INFORMS Journal on Data Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Achieving Reliable Causal Inference with Data-Mined Variables: A Random Forest Approach to the Measurement Error Problem

Mochen Yang, Edward McFowland, III, Gordon Burtch, Gediminas Adomavicius

To cite this article:
Mochen Yang, Edward McFowland, III, Gordon Burtch, Gediminas Adomavicius (2022) Achieving Reliable Causal Inference with Data-Mined Variables: A Random Forest Approach to the Measurement Error Problem. INFORMS Journal on Data Science 1(2):138-155. https://doi.org/10.1287/ijds.2022.0019

https://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Vol. 1, No. 2, October–December 2022, pp. 138–155
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

Achieving Reliable Causal Inference with Data-Mined Variables: A Random Forest Approach to the Measurement Error Problem

Mochen Yang,[a,*] Edward McFowland, III,[b] Gordon Burtch,[c] Gediminas Adomavicius[a]

[a] Department of Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455; [b] Department of Technology and Operations Management, Harvard Business School, Boston, Massachusetts 02163; [c] Department of Information Systems, Questrom School of Business, Boston University, Boston, Massachusetts 02215
*Corresponding author

Contact: yang3653@umn.edu, https://orcid.org/0000-0001-5101-9041 (MY); mcfowland@hbs.edu, https://orcid.org/0000-0001-5249-7117 (EM); gburtch@bu.edu, https://orcid.org/0000-0001-9798-1113 (GB); gedas@umn.edu, https://orcid.org/0000-0001-5251-5098 (GA)

Received: February 14, 2022
Revised: May 18, 2022
Accepted: June 27, 2022
Published Online in Articles in Advance: September 21, 2022

https://doi.org/10.1287/ijds.2022.0019

Abstract. Combining machine learning with econometric analysis is becoming increasingly prevalent in both research and practice. A common empirical strategy uses predictive modeling techniques to "mine" variables of interest from available data and then includes those variables into an econometric framework to estimate causal effects. However, because the predictions from machine learning models are inevitably imperfect, econometric analyses based on the predicted variables likely suffer from bias due to measurement error. We propose a novel approach to mitigate these biases, leveraging the random forest technique. We propose using random forest not just for prediction but also for generating instrumental variables for bias correction. The random forest algorithm performs best when comprised of a set of trees that are individually accurate in their predictions, yet which also make "different" mistakes, that is, have weakly correlated prediction errors. A key observation is that these properties are closely related to the relevance and exclusion requirements of valid instrumental variables. We design a data-driven procedure to select tuples of individual trees from a random forest, in which one tree serves as the endogenous covariate and the others serve as its instruments. Simulation experiments demonstrate its efficacy in mitigating estimation biases and its superior performance over alternative methods.

History: David Martens served as the senior editor for this article.
Data Ethics & Reproducibility Note: The code capsule is available on Code Ocean at https://codeocean.com/capsule/7039927/tree/v1 and in the e-Companion to this article (available at https://doi.org/10.1287/ijds.2022.0019).

Keywords: machine learning • econometric analysis • instrumental variable • random forest • causal inference

1. Introduction

Advances in predictive machine learning have enabled researchers to extract useful information from various types of data, such as text and images, which would otherwise be difficult or costly to codify at scale. For example, recent academic work has highlighted that cutting-edge prediction techniques are now capable of inferring the socioeconomic properties of a localized population (e.g., income/racial distribution) from the models and makes of cars appearing in Google Street View images (Gebru et al. 2017) and detecting adverse drug events based on drug attributes (Ryu et al. 2018). These measurements, now available at scale and with little cost, can *enable* empirical investigations of important questions in economics, healthcare, and many other domains.

Indeed, many researchers have begun doing exactly that, first using predictive machine learning to construct or populate a variable of interest, for example, using text mining tools to predict text sentiment and then including that variable into econometric models as an independent covariate. This practice has become prevalent in multiple social science domains, including economics (Jelveh et al. 2015), political science (Fong and Tyler 2021), and management (Yang et al. 2018). We summarize a number of example studies that have used this approach in Online Appendix F.

