $L_2(\mathbb{P}_n)$ seminorm

$$\|f\|_n = \left(\frac{1}{n}\sum_{i=1}^{n} f^2(z_i)\right)^{1/2}.$$

In fact, note that for $f, g \in \mathcal{F}$ we can compute the standard deviation metric

$$\sigma_{\xi,\sigma}(f-g) = \left(\mathrm{Var}\left(\frac{1}{\sqrt{n\,\pi_n}}\sum_{i=1}^{n}\xi_i\,\sigma_i\,(f(x_i)-g(x_i))\,\middle|\,x_1,\ldots,x_n\right)\right)^{1/2} = \left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-g(x_i))\right)^{1/2}.$$

Suppose that $N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n))$ is the $L_2(\mathbb{P}_n)$-covering number of the class of functions $\mathcal{F}$. Suppose that $\eta_n = \sup_{\mathcal{F}} \|f\|_n$. Then we can apply Dudley's approach to bound the Orlicz norm conditional on the observations $x_1, \ldots, x_n$ as

$$\left\|\sup_{\mathcal{F}}(f)\right\|_{\psi_2} \leq K \int_0^{\eta_n} \sqrt{1+\log\,N(\epsilon,\mathcal{F},L_2(\mathbb{P}_n))}\,d\epsilon,$$

for some constant K. Then let $\epsilon^* = \epsilon/\|F\|_n$, which leads to

$$\int_0^{\eta_n} \sqrt{1+\log\,N(\epsilon,\mathcal{F},L_2(\mathbb{P}_n))}\,d\epsilon = \|F\|_n \int_0^{\theta_n} \sqrt{1+\log\,N(\epsilon^*\|F\|_n,\mathcal{F},L_2(\mathbb{P}_n))}\,d\epsilon^*.$$

Recall that the uniform entropy integral is defined as

$$J(\theta,\mathcal{F}) = \sup_Q \int_0^{\theta} \sqrt{1+\log\,N(\epsilon\|F\|_Q,\mathcal{F},L_2(Q))}\,d\epsilon,$$

where the supremum is taken over all measures Q. Noticing that $\theta_n \leq 1$, we can bound

$$E_{\xi,\sigma}\left[\sup_{\mathcal{F}}(f)\right] \leq K\,J(1,\mathcal{F})\,\|F\|_n.$$

Finally, taking the expectation over the sample and applying Jensen's inequality allows us to evaluate

$$E\left[\sup_{\mathcal{F}} \frac{1}{\sqrt{n\,\pi_n}} \sum_{i=1}^{n} \xi_i\,\sigma_i\,f(x_i)\right] \leq K\,J(1,\mathcal{F})\,E[F(Z)^2]^{1/2}.$$

As a result, we can bound the Bernoulli downsampled Radamacher complexity by

$$\mathcal{B}_{\pi_n}(\mathcal{F}) \leq K\frac{J(1,\mathcal{F})\,E[F(Z)^2]^{1/2}}{\sqrt{n\,\pi_n}}$$

Q.E.D.

Lemma 1 indicates that the notion of Bernoulli-downsampled Radamacher complexity is well-defined and can be applied to the low complexity classes of moment functions. Our next step is to establish that Assumption 2 guarantees the validity of the uniform law of large numbers within the class of moment functions induced by classification trees.

LEMMA 2. *Suppose that function $m(x;\cdot,\eta)$ is Lipschitz-continuous for each $x \in \mathcal{X}$ and $\eta \in \mathcal{H}$ with a universal Lipschitz constant. Under Assumption 2 and the leaf of classification tree L such that $P_Z(L) > 0$ for functions (3.14) and (3.13), $\sqrt{n\,P_Z(L)}\gamma\,(\delta; P_Z(L),\,n) = o(\delta)$ as $n \to \infty$ and each $\theta \in \Theta$*

$$\sup_{d_\Theta(\theta',\theta)<\delta,\,\eta\in\mathcal{H}} \left|\widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta)\right| = o_p(1).$$

Proof:
To establish the uniform law of large numbers we need to derive a large deviation bound for

$$\begin{aligned}
&\sup_{d_\Theta(\theta',\theta)<\delta,\,\eta\in\mathcal{H}} \left|\widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta)\right| \\
&\qquad \leq \max\left\{ \sup_{d_\Theta(\theta',\theta)<\delta,\,\eta\in\mathcal{H}} \left(\widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta)\right),\right. \\
&\qquad\qquad \left. \sup_{d_\Theta(\theta',\theta)<\delta,\,\eta\in\mathcal{H}} \left(\widehat{M}_L(\theta,\eta) - M_L(\theta,\eta) - \widehat{M}_L(\theta',\eta) + M_L(\theta',\eta)\right) \right\}.
\end{aligned}$$

Applying the union bound, we have

$$
\begin{aligned}
&P\Bigg(\sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \Big| \widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta) \Big| \geq t \Bigg) \\
&\quad \leq P\Bigg(\sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \Big(\widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta) \Big) \geq t \Bigg) \\
&\quad + P\Bigg(\sup_{\theta\in\Theta} \sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \Big(\widehat{M}_L(\theta,\eta) - M_L(\theta,\eta) - \widehat{M}_L(\theta',\eta) + M_L(\theta',\eta) \Big) \geq t \Bigg).
\end{aligned}
$$

We focus on the first term of this evaluation and the second term can be bounded in an analogous way.

To do that consider

$$
\begin{aligned}
&E_{\xi_i,x_i}\left[\sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \Big(\widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta) \Big) \right] \\
&\quad = E_{\xi^n,x^n}\Bigg[\sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} E_{\xi_i',z_i'}\Bigg[\frac{1}{nP_Z(L)} \sum_{i=1}^n \xi_i\, A(x_i)\, (m(x_i;\theta',\eta) - m(x_i;\theta,\eta)) \\
&\qquad\quad - \frac{1}{nP_Z(L)} \sum_{i=1}^n \xi_i'\, A(x_i')\, (m(x_i';\theta',\eta) - m(x_i';\theta,\eta)) \Bigg]\Bigg] \\
&\quad \leq E_{\xi_i,x_i,\xi_i',x_i'}\Bigg[\sup_{\theta\in\Theta,\eta\in\mathcal{H}} \frac{1}{nP_Z(L)} \sum_{i=1}^n \Bigg(\xi_i\, A(x_i)\ (m(x_i;\theta',\eta) - m(x_i;\theta,\eta)) \\
&\quad - \xi_i'\, A(x_i')\, (m(x_i';\theta',\eta) - m(x_i';\theta,\eta)) \Bigg)\Bigg],
\end{aligned}
$$

where $x^n = \{x_1,\ldots,x_n\}$ and $\xi^n = \{\xi_1,\ldots,\xi_n\}$ and $\xi_i = \mathbf{1}\{z_i \in L\}$.

Given that the data are i.i.d., the distribution of pairs of random variables in the supremum is symmetric. Therefore, we can apply the standard symmetrization argument and use Radamacher random variables σ_i to obtain

$$
\begin{aligned}
&E_{\xi_i,x_i}\left[\sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \Big(\widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta) \Big) \right] \\
&\quad \leq 2\, E_{\xi_i,x_i}\left[\sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \frac{1}{nP_Z(L)} \sum_{i=1}^n \sigma_i\, \xi_i\, A(x_i)\ (m(x_i;\theta',\eta) - m(x_i;\theta,\eta)) \right]
\end{aligned}
$$

The bound on the right is the Bernoulli-downsampled Radamacher complexity. Let $x^{n,i}(x_i') =$

$\{x_1, \ldots, x_i', \ldots, x_n\}$, where we replace x_i in S by x_i'. Denote

$$Q(x^n) = \sqrt{nP_Z(L)} \sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \left(\widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta)\right)$$

when sample x^n is used. Let $\widehat{M}_L'(\theta,\eta)$ correspond to the empirical moment computed from sample $x^{n,i}(x_i')$. Then

$$\begin{aligned}
|Q(x^n) - Q(x^{n,i}(x_i'))| &\le \sqrt{nP_Z(L)} \Bigg| \sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \left(\widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta)\right) \\
&\quad - \sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \left(\widehat{M}_L'(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L'(\theta,\eta) + M_L(\theta,\eta)\right) \Bigg| \\
&\le \sqrt{nP_Z(L)} \sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \left|\widehat{M}_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) - \widehat{M}_L'(\theta',\eta) - \widehat{M}_L'(\theta,\eta)\right| \\
&= \sqrt{nP_Z(L)} \sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \Bigg| \frac{1}{nP_Z(L)} A(x_i) \left(m(x_i;\theta',\eta) - m(x_i;\theta,\eta)\right) \\
&\quad - \frac{1}{nP_Z(L)} A(x_i') \left(m(x_i';\theta',\eta) - m(x_i';\theta,\eta)\right) \Bigg| \\
&\le \frac{2\lambda B\delta}{\sqrt{nP_Z(L)}},
\end{aligned}$$

where B is the universal Lipschitz constant. Thus, by McDiarmid's inequality, we have

$$\begin{aligned}
P(Q(x^n) - E[Q(x^n)] \ge t) &\le \exp\left(\frac{-2t^2}{\sum_{i=1}^n (2B\lambda\delta)^2/(nP_Z(L))}\right) \\
&= \exp\left(\frac{-P_Z(L)^2 t^2}{2B^2\lambda^2\delta^2}\right).
\end{aligned}$$

Thus, with probability at least $1 - \alpha/2$, we have

$$Q(x^n) \le E[Q(x^n)] + B\lambda\delta\sqrt{\frac{2\log(2/\alpha)}{P_Z(L)}}.$$

Given that identical analysis applies to $\sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} (M_L(\theta,\eta) - \widehat{M}_L(\theta,\eta))$, we can combine

the results to establish that

$$P\Bigg(\sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \left| \widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta) \right| \\ \geq \gamma\left(\delta; P_Z(L),\, n\right) + \frac{B\lambda\delta}{P_Z(L)}\sqrt{\frac{2\log(2/\alpha)}{n}} \Bigg) \leq \alpha.$$

Therefore, it is sufficient for the uniform law of large numbers to hold when $\alpha\, n^r \to \infty$ for some $r > 1$ when

$$\sup_{d_\Theta(\theta',\theta)<\delta,\, \eta\in\mathcal{H}} \left| \widehat{M}_L(\theta',\eta) - M_L(\theta',\eta) - \widehat{M}_L(\theta,\eta) + M_L(\theta,\eta) \right| = O_p\left(\gamma\left(P_Z(L),\, n\right)\right) + o_p\left(\frac{1}{P_Z(L)}\sqrt{\frac{\log\, n}{n}} \right).$$

By the Borel-Cantelli lemma the convergence will also occur almost surely.
Q.E.D.