However, recent studies have also noted that attempts to draw inferences based on this recipe are likely to suffer from endogeneity due to measurement error (Yang et al. 2018). This is because predictions from machine learning models are inevitably imperfect, and prediction errors carry over to subsequent econometric models as measurement error, leading to biased and inconsistent parameter estimates. Measurement error may lead to overestimation or underestimation of coefficients (Loken and Gelman 2017), and the degree of bias can be substantial, even when the machine learning model achieves reasonable predictive performance (Yang et al. 2018). The estimation biases stemming from measurement error in machine learning–generated covariates can thus undermine the

validity of subsequent causal inferences and decision making.

In this paper, we propose a novel approach to address this problem. Our approach is based on the notion of instrumental variable regression, a well-established method of resolving endogeneity in the econometrics literature, including endogeneity deriving from measurement error (Greene 2003). We make use of a notably unique feature of this problem setting of applying machine learning followed by regression. Specifically, we leverage the fact that predictive machine learning models are typically trained and evaluated using data for which true labels (assumed to be perfectly measured) are available and which are used to quantify prediction error and model performance. This perfectly measured set of data offers a *unique opportunity* to overcome difficulties commonly associated with evaluating the validity of instruments.

To find candidate instruments, we rely on *random forest* (Breiman 2001), an ensemble learning approach that aggregates a collection of individual decision trees (weak learners) to arrive at accurate predictions. Prior work has demonstrated that the performance of random forests depends jointly on (i) the degree to which trees comprising the forest yield correlated predictions and (ii) the degree to which the trees yield weakly correlated prediction errors (Breiman 2001, Bernard et al. 2010). We demonstrate that these notions are closely related to the relevance and exclusion criteria that underpin valid instrumental variables. Based on this, we look to explore the random forest ensemble to identify sets of individual trees such that one tree's predictions might serve as the endogenous covariate in the econometric model of interest, while other trees' predictions serve as its instruments, thereby alleviating estimation biases due to measurement error. Drawing on theories from the instrumental variable and random forest literature, we develop algorithms to implement this idea, empirically selecting the best set of individual trees to mitigate biases in coefficient estimates. We term our procedure *ForestIV*.[1]

We conduct multiple sets of comprehensive simulation experiments, considering the cases where a data-mined covariate is *continuous* versus *binary* and thus, respectively, suffers from either continuous measurement error or misclassification. In both cases, we show that ForestIV can effectively mitigate estimation biases. We also report sensitivity analyses of ForestIV and benchmark its performance against three alternative bias correction methods.

It should be noted that ForestIV provides a *general-purpose* method for correcting bias with machine learning–generated covariates, whether derived from structured or unstructured data (e.g., text or images). For scenarios involving structured data, random forest is widely applicable to various supervised machine learning problems, and, for a large number of real-world prediction problems, random forest is among the most accurate techniques (Fernández-Delgado et al. 2014). Meanwhile, even in scenarios involving unstructured data, where other techniques (e.g., deep neural networks) may be the state-of-the-art, random forest can be usefully combined with them. For example, random forest can be stacked with the intermediate representations learned by a neural network; that is, the output of an intermediate layer of the network, which encodes informative, high-level features learned from the unstructured data, can serve as the input features to the random forest algorithm. Notably, this practice is quite common in transfer learning, where a supervised machine learning model is built based on features produced by another technique (Goodfellow et al. 2016).

Our paper makes several notable contributions. First, we theoretically and empirically show that the proposed ForestIV approach effectively addresses the estimation biases in econometric models due to measurement error in machine-learning-generated covariates. Therefore, ForestIV improves the robustness of causal inferences and decisions derived from procedures combining machine learning with econometric analyses. Second, we design data-driven procedures that leverage the labeled data (used to build and evaluate the machine learning model) to empirically select instruments that are most suitable for bias correction purposes. Third, ForestIV represents a novel method to automatically obtain candidate instruments from the output of the random forest technique. This provides a viable solution to the often-challenging problem of identifying valid instruments.