3.4 Consistency and asymptotic normality of the target parameter

Having established uniform consistency of the empirical moment function, we can use the standard arguments from the theory of semiparametric Z-estimators to establish consistency and asymptotic normality of the target parameter.

Lemma 2 ensures that whenever $L \subseteq \mathcal{Z}^{kK}$ then the uniform law of large numbers holds for $\theta = \theta^{k*}$ and, as a result, for the Z-estimator. The Z-estimator is constructed by finding a solution to

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

producing an estimator $\widehat{\theta}^{kK}$ that is consistent for θ^{k*}, which is justified given that $a_K \to 0$ as K grows (and thus the moment function provides a higher quality uniform approximation).

We impose conditions on the behavior of the population moment function (3.13) in the leaves of the classification tree.

ASSUMPTION 3. *For any $\theta_0 \in \Theta$ and $\eta_0 \in \mathcal{H}$ there exists $\bar{\delta} > 0$ such that for all $0 < \delta < \bar{\delta}$ and all $(\theta, \eta) \in \Theta \times \mathcal{H}$ such that $d_\Theta(\theta, \theta_0) \leq \delta$ and $d_\mathcal{H}(\eta, \eta_0) \leq \delta$ there exists a non-singular matrix $J(\theta_0, \eta_0)$ and linear functional $I(\theta_0, \eta_0)[\cdot]$ such that*

$$M_L(\theta, \eta) = M_L(\theta_0, \eta_0) + J(\theta_0, \eta_0)(\theta - \theta_0) + I(\theta_0, \eta_0)[\eta - \eta_0] + R(\delta)$$

where the eigenvalues of $J(\theta_0, \eta_0)$ *and the norm of* $I(\theta_0, \eta_0)[\cdot]$ *are universally bounded for all* $\theta_0 \in \Theta$ *and* $\eta_0 \in \mathcal{H}$ *for each* L *such that* $P_Z(L) > 0$ *(where the bound may depend on* L*) and* $R(\delta) = O(\delta^2)$.

Assumption 3 requires the existence of the local linear approximation for the population moment function $M_L(\theta, \eta)$. This assumption may be directly verified for specific moment functions. In general, it will be satisfied whenever either the distribution W or the moment function $\rho(\cdot, \cdot)$ is sufficiently smooth.

Our next assumption requires the existence of a "high quality" estimator for parameter η.

ASSUMPTION 4. *There exists an estimator* $\widehat{\eta}$ *that converges to its population counterpart* η_0 *in* $\mathcal{H}$ *and for sequence* $r_n \to \infty$ *such that* $\sup_{\theta \in \Theta, x \in \mathcal{X}} \| r_n \left(m(x; \theta, \widehat{\eta}) - m(x; \theta, \eta_0) \right) \| = o_p(1)$.

This assumption allows us to evaluate

$$\left\| \widehat{M}_L(\widehat{\theta}, \widehat{\eta}) - \widehat{M}_L(\widehat{\theta}, \eta_0) \right\| \leq \sup_{\theta \in \Theta, x \in \mathcal{X}} \| m(x; \theta, \widehat{\eta}) - m(x; \theta, \eta_0) \| \frac{1}{n P_Z(L)} \sum_{i=1}^{n} A(x_i) \mathbf{1}\{z_i \in L\},$$

which ensures that the corresponding term vanishes.

The last assumption is a generalization of the Central Limit Theorem for empirical functions $\widehat{M}_L(\theta, \eta)$. This assumption can be verified by directly computing the moments of random variable $A(X) m(x; \theta, \eta)$ for each fixed η and θ.

ASSUMPTION 5. *For each* $(\theta, \eta) \in \Theta \times \mathcal{H}$ *and sequence of leaves* L_n *such that* $P_Z(L_n) > 0$ *and the limit* $P_Z(L_n)$ *exists, we can find a positive definite matrix* Ω *such that*

$$\sqrt{n\, P_Z(L_n)} \left(\widehat{M}_L(\theta, \eta) - M_L(\theta, \eta) \right) \xrightarrow{d} N(0,\, \Omega).$$

Now we derive the asymptotic distribution for Z-estimator $\widehat{\theta}_L$ for parameter θ^{kK} when we found the leaf $L \subseteq \mathcal{Z}^{kK}$. We assume now that K grows and Assumption 1 is satisfied.

THEOREM 1. *Under conditions of Lemma 2 and Assumptions 1-5 for* $K \to \infty$ *such that* $a_K \sqrt{n/K} \to 0$ *and* $b_K \sqrt{n/K} \to \infty$ *for each* $L \in \mathcal{Z}^{kK}$

$$\sqrt{\frac{n}{K}} \left(\widehat{\theta}_L - \theta^{*k} \right) \xrightarrow{d} N(0,\, J(\theta^{*k}, \eta_0)^{-1} \Omega\, J(\theta^{*k}, \eta_0)^{-1\prime})$$

The result of the theorem follows immediately from Theorem 3.3.1 in Van Der Vaart and Wellner (1996) applied to the mapping $\widehat{M}_L(\cdot, \eta_0)$ in light of our assumption regarding the estimator $\widehat{\eta}$ taking into account that $P_Z(L) = O(1/K)$.

We define a global estimator $\widehat{\theta}(z) = \theta_{\mathcal{S}}(L)$ if $z \in L$ and the population object

$$\theta^*(z) = \begin{cases} \theta^{*k}, \text{ if } z \in L \text{ and } L \subseteq \mathcal{Z}^{kK}, \\ \emptyset, \text{ otherwise.} \end{cases}$$

Then Theorem 1 is the uniform convergence result that the global estimator over $\mathcal{Z}$ converges to the population object $\theta^*(z)$ which defines the heterogeneity structure.

We proved this for a fixed partition $\mathcal{S}$ and the leaf of this partition L. The classification tree uses an independent training sample $\mathcal{D}_t$ to implement the partition of $\mathcal{Z}$ into rectangular areas. Theorem 1 specifies the number of sets K (depending on n) in the set system that define the structure of heterogeneity. The classifier then takes a given leaf L and identifies whether $L \subseteq \mathcal{Z}^{kK}$ for some K if $\|\widehat{M}_L(\widehat{\theta}, \widehat{\eta})\| = O(a_K)$. The probability of misclassification approaches 0 given that $\|\widehat{M}_L(\widehat{\theta}, \widehat{\eta})\| \gg O(a_K)$ whenever $L \not\subset \mathcal{Z}^{kK}$ for some k. Wager and Walther (2015) characterize the class of rectangles that result from recursive partitioning and show that for any rectangular set one can construct a rectangular subset from this class. Walther et al. (2010) shows that a fixed rectangular subset can be approximated with probability approaching 1 by finding the maximum of the blocked scan statistic over rectangles whenever $P_Z(\mathcal{Z}^{kK}) > O(\log n/n)$. As a result, the classifier that generates the partition $\mathcal{S}$ produces consistent subsets $\mathcal{Z}^{kK}$.

3.5 Extension to Moment Forests

Following the insights of Breiman (1996, 2001), we extend our moment trees to moment forests by generalizing the single estimator described above to an ensemble and introducing randomness. Moment forests are composed of B moment trees, where each tree is estimated on resampled data. Each tree has one additional difference: the set of covariates considered for splitting at each node is restricted to be a random subset of the full set of covariates. The estimated parameters are then (inverse-variance weighted) averages of the estimated parameter in each tree. This extension introduces randomness along two dimensions: the resampling of data and the random set of covariates considered at each split when growing the trees. Both of these injections of randomness reduce the correlation across trees, which theoretically can lower the variance of the estimated parameters. Furthermore, by resampling

all of the data across the trees in the forest, all of the statistical information in the underlying sample is used to construct estimates, which increases overall efficiency. This is opposed to a single tree, which throws out some of the data used to estimate the structure of the tree in the first stage when estimating the parameters in the second stage. Restricting the number of covariates considered at each split also has a computational benefit as well, as each additional covariate added to a forest does not increase the runtime multiplicatively.

While the results above are written in terms of a single moment tree, our results also extend to a moment forest. From a theoretical perspective, a moment forest can be viewed as a collection of classification trees that are constructed by independent binary perturbations of split locations, both by the resampling of the input data but also by the restricted set of splits considered in each leaf. The moment forest can then be generically represented as a moment tree where each possible split occurs based on the realization of an independent sequence of Bernoulli random variables. Due to the independence of this sequence (by construction), we can consider the expectation of the empirical moment function conditional on that sequence. The resulting analysis produces the uniform evaluation for the empirical moment function which is identical to our analysis for a single tree in the previous subsection. Since that uniform evaluation holds for any given sequence of splits, we can take expectations over the random variables generating the splits and the bound remains valid. That bound is exactly the uniform bound for the estimator produced by the moment forest.

4 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.

For all of the following simulations, we use the following process. We run a Monte Carlo simulation 500 times for each set of parameters. In each simulation, we first generate data according to a DGP specified below, with some sample size n. Second, we follow Algorithm 2 to generate a random forest: we resample datasets of size n from the generated data 50 times, each time generating one structure tree and one estimation tree. We then take the weighted mean of all of the tree estimates, which gives us a parameter estimate for each observation in

the data. The three tuning parameters—k, the minimum number of observations in each leaf; α, the minimum proportion of data in each leaf; and $\overline{MSE}$, the minimum improvement in MSE after each split—are determined through cross-validation from the raw data.[11] Because the number of X variables in these examples is small, we include all of them as potential splitting variables in each tree; Section 5 presents an empirical example where trees are generated from splits on subsets of the X variables.

As a benchmark, we compare our moment forest estimates to set of conventional estimates which are calculated by running OLS on each discrete subgroup of the data.[12] To evaluate each approach, we calculate mean squared prediction error (MSPE) between the estimated parameters of interest and the true parameters. This prediction error captures both statistical sampling error and model specification error.

4.1 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:

$$Y = \tau(X) \cdot W + \epsilon, \tag{4.15}$$

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, x_1 and x_2, 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.

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. Table 1 shows the results. We highlight two key features. First, Column 2 (Num. Leaves)

[11] Effects are similar if we recalculate the tuning parameters for each new resampling of the data.

[12] Alternatively, we could benchmark against simple pooled OLS estimates. These have relatively lower error in models with no heterogeneity, and higher error in models with substantial heterogeneity. While one could run other model selection methods and compare those here, our goal is to highlight the efficiency advantages of grouping together observations with the same treatment effects. The naive subgroup estimator is thus a natural baseline for comparison.

Table 1: Monte Carlo: Uniform RCT: $\tau(X) = 10$

n / $(\alpha, k, \overline{MSE})$	Num. Leaves	MSPE (Forest)	MSPE (OLS)
50	1.000	0.060	NaN
(0.01, 7, 1e-03)	(0.000)	(0.089)	(NaN)
100	1.000	0.044	NaN
(0.01, 13, 1e-03)	(0.000)	(0.076)	(NaN)
200	1.000	0.021	NaN
(0.01, 31, 1e-03)	(0.000)	(0.029)	(NaN)
400	1.000	0.008	0.527
(0.01, 56, 1e-03)	(0.000)	(0.012)	(0.156)
800	1.000	0.005	0.339
(0.01, 106, 1e-03)	(0.000)	(0.007)	(0.068)
1600	1.000	0.002	0.169
(0.01, 206, 1e-03)	(0.000)	(0.002)	(0.031)
3200	1.000	0.002	0.084
(0.01, 406, 1e-03)	(0.000)	(0.002)	(0.015)

shows the average number of leaves at the bottom of each moment tree; this is the number of different treatment effects that the moment forest will estimate. In this case, the moment forest identifies no meaningful treatment heterogeneity across all Monte Carlo runs, hence the estimated trees have only a single leaf. The moment forest thus consistently recovers the true underlying model. Note that the consistent rejection of heterogeneity across different simulations in this case comes in part because of the high value of k selected by the initial cross-validation.

Second, Column 3 (MSPE (Forest)) is a useful baseline against which to compare the more complex models which follow. In this simple example, the MSPE reflects only the statistical sampling error, because the moment forest has calculated the correct model in every case. The more complex models we consider next have a combination of statistical sampling and model misspecification.

The final column shows the performance of OLS run on each subgroup separately.[13] 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 each subgroup separately. The inability of OLS estimator to group together similar observations gives it

[13]There is no column analogous to "number of leaves" for OLS, because OLS is estimating the same model every time—a model with 64 different treatment effects. The OLS errors thus decline at a parametric rate in all examples.

Table 2: Monte Carlo: Group RCT
$\tau(X) = 10$ if $x_1 = 1$, $\tau(X) = 0$ otherwise

n / $(\alpha, k, \overline{MSE})$	Num. Leaves	MSPE (Forest)	MSPE (OLS)
50	3.338	5.733	NaN
(0.01, 1, 1e-03)	(0.253)	(2.248)	(NaN)
100	6.426	1.643	NaN
(0.01, 1, 1e-03)	(0.549)	(1.513)	(NaN)
200	9.762	0.157	NaN
(0.01, 1, 1e-01)	(0.916)	(0.109)	(NaN)
400	4.991	0.029	0.527
(0.01, 1, 1e-01)	(0.819)	(0.018)	(0.156)
800	4.336	0.021	0.339
(0.01, 1, 1e-01)	(0.313)	(0.016)	(0.068)
1600	2.939	0.007	0.169
(0.01, 1, 1e-01)	(0.250)	(0.007)	(0.031)
3200	2.067	0.003	0.084
(0.01, 1, 1e-01)	(0.063)	(0.003)	(0.015)
6400	2.000	0.001	0.043
(0.01, 1, 1e-01)	(0.000)	(0.001)	(0.007)

a large precision penalty at every sample size. This highlights a benefit of using the tree method even when the true model is a single treatment effect. Finally, note that the moment forest can estimate a potentially saturated heterogeneous treatment effect even with only 50 observations, whereas estimating OLS in every subgroup is not even possible without a higher observation count.

To assess the performance of the estimator when we introduce observable heterogeneity, we next set the treatment effect to equal ten if the observation has $x_1 = 1$, and zero otherwise, generating two treatment effects. Table 2 reports the results. The forest initially estimates too many splits, compared to the one split and two leaves required to fit the true model, but converges to the true model for $n > 1600$.

The MSPE 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 MSPE is driven by reductions in both errors, hence the faster-than-parametric rate of decline while the forest is converging to the true model. The forest then decreases at a parametric rate of $\sqrt{n}$ once the true model is obtained at

Table 3: Monte Carlo: Sparse RCT
$\tau(X) = 10$ if $x_1 = 1$ and $x_2 = 1$, $\tau(X) = 0$ otherwise

n / $(\alpha, k, \overline{MSE})$	Num. Leaves	MSPE (Forest)	MSPE (OLS)
50	1.679	1.649	NaN
(0.01, 5, 1e-03)	(0.112)	(0.164)	(NaN)
100	6.743	1.511	NaN
(0.01, 1, 1e-03)	(0.356)	(0.241)	(NaN)
200	13.346	1.172	NaN
(0.01, 1, 1e-02)	(0.543)	(0.347)	(NaN)
400	26.054	0.753	0.527
(0.01, 1, 1e-03)	(0.696)	(0.325)	(0.156)
800	43.860	0.359	0.339
(0.01, 1, 1e-02)	(0.991)	(0.196)	(0.068)
1600	44.984	0.105	0.169
(0.01, 1, 1e-02)	(2.283)	(0.028)	(0.031)
3200	14.131	0.011	0.084
(0.01, 1, 1e-02)	(3.149)	(0.006)	(0.015)
6400	5.760	0.003	0.043
(0.01, 1, 1e-02)	(0.274)	(0.002)	(0.007)
12800	4.277	0.001	0.020
(0.01, 1, 1e-02)	(0.216)	(0.001)	(0.003)
25600	3.250	0.001	0.010
(0.01, 1, 1e-02)	(0.111)	(0.000)	(0.002)

$n = 3200$. The OLS estimates have larger standard errors than the moment forest at all sample sizes—even before the forest has converged to the right model—due to the inability of OLS to group observations that have the same parameters.[14]

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 zero treatment effects which may appear to be true effects due to within-group statistical errors. Table 3 reports the results. There are several notable features. The true model has two branches (and thus three leaves): a split on $x_1 = 1$, and then a split on $x_2 = 1$ in the $x_1 = 1$ branch (or vice versa). There is an increase in the complexity of the tree at small to moderate samples sizes before the estimator begins to converge back to the simpler true model starting at $n = 3200$. This pattern results from the optimal tradeoff of

[14]Note that the OLS estimates are identical in Tables 1 through 4 because we use the same randomization seed, and the DGP in each subgroup only differs by a constant shift in the treatment effect.

Table 4: Monte Carlo: RCT with Saturated Heterogeneity: $\tau(X) = x_1 + 8 * (x_2 - 1)$

n / $(\alpha, k, \overline{MSE})$	Num. Leaves	MSPE (Forest)	MSPE (OLS)
50	3.720	112.964	NaN
(0.01, 1, 1e-03)	(0.267)	(56.127)	(NaN)
100	7.186	23.799	NaN
(0.01, 1, 1e-03)	(0.316)	(21.953)	(NaN)
200	13.832	4.718	NaN
(0.01, 1, 1e-03)	(0.458)	(1.028)	(NaN)
400	26.029	1.812	0.527
(0.01, 1, 1e-03)	(0.684)	(0.334)	(0.156)
800	46.237	0.458	0.339
(0.01, 1, 1e-03)	(0.964)	(0.145)	(0.068)
1600	61.930	0.171	0.169
(0.01, 1, 1e-03)	(0.314)	(0.032)	(0.031)
3200	63.860	0.089	0.084
(0.01, 1, 1e-03)	(0.058)	(0.016)	(0.015)
6400	63.964	0.045	0.043
(0.01, 1, 1e-03)	(0.029)	(0.007)	(0.007)
12800	63.998	0.021	0.020
(0.01, 1, 1e-03)	(0.006)	(0.003)	(0.003)
25600	64.000	0.010	0.010
(0.01, 1, 1e-03)	(0.000)	(0.002)	(0.002)

variance (too few observations in each leaf) versus bias (not enough partitions of the data to capture the true number of effects). The faster-than-parametric decrease in the MSPE at that threshold again reflects the decline in specification error.