2. Relevant Literature

Our work is informed by the econometrics literature on measurement error, instrumental variables, and generated regressors, as well as the machine learning literature on random forests. We provide a review of relevant literature in this section.

The problem of measurement error has been studied extensively in the econometrics literature. In regression models, measurement error in covariates (regressors) is a form of endogeneity (Greene 2003). This endogeneity is known to lead to biased coefficient estimates, not only for the mis-measured covariate, but also for coefficients associated with other (precisely measured) covariates appearing in the same regression (unless the precisely measured covariates are strictly independent of the measurement error). In contrast to the common (mis-held) belief that measurement error only leads to attenuation of coefficient on the mis-measured covariate (i.e., bias toward zero), the actual direction of bias is difficult to anticipate, particularly as the econometric specification or the structure of the measurement error grow more complicated (Gustafson 2003, Schennach

2016, Yang et al. 2018). In general, ignoring measurement error may lead to errors in sign, magnitude, and statistical significance of coefficient estimates.

2.1. Instrumental Variable Approach

Instrumental variables are a standard approach to addressing the measurement error problem; they can be used in a two-stage least-squares estimation to mitigate associated estimation biases (Carroll and Stefanski 1994, Buzas and Stefanski 1996, Greene 2003, Hu and Schennach 2008). A valid instrument in this case will be correlated with the mis-measured covariate but not its measurement error. However, researchers using the instrumental variable approach often face two significant challenges. First, valid instruments are not easy to locate. Typically, instruments cannot be identified absent knowledge of the underlying data-generation process, that is, the nature of the endogeneity at play.[2] Second, to justify the validity of a proposed instrument, researcher needs to provide convincing evidence that the instrument satisfies two criteria, namely *relevance* and *exclusion*. Although the former criterion can be evaluated by empirically examining strength of the association between the endogenous covariate and the instrument, the latter is often untestable and thus depends on the researcher offering a convincing qualitative, conceptual argument that the instrument has no association with the final outcome of interest, except via its influence on the endogenous variable.

Our algorithm offers a novel opportunity to achieve both requirements, through quantitative means, as it leverages the availability of a perfectly measured set of data, that is, the predictive model's labeled data, reducing the need for qualitative arguments (or restrictive assumptions).

2.2. Ensemble Learning and Random Forest

In the machine learning literature, ensemble learning represents an important paradigm in the formulation of predictive models. Instead of building one model to solve a prediction problem, ensemble learning aims to build multiple individual models, that is, an ensemble of models, and to combine their individual predictions to arrive at a more accurate and stable aggregate prediction (Aggarwal 2015). Some typical ensemble learning methods include bagging (Breiman 1996), boosting (Freund and Schapire 1996), and random forest (Breiman 2001, Denisko and Hoffman 2018), of which the latter is particularly relevant to our paper. A random forest is an ensemble of decision trees. Each tree is built on a random sample of the training data, and a random subset of features is considered for each split (node) in a tree (Breiman 2001). The forest's prediction for an observation results from aggregating predictions from individual trees, for example, majority voting for classification tasks or averaging for numeric prediction tasks.

Random forests have proven extremely useful in a variety of fields of study due to their commonly high predictive accuracy (Verikas et al. 2011, Denisko and Hoffman 2018). The predictive performance of a random forest is positively associated with the accuracy of each individual tree and negatively associated with intertree correlations in prediction errors (Breiman 2001, Bernard et al. 2010). Intuitively, the performance of a random forest increases as a joint function of the individual prediction accuracy of trees comprising the forest, and the degree to which constituent trees make" different" prediction errors. Based on the observation that these objectives are closely analogous to the relevance and exclusion restrictions that underpin valid instrumental variables, there is some face validity to the idea that these individual trees may serve as candidate instruments for one another to resolve endogeneity in later regressions arising from predictive (measurement) error.