It is instructive to contrast the performance of the tree against the OLS estimator. Initially, OLS cannot estimate the model at all because, unlike the moment forest, OLS requires a minimum number of observations in each subgroup. At all but the smallest sample sizes, the moment forest dominates the comparison of MSPE, even at $n = 1600$ where the forest is estimating an excessively complex model. Once the moment forest converges to the correct model, it generates far more precise estimates than OLS due to its ability to group together observations with similar treatment effects.

Finally, we consider a specification at the other extreme, where each subgroup has a different

treatment effect. We modify the data-generating process for $\tau(X)$ to be:

$$\tau(X) = x_1 + 8 * (x_2 - 1). \tag{4.16}$$

This results in 64 treatment effects as a combination of x_1 and x_2. Table 4 reports the results. Note that the optimal k is set to 1, which was 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 can 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.

At small sample sizes, the forest cannot grow complex enough to estimate the true model, leading to considerably higher prediction error than the earlier specifications. As above, MSPE converges at a faster-than-parametric rate until the moment forest approximately recovers the true model around $n = 3200$. Parametric error rates obtain after that point, reflecting the independence of the estimation of tree structure and the estimation of treatment effects, as desired.

The OLS estimator in this case has a performance as good or better than the moment tree, for all sample sizes for which it can be estimated. This is expected, because the OLS estimator in this case is the true model. Once the tree has found the true model, prediction errors are essentially identical across adjacent rows, again highlighting the independence of the honest forest's predictive performance from the model selection step. The MSPE of the two models converge once the true model is recovered in the first step. We note that there is no MSPE penalty to using the moment forest against the true model even though the forest splits the sample when estimating effects. This is because the moment forest resamples the entire data set before making the splits within each moment tree, effectively using all of the statistical information in the sample. The end result is statistical performance in the limit as good as the infeasible oracle model.

4.2 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.

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

$$Y = \tau(X) \cdot W(R) + \epsilon. \tag{4.17}$$

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

$$W = \begin{cases} 0 & \text{if } R < \bar{R}, \\ 1 & \text{else.} \end{cases} \tag{4.18}$$

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:

$$\tau(X) = \begin{cases} 5 & \text{if } X_2 < 0.67, \\ -2 & \text{else.} \end{cases} \tag{4.19}$$

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).

In contrast to the previous section, which used a moment forest, we report here the results for a single moment tree. We do this to highlight the underlying structure more clearly, as one can examine the number of splits and where they occur more clearly with a single moment tree versus the ensemble forest. All of the results presented here will be identical for the moment forest with the exception of having larger standard errors. As before, we generate 500 draws of each sample size, estimate tree structure on a 50% subsample, and estimate the RDD model in the second subsample according to that structure. We calculate mean squared prediction error by summing the squared difference from the true value, dividing by sample size, and taking the square root. Table 5 shows the results for the model above when using all the data in sample on either side of the treatment threshold ($h = 0.5$); the threshold for improvement in the tree $\overline{MSE}$ is set to 0.1.[15]

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 structure of the data without error. The RMSPE column reports the root mean squared prediction error on data generated from

[15] In principle, we could also select an optimal window size for the local linear estimator using cross-validation or an optimal bandwidth selection rule.

Table 5: Monte Carlo: RDD

n	Dim(τ)	Count Discrete	Count Continuous	Mean X_2	RMSPE
500	6.587	3.207	2.380	0.539	0.918
	(0.785)	(0.480)	(0.772)	(0.252)	(0.393)
1000	6.107	3.113	1.993	0.671	0.669
	(0.449)	(0.317)	(0.337)	(0.027)	(0.303)
2000	6.133	3.087	2.047	0.675	0.497
	(0.499)	(0.281)	(0.291)	(0.026)	(0.273)
4000	6.040	3.020	2.020	0.671	0.321
	(0.280)	(0.140)	(0.140)	(0.014)	(0.187)
8000	6.027	3.013	2.013	0.670	0.246
	(0.229)	(0.115)	(0.115)	(0.013)	(0.213)
16000	6.000	3.000	2.000	0.669	0.198
	(0.000)	(0.000)	(0.000)	(0.006)	(0.174)

the DGP without noise, uniformly distributed over the X variables.[16]

4.3 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:

$$\tau(x_i) = \sin(4\pi x_i), \tag{4.20}$$

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 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 $\alpha = 0.1$ 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.

[16]We report RMSPE rather than MSPE in this and the following table because the squared error terms become very small.

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)

4.3.1 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.

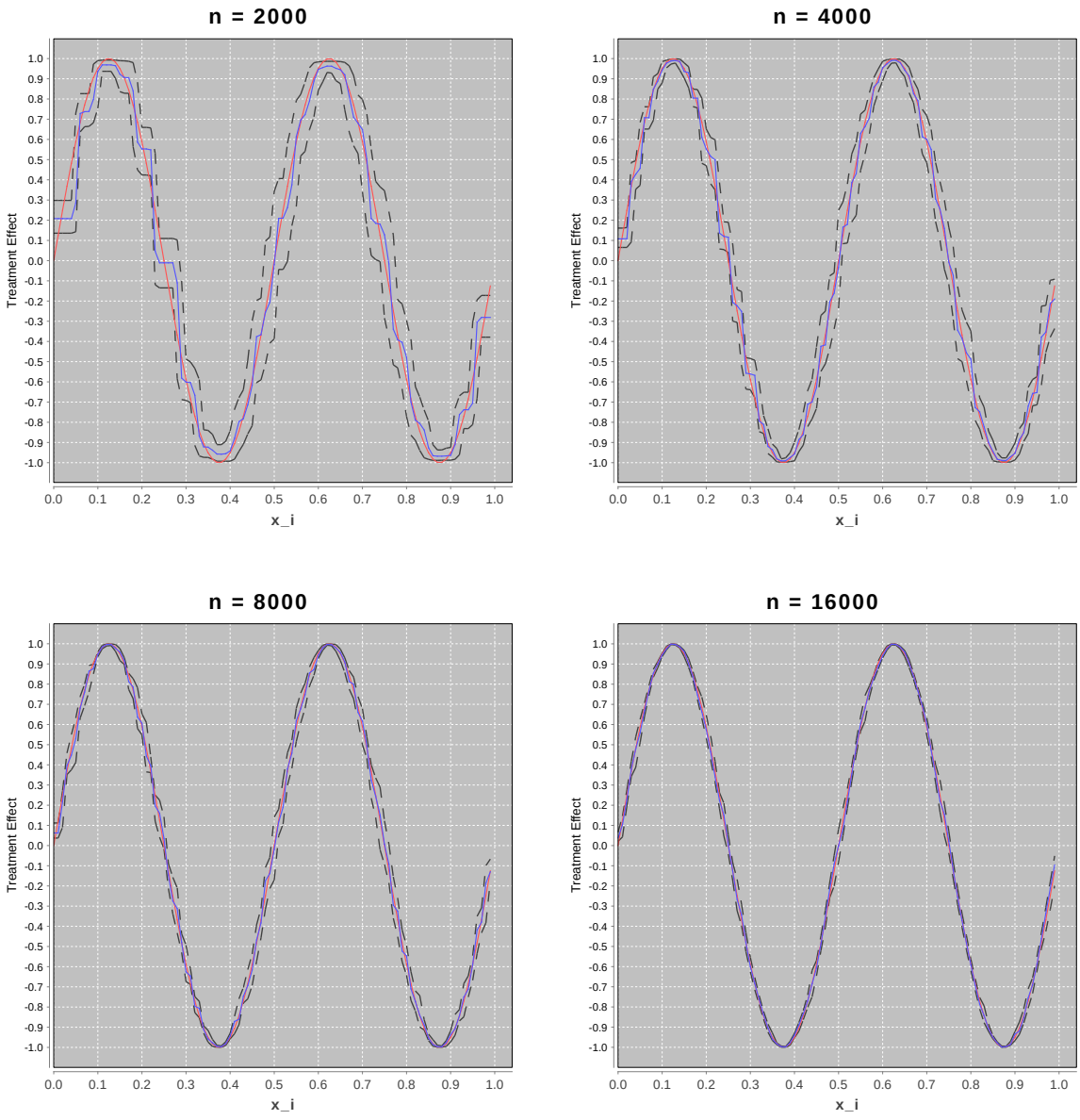
The left panel of Table 6 reports the results from this experiment. As the dimension of τ shows, the model fits an increasingly complex model to the data as sample size rises, resulting in a substantial decrease in mean squared prediction error.

Figure 1 shows the fit of the moment tree in this case. The red line shows the true model, while the blue line shows the moment tree estimate, and the dashed lines the 95% confidence interval. 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.

4.3.2 With Measurement Error

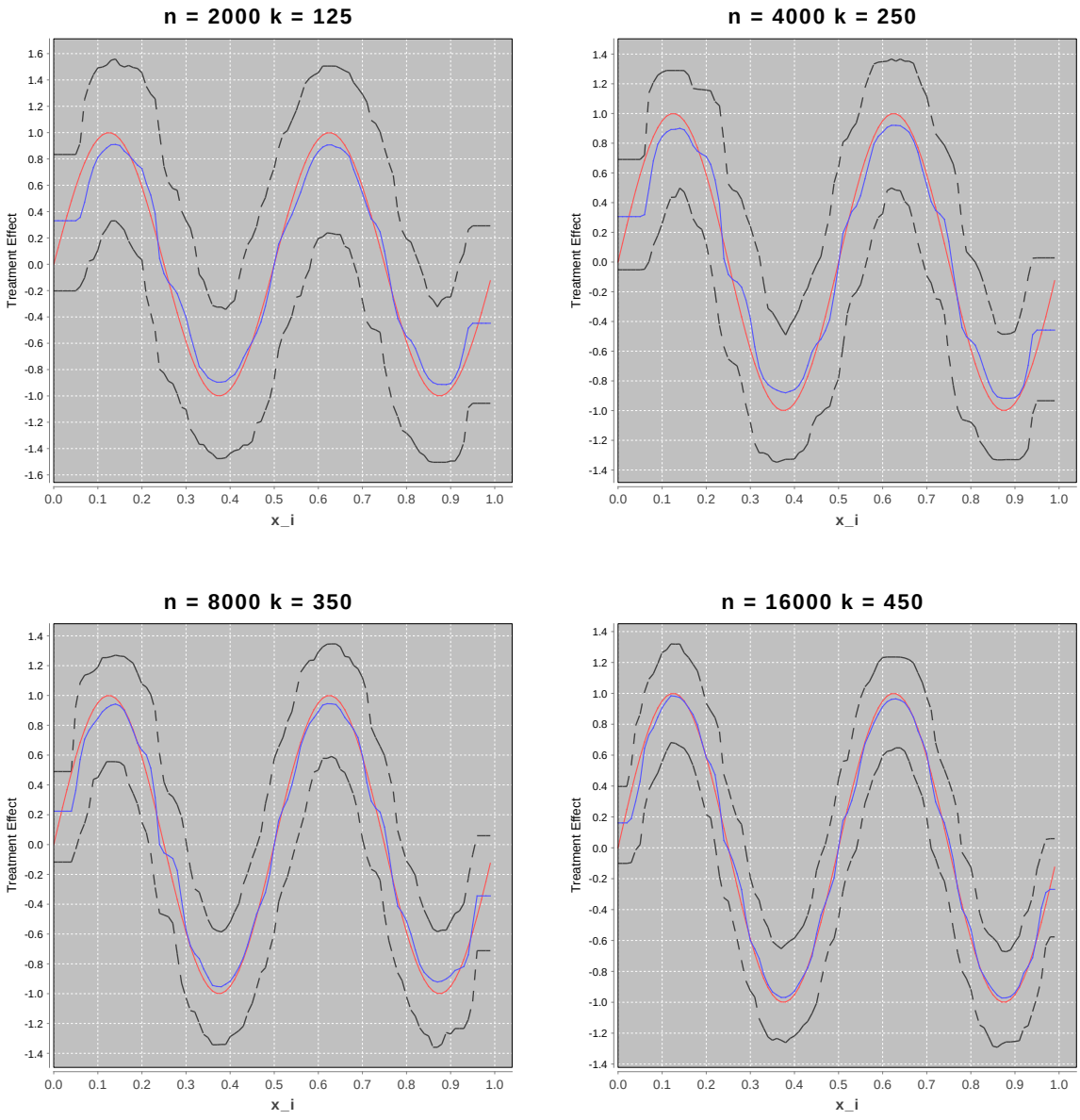
We now allow the error term to be drawn from a standard normal. The right panel of the table shows the results. The trees are simpler in this case, as the estimator has to balance variance against bias. The MSPE is substantially larger, although it rapidly shrinks at higher sample sizes. Figure 2 shows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

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



Notes: Each figure plots the mean estimated (blue) and true treatment effect (red) functions, $\tau(x_i)$, 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 were computed using 500 Monte Carlo experiments.

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 were computed using 500 Monte Carlo experiments.

4.3.3 Normally-Distributed Data

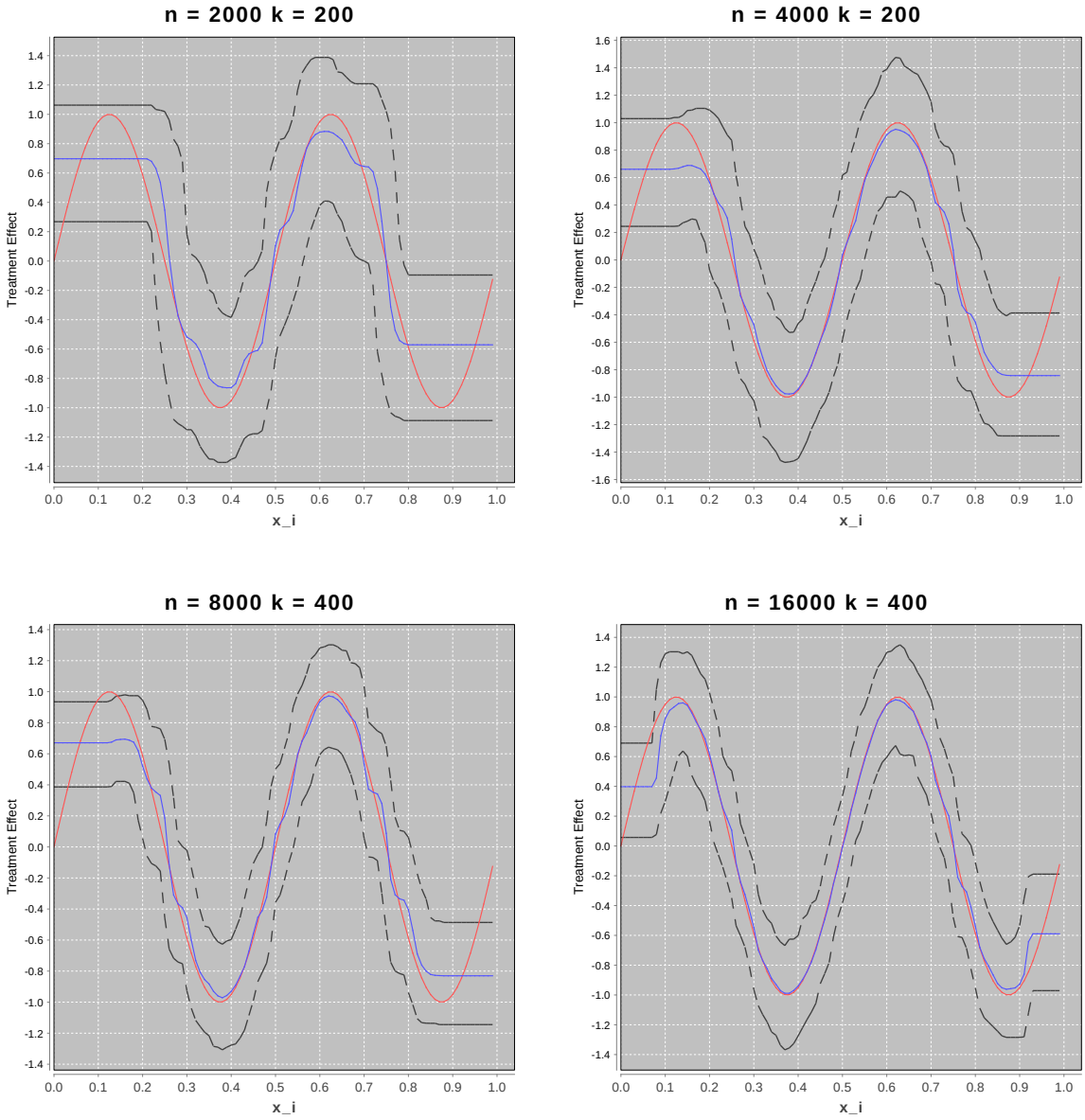
In this section, we show that the method works well even when data is not distributed uniformly across the domain of interest. We draw x from a normal distribution with mean one-half and standard deviation equal to 0.25, and truncate at zero and one. Figure 3 plots the estimated functions and the 95 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 hone 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.

5 Empirical Application

This section presents an illustrative example of our estimator in an empirical regression discontinuity context with treatment heterogeneity. We examine the impacts of India's large scale rural road construction program, under which over 115,000 new roads were built from 2000 to 2015, on the emergence of new bus routes to villages. The program's implementation rules dictated that villages were to be prioritized for new roads based on population thresholds, such that villages with populations just above 500 or 1000 were about 20% more likely to be treated than villages with populations just below those thresholds. Asher and Novosad (2019) exploit these discontinuities in a fuzzy regression discontinuity framework to evaluate the impacts of new roads. Because we are focused on the application of the method, we forgo the usual specification checks and assume that the assumptions required for causal identification are upheld; see Asher and Novosad (2019) for more details.

We focus on bus routes both because they are important economically, and because they showcase the challenge of selecting a good specification for examining treatment heterogeneity. Access to bus routes is a critical determinant of individuals' ability to take advantage of new rural roads, as few individuals in remote Indian villages own vehicles. Recent research has suggested that rural demand may not be sufficient to support the provision of bus services, which may mitigate the value of new roads Raballand, Thornton, Yang, Goldberg, Keleher, and Müller (2011). Bryan, Chowdhury, and Mobarak (2014) find very large returns to subsidizing bus travel to nearby cities, so much that their intervention is being scaled up into a major anti-poverty program in Bangladesh. Given the high cost of road construction,

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 were computed using 500 Monte Carlo experiments.

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. The typical bus in the Indian context is a privately operated vehicle that tightly fits fifteen to twenty people.