2.3. Generated Regressors Literature

The measurement error problem we study here is closely related to a large body of econometrics literature on "generated regressors," where certain covariates in econometric estimations are not directly observed; rather, they are first estimated. In fact, two-stage least-square (2SLS) estimation with instrumental variables is one such generated regressor model, where predicted values of the endogenous covariate used in the second-stage regression are generated from the first-stage regression. Researchers have examined the theoretical properties of econometric models incorporating generated regressors in cases where the generating function or the final estimation are parametric (Newey 1984, Murphy and Topel 1985), semiparametric (Blundell and Powell 2004, Mammen et al. 2016), or nonparametric (Sperlich 2009, Mammen et al. 2012). For in-depth reviews of this extensive literature, we refer the reader to Pagan (1984) and Oxley and McAleer (1993). In some of this work, researchers have noted that the generating function may yield biased estimates of the regressor in question, which in turn will yield bias and inconsistency in a second-stage regression, discussing (typically theoretical) approaches to resolving the problem.

Our context and the problem we are seeking to resolve bear obvious similarities to the generated regressor problem. However, the measurement error problem that we address here nonetheless has some unique characteristics that differentiate it. In particular, in our setting, the measurement error stems from predictions of a machine learning model, which is built using a set of labeled data on which the covariate of interest is (assumedly) perfectly observed. In other words, the covariate to be generated is only *partially* unobserved. This is different from the typical setup of a generated regressors model in the literature. This partial observation via a

labeled data set enables objective quantification of measurement error and potentially more effective bias correction. As will be discussed later, our proposed method makes use of this labeled data set to achieve bias correction.

Based on the previous statements, some recent work in the generated regressors literature has proposed methods to correct for bias in generated regressors in the presence of distributional information about that bias (Meng et al. 2016). Accordingly, as part of our benchmarking analyses, we seek to compare the relative performance of our method with that approach (Meng et al. 2016).

3. ForestIV with Continuous Endogenous Covariate

In this section, we provide a description of the measurement error problem resulting from machine learning predictions of a *continuous* covariate. We then introduce the theoretical justifications and implementation details of our proposed ForestIV solution.

3.1. Continuous Measurement Error Problem Formulation

We begin by setting up the measurement error problem for a continuous covariate. For expositional simplicity, we use a simple linear regression as a representation of the econometric model to be estimated, but the underlying theoretical arguments can be generalized to other econometric specifications, for example, generalized linear models.[3]

Consider a linear regression model,

$$Y = X\beta_X + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}} + \varepsilon, \tag{1}$$

where Y represents the dependent variable, $\{X, \boldsymbol{Z}\}$ represent the independent covariates ($\boldsymbol{Z}$ includes control variables and a constant term), ε is the *exogenous* random error term, and $\boldsymbol{\beta} = \{\beta_X, \boldsymbol{\beta}_{\boldsymbol{Z}}\}$ denotes the model coefficients to be estimated. Importantly, X is not directly observed in the data, and we instead rely on its surrogate, $\widehat{X}$, which is based on the *predictions* of a machine learning model, for example, random forest. For instance, if X represents poverty level in a neighborhood, then $\widehat{X}$ could be the predicted poverty level mined from Google Street View images. In contrast, covariates $\boldsymbol{Z}$ are directly observed in the data and are measured precisely (with no measurement error). Therefore, the actual estimation being conducted is a regression of Y on $\{\widehat{X}, \boldsymbol{Z}\}$. In this section, we assume X and $\widehat{X}$ to be continuous variables, whereas $\boldsymbol{Z}$ can contain both continuous and categorical variables. We discuss the case when X and $\widehat{X}$ are binary variables in later sections.