We choose this example because geographic variables are natural candidates as predictors for heterogeneity in the appearance of new bus routes, and such variables are easy to represent visually. The representation and visualization of heterogeneous treatment effects in random forest settings is otherwise challenging, because each unit has a distinct treatment effect; we avoid these challenges by choosing variables that can be easily represented in a heat map. A policy-focused analysis could saturate the model with an arbitrary number of location characteristics, even with more characteristics than observations; we leave this to future research.

A priori, one could predict that distance to towns plays a key role in whether a village with a new road also gets serviced by a bus route. However, the multidimensional nature of the town and highway network forces the econometrician to make many arbitrary decisions in estimating treatment heterogeneity. Are distance effects linear, or do they take a specific non-linear form? Is heterogeneity in distance to small towns equivalent to heterogeneity in distance to large towns? What town sizes should be used? The most common approach is to collapse the multidimensional town distance matrix (i.e. town size x distance to town) into a single scalar market access variable with an elasticitiy parameter (e.g. Donaldson and Hornbeck (2016)) but this is not guaranteed to be the empirically correct functional form. Our estimator allows for the discovery of a complex structure of treatment heterogeneity from a finite sample, while maintaining correct standard errors.

We combine data from the 2001 and 2011 population censuses on bus routes and other public goods in the universe of Indian villages, with administrative records from the national rural road construction program, which provides initial access conditions and road completion dates. Baseline covariates are measured in 2001 before any roads under the program were built. Data and sample construction are described in detail in Asher and Novosad (2019).

We implement a fuzzy regression discontinuity estimator that identifies the impact of new roads by examining changes in outcomes across the population treatment thresholds. We use a local linear estimator within the optimal population bandwidth, following the recommendations of Imbens and Lemieux (2008) and Imbens and Wooldridge (2009).

The basic two stage IV estimator takes the form:

$$Y_{v,j} = \gamma_0 + \gamma_1 newroad_{v,j} + \gamma_2(pop_{v,j} - T_j) + \gamma_3(pop_{v,j} - T_j) * 1\{pop_{v,j} \geq T_j\} + \upsilon_{v,j}, \quad (5.21)$$

where $Y_{v,j}$ is the outcome of interest in village v in state j, T_j is the population eligibility threshold in state j, $pop_{v,j}$ is village population measured at baseline and $\upsilon_{v,j}$ is an error term. $newroad_{v,j}$ is instrumented by an indicator variable that takes the value one for villages above the population threshold. γ_1 captures the causal effect of being treated with a new road, for a village with population at the treatment threshold T. The sample consists of the set of states that followed program implementation rules regarding population thresholds.[17]

5.1 Results

As a preliminary step, we present standard regression discontinuity estimates on the impact of new roads on bus routes. Column 1 of Table 7 reports the first stage estimate, where the outcome variable is an indicator that takes the value one if a village received a new road by 2010. Villages just above the treatment threshold are 20% more likely to receive a new road. Column 2 shows the IV estimates; the causal effect of a newly-built road to a rural village is to increase bus availability by 17 percentage points, a statistically significant estimate at the 5 percent level. Figure 4 shows a graphical representation of the first stage effect of crossing the population threshold on the probability that a village gets a new road (left panel) and the reduced form form effect of the same threshold on the presence of a bus route by the end of the sample (right panel). These tables and graphs describe an average treatment effect across the entire sample, but there may be heterogeneity in the correctly specified model, or some specification error. To assess whether we can improve on this estimate by accounting for unobserved heterogeneity, we next run the moment forest estimator on the same data.

We consider heterogeneity on several dimensions; the Z vector includes state fixed effects; distances to the nearest cities with 10,000 people, 100,000 people, and 500,000 people, respectively; and an indicator for existence of a bus route in 2001, prior to the road-building program.[18]

We follow the procedure described in Section 2. Within each of 50 bootstrap samples, we

[17] Different states used different implementation thresholds, as described in Asher and Novosad (2019). The sample includes villages from Chhattisgarh, Madhya Pradesh, Orissa and Rajasthan.

[18] Bus routes can exist at baseline in villages without paved roads, because some of these villages could be reached on dirt roads, though not necessarily in all seasons.

Table 7: Regression Discontinuity Estimates

	(1)	(2)
Road Priority	0.196***	
	(0.019)	
New Road		0.172**
		(0.086)
Population * 1(Pop < Cutoff)	0.019	-0.022
	(0.042)	(0.040)
Population * 1(Pop ≥ Cutoff)	0.077*	0.001
	(0.042)	(0.043)
N	9996	9996

Standard Errors in Parentheses

Figure 4: Graphical Regression Discontinuity

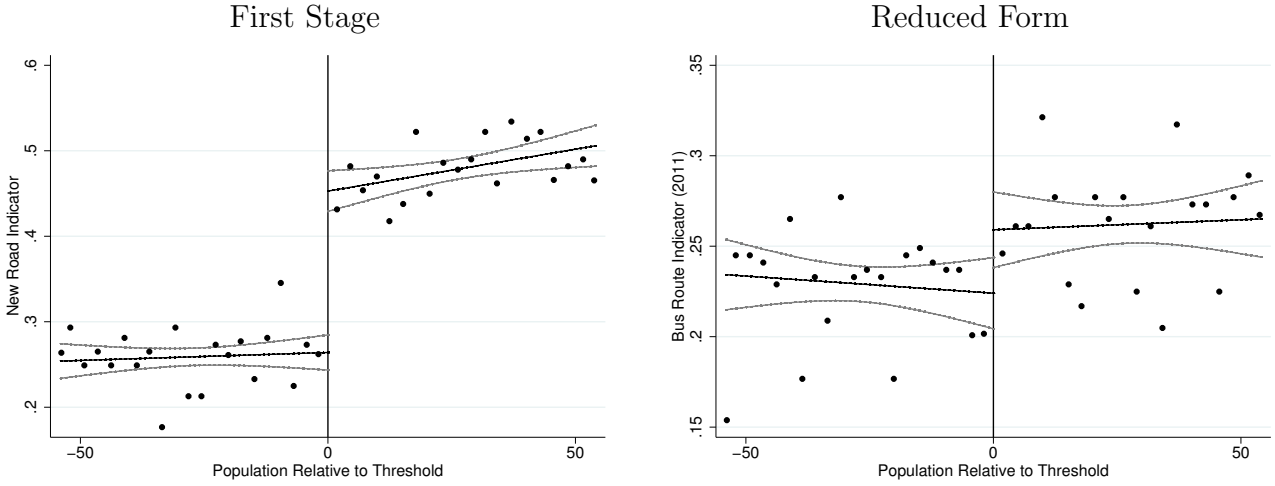
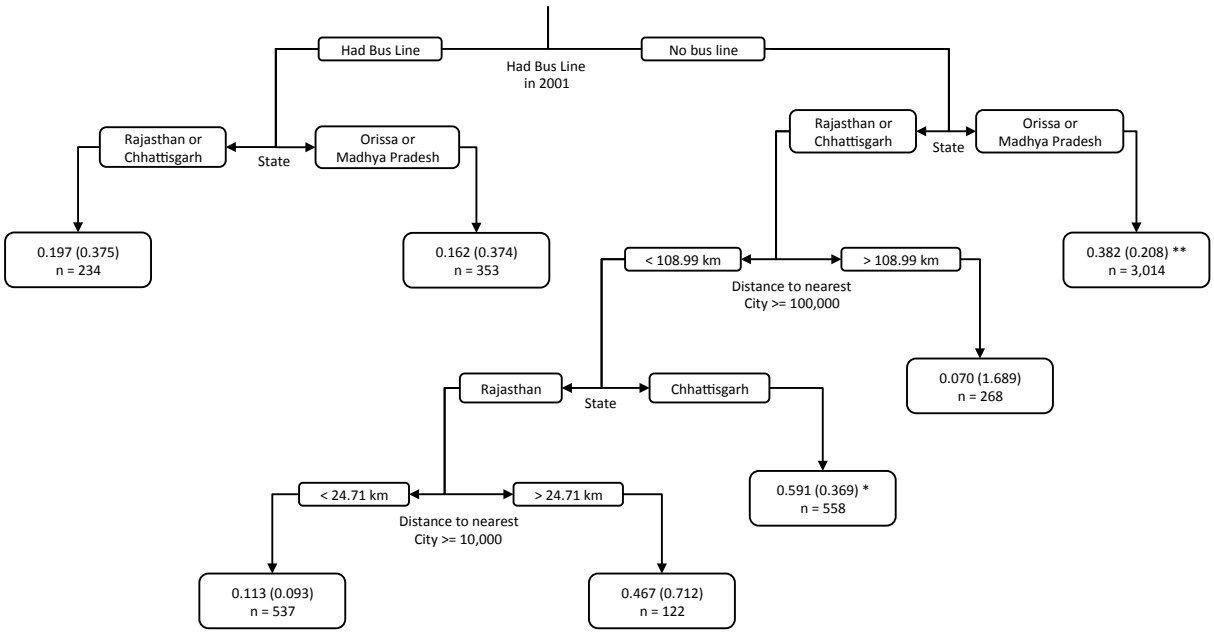


Figure 5: Sample Tree (Structure Subset)



grow 50 moment trees. To grow each tree, we sample with replacement (from the underlying bootstrap sample), and then partition the data into two subsets. We estimate the structure of the tree on the first subset, and estimate the RDD effect within each leaf of that tree in the complementary subset of the data.[19]

To illustrate what is happening in our estimation, it is useful to consider the output of a single tree. The structure selected for a single tree (generated on the first subset of the data) may look like the tree depicted in Figure 5.

Each binary split in the graph partitions the remaining data into two subsets. The cell beneath each leaf shows the treatment effect, standard error and number of observations in the leaf. In this tree, there are statistically significant effects in two of the leaves. The first is for the states Orissa and Madhya Pradesh (MP), in villages that previously did not have bus routes; the estimated treatment effect is a 38.2 percentage point increase in the probability of

[19] As in the Monte Carlo, we set the stopping criteria for growing the tree (k, α, and $\overline{MSE}$) using cross-validation. We used a holdout sample of 1,000 observations to calculate prediction error. Because the model selection step has much faster convergence rates than the estimation step, we used unequal sample sizes; we grew trees using 25% of the data and estimated treatment effects on the remaining 75% of observations. In principle these sample sizes could also be optimized with cross-validation.

receiving a bus route after building a road. The second effect is a more complex splitting of the data: for the state Chhattisgarh, 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 percentage points with a standard error of 39.6 percentage points, indicating this effect is marginally significant at the 10 percent level. For comparison, a baseline RDD on the estimation sample for this tree finds a treatment effect of 28.8 percentage points with a standard error of 11.6 percentage points.[20] The presence of prior bus routes is associated with low and statistically insignificant treatment effects, following the intuition given that it is not possible for the outcome variable to grow in this subsample, though it could decline. Note that the tree did not split on distance to the largest size of town, indicating that in this subsample it was not a key predictor of treatment effect sizes.

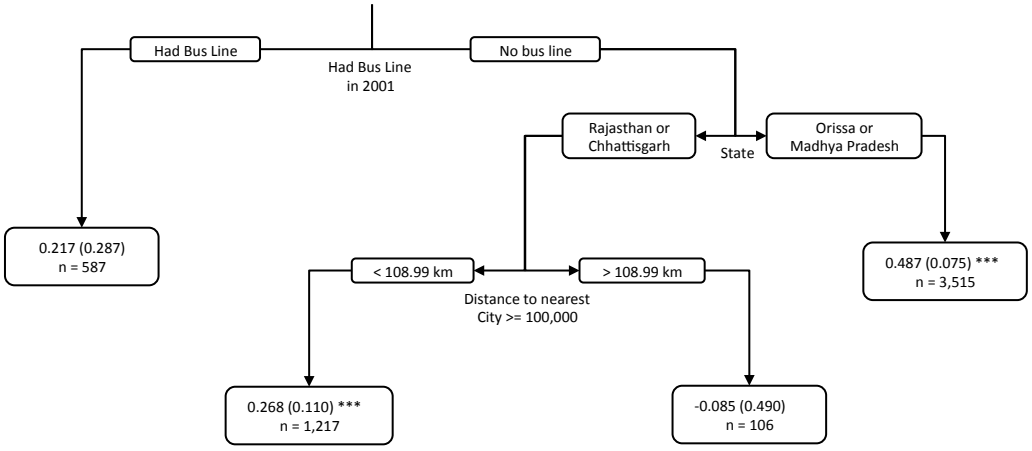
This tree from the first sample subset is only used for its structure; the standard errors are not correct because they do not take into account the process of selecting where to split the data. The second tree, estimated on the remaining subset of the sample, generates unbiased treatment effects. The second tree is estimated using the structure defined by the first tree above, but we "prune" leafs where first stage regression discontinuity estimates do not produce valid estimates; this could be because the first stage F statistic is too low, or because there are not enough observations in a given leaf of the second sample to run the estimation at all. We combine these pruned leafs at the next highest level of aggregation. The second tree is depicted in Figure 6.

The second tree finds an even stronger effect in Orissa and MP in the absence of prior bus routes. The two leaves with prior bus routes are now combined into one; there is still no significant effect. All leafs below the Rajasthan/Chhattisgarh split are also combined, resulting in a more precise joint estimate for villages within 108 km of cities with more than 100,000 people in these two states.

We highlight two outcomes of this process. First, the average treatment effect in this subsample of 0.288 is a composite estimate mixing together the weak and heterogeneous effects effects in Rajasthan and Chhattisgarh with the strong and precise effects in MP and Orissa, along with the zero effects in places already on bus routes. Second, even if the researcher had the intuition that distance to cities would be more important in the states with lower population density (Chhattisgarh and Rajasthan), this specific model is not something that

[20] This is different from the grouped RDD effect on all the data because this is a bootstrap sample and we are estimating the RDD only on the subset of the data not used to estimate the structure of the tree.

Figure 6: Sample Tree (Pruned Estimation Subset)

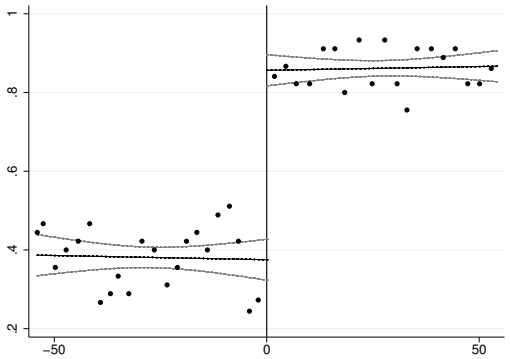


could plausibly have been pre-specified: the lowest leafs of the tree are based on splits of a continuous variable interacted with two other discrete variables. Even if such a specification was stumbled upon, the standard errors would not correctly account for the process of finding this specification. Note that this is a single tree out of 2,500 that are used to produce the final estimates, and different splits emerge in other subsamples.

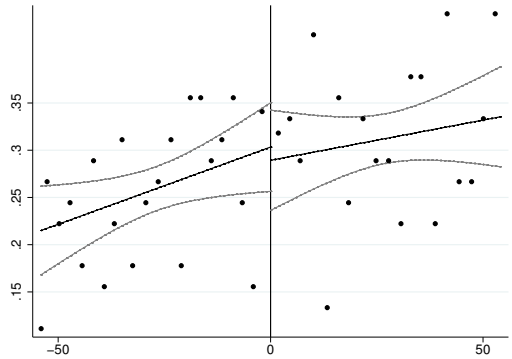
We estimate the full bootstrapped moment forest model under two sets of possible splitting variables. We first restrict each tree to only split on the state indicators. Under this restriction we obtain four different treatment effects: 0.292 (s.e. 0.157) in Orissa, 0.232 (s.e. 0.126) in Chhattisgarh, 0.301 (s.e. 0.147) in MP and -0.069 (s.e. 0.086) in Rajasthan. The first three are statistically different from zero at the 5% level. These results contrast with the baseline homogeneous estimate in two important ways. First, the moment tree estimator finds statistically significant effects for only 7389 of the 9996 sample villages. Second, it estimates three different effects within those 7389 villages. The baseline estimate combines both kinds of heterogeneity, resulting in a lower point estimate with a larger standard error. Figure 7 shows reduced form and first stage binscatter estimates separately for the set of states with significant treatment effects and the complementary set without. The figures make clear that the sample was split based on treatment effect size, not based on the size of the first stage; the insignificant sample has an even larger first stage than the significant sample.

Figure 7: Graphical Regression Discontinuity

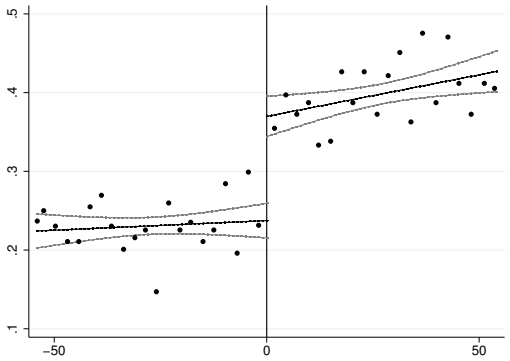
First Stage
Insignificant States



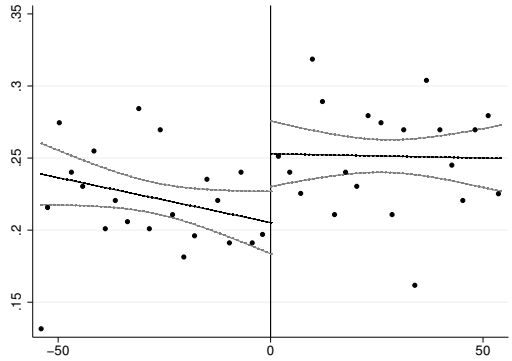
Reduced Form
Insignificant States



First Stage
Significant States



Reduced Form
Significant States



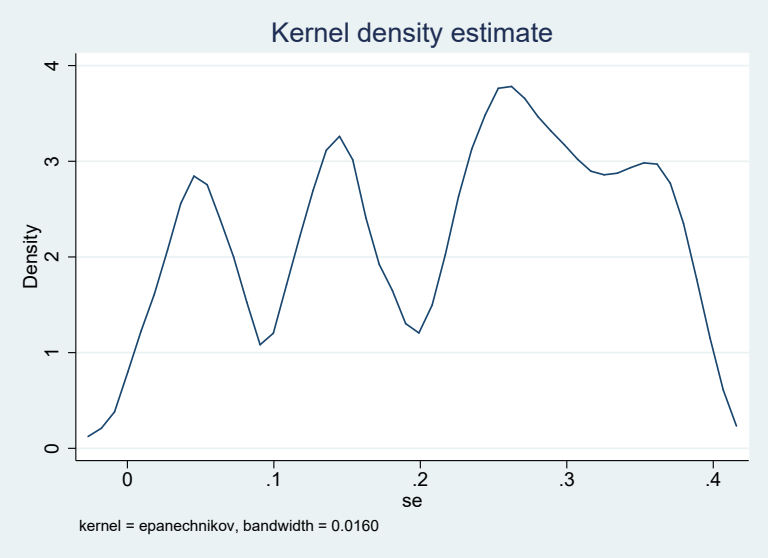
We next estimate the full moment tree model, allowing for continuous variables to enter the tree, specifically the distances to towns of different sizes. The estimator finds 3,525 statistically significant unique treatment effects across 9,991 villages. The treatment effect for each village is a weighted mean of treatment effects in every leaf in which that village appeared across 2,500 estimated moment trees. A density plot of the range of estimates is provided in Figure 8; the vast majority of estimated effects range from 0 to 0.4. This density plot understates the incredible richness of the estimator, however, as can be seen by plotting the distribution of treatment effects across space.

Figure 9a shows a heat map of treatment effects for all sample villages, both treated and untreated. Treatment effects are divided into deciles, where the black areas show the smallest treatment effects, and the bright red areas show the largest. State effects are plainly visible; as above, treatment effects are smallest in Rajasthan and largest in MP. Urban proximity stands out as well; the major cities of Bhubaneswar, Gwalior, Indore, Bhopal and Jabalpur stand out as places where nearby connected villages are most likely to benefit from new bus routes. Proximity to smaller cities does not appear to drive substantial variation in treatment effects, and Rajasthan and Chhattigarh (where treatment effects are respectively insignificant and marginally significant) do not show strong heterogeneity in urban proximity. Figure 9b presents another heatmap of treatment effects, this time for the subset of villages with statistically significant treatment effects (all of them positive). The figure uses a different gradient scale to highlight the substantial heterogeneity even among villages with large and significant treatment effects. The proximity effect is clearly non-linear—treatment effects become smaller at extremely close proximity to towns, perhaps because even villages with very poor quality access roads were already connected to bus routes in these periurban areas.

We can draw two conclusions from this exercise. First, treatment heterogeneity is substantial; many places have double the grouped treatment estimate, and many places have zero estimates. Second, the partitioning of the data is complex and nonlinear even with only a handful of variables, highlighting the fact that it is highly unlikely anyone would ever be able to guess that this was the true model, even though the pattern of treatment effects is sensible and can be understood ex post.

Trying to achieve a similar result by saturating a conventional regression specification would not be possible. There is no obvious way to cut the continuous variables; without some kind of grouping of similar estimates, one would quickly run out of data since there are literally an infinite number of potential cuts. A fully nonparametric approach would be fully general,

Figure 8: Density Plot of Treatment Effects



but converges 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.

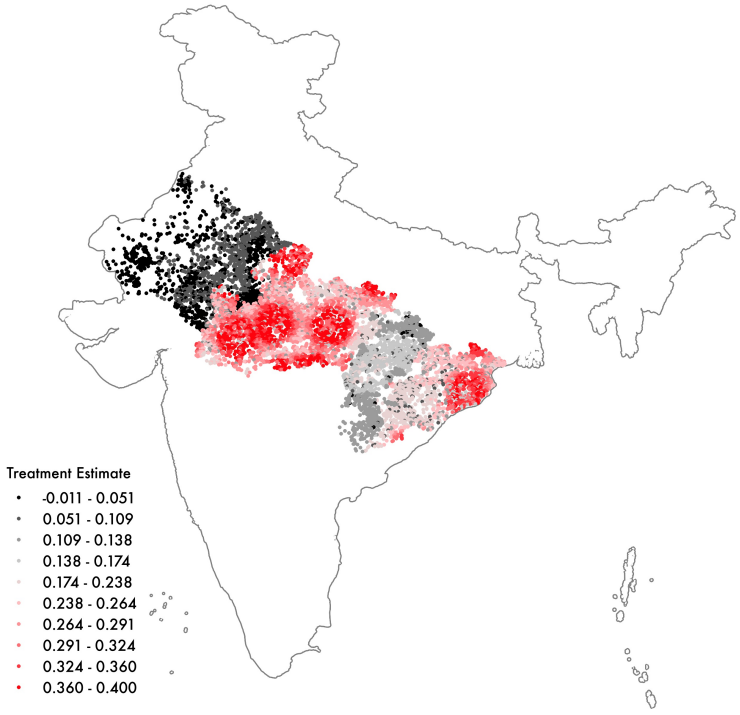
6 Conclusion

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 estimate parameters of those models within each assignment. We derive a set of econometric results showing 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.

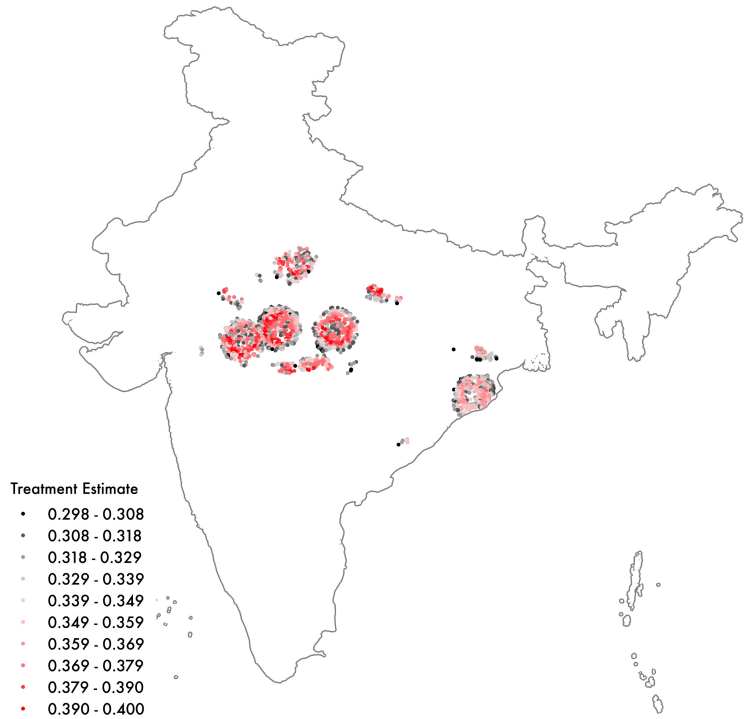
We apply our estimator to a road building project in India. Using a bootstrapped moment forest, we estimate a model that produces a range of statistically significant treatment effects spread across 3,603 villages. The results highlight the heterogeneity in treatment effects found using a regression discontinuity framework across these villages, including the impor-

Figure 9: Heatmap of Treatment Effects

(a) All Villages



(b) Statistically Significant Villages



tance of proximity to large urban spaces and the variation across Indian states. In future work, we plan to expand on these preliminary results and 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 (2019): "Rural Roads and Local Economic Development," *American Economic Review*, Forthcoming.

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*.

Athey, S., J. Tibshirani, S. Wager, et al. (2019): "Generalized random forests," *The Annals of Statistics*, 47(2), 1148–1178.

Bajari, P., C. L. Benkard, and J. Levin (2007): "Estimating dynamic models of imperfect competition," *Econometrica*, 75(5), 1331–1370.

Banerjee, A., S. Barnhardt, and E. Duflo (2018): "Can iron-fortified salt control anemia? Evidence from two experiments in rural Bihar," *Journal of Development Economics*, 133, 127–146.

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.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993): *Efficient and adaptive estimation for semiparametric models*, vol. 4. Johns Hopkins University Press Baltimore.

Breiman, L. (1996): "Bagging predictors," *Machine learning*, 24(2), 123–140.

——— (2001): "Random forests," *Machine learning*, 45(1), 5–32.

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.

Chamberlain, G. (1986): "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32(2), 189–218.

Chetty, R., N. Hendren, and L. F. Katz (2016): "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment," *The American Economic Review*, 106(4), 855–902.

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.

Donaldson, D., and R. Hornbeck (2016): "Railroads and American Economic Growth: A "Market Access" Approach," *Quarterly Journal of Economics*, 131(2).

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.

Greenstone, M., S. P. Ryan, M. Yankovich, and K. Greenberg (2017): "The Value of a Statistical Life: Evidence from Military Retention Incentives and Occupation-Specific Mortality Hazards," Discussion paper, University of Chicago Working Paper.

Gu, Q., T. Komarova, and D. Nekipelov (2017): "Differentially Private Empirical Risk Minimization with Non-Convex and Non-Lipschitz Loss functions," *University of Virginia working paper*.

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.

Leo, B., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984): "Classification and regression trees," *Wadsworth International Group*.

Loh, W.-Y. (2014): "Fifty years of classification and regression trees," *International Statistical Review*, 82(3), 329–348.

Messenger, R., and L. Mandell (1972): "A modal search technique for predictive nominal scale multivariate analysis," *Journal of the American statistical association*, 67(340), 768–772.

Morgan, J. N., and J. A. Sonquist (1963): "Problems in the analysis of survey data, and a proposal," *Journal of the American statistical association*, 58(302), 415–434.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, pp. 1349–1382.

Otsu, T., M. Pesendorfer, and Y. Takahashi (2016): "Pooling data across markets in dynamic Markov games," *Quantitative Economics*, 7(2), 523–559.

Powell, J. L. (1994): "Estimation of semiparametric models," *Handbook of econometrics*, 4, 2443–2521.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica: Journal of the Econometric Society*, pp. 1403–1430.

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).

Robinson, P. M. (1988): "Semiparametric econometrics: A survey," *Journal of Applied Econometrics*, 3(1), 35–51.

Ryan, S. P. (2012): "The costs of environmental regulation in a concentrated industry," *Econometrica*, 80(3), 1019–1061.

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*.

Walther, G., et al. (2010): "Optimal and fast detection of spatial clusters with scan statistics," *The Annals of Statistics*, 38(2), 1010–1033.