Because predictions of a machine learning model inevitably have some degree of error, $\widehat{X}$ is in general an imperfect surrogate for X and contains continuous measurement error. The existence of measurement error in an independent covariate is known to result in biased regression estimates. As an illustrative example, consider a particular case of additive measurement error, where $\widehat{X} = a + bX + e$, where a, b represent the additive and multiplicative systematic errors and e represents the random error. The estimated regression equation is

$$\begin{aligned} Y &= \widehat{X}\beta_X + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}} + \left(\varepsilon - \beta_X \frac{(b-1)\widehat{X} + a + e}{b}\right) \\ &:= \widehat{X}\beta_X + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}} + \omega. \end{aligned} \tag{2}$$

Because $Cov(\widehat{X}, \omega) = -\beta_X\left(\frac{b-1}{b}\sigma_{\widehat{X}}^2 + \frac{1}{b}\sigma_{\widehat{X}e}\right) \neq 0$, the regression suffers from endogeneity due to measurement error in $\widehat{X}$, resulting in biased coefficient estimates (Greene 2003). (See Online Appendix B for a more formal setup of this problem.)

3.2. Building Random Forest

Consider the task of building a random forest (or any predictive machine learning model) to predict X. A typical approach is to collect some amount of *labeled data*, where the outcome to be predicted is actually observed. More concretely, denote the entire data set that a researcher has access to as D. Suppose the researcher takes a random subsample from D, denoted as D_{label}, and obtains the (precisely measured) ground truth for that subsample, for example, via manual labeling. Then, D_{label} would be randomly partitioned into D_{train} and D_{test}, where D_{train} would be used to build the random forest model and D_{test} to evaluate the resulting model's performance. The random forest model would then be deployed on the remaining unlabeled data set, $D_{unlabel} = D \backslash D_{label}$, to generate predictions $\widehat{X}$.

As is common in a growing number of policy-relevant contexts, the data set of interest often includes a large amount of unlabeled data. Due to the cost of obtaining ground truth labels, the size of D_{label} is typically much smaller than the size of $D_{unlabel}$. As a result, researchers may not be able to estimate their econometric models of interest on D_{label} with a satisfactory degree of statistical power given the likely large variance. That is, estimating the econometric models using information in $D_{unlabel}$ presumably has the potential to deliver substantial improvements in the precision of estimates (due to its larger sample size).

3.3. Generating Candidate Instruments from Random Forest

In this section, we consider the generation of instruments from random forest, which is at the core of the

ForestIV approach. We first establish the setting and list assumptions, before providing formal theoretical results. We consider a random forest model with M individual trees, indexed by $\{1,\ldots,M\}$, which is built based on n training samples and p features. On a new data point represented by feature vector $\mathbf{f} = \{f_1,\ldots,f_p\}$, denote the prediction of individual tree $i \in \{1,\ldots,M\}$ as $\widehat{X}^{(i)}$, and the prediction of the forest as $\widehat{X}$, where $\widehat{X} = \frac{1}{M}\sum_{i=1}^{M}\widehat{X}^{(i)}$. Given the ground truth, X, the prediction error is correspondingly defined as $e^{(i)} = X - \widehat{X}^{(i)}$. To establish our theoretical results, we adopt two assumptions from Scornet et al. (2015) that were used to show the consistency of random forest predictions.

Assumption 1 (Ground Truth Function, Scornet et al. 2015). *The ground truth can be expressed as* $X = \sum_{k=1}^{p} m_k(f_k) + \zeta$, *where features* $\{f_1,\ldots,f_p\}$ *are uniformly distributed over* $[0,1]^p$, ζ *represents independent, centered Gaussian noise with finite variance, and each component* $m_k(.)$ *is continuous. This assumption states that the ground truth is the sum of univariate functions of input features. Although random forest is a nonparametric model, analysis of its properties is often facilitated within the framework of additive models (Scornet et al. 2015).*

Assumption 2 (Tree Growth, Scornet et al. 2015). *Denote* t_n *as the number of leaves in each tree, and* a_n *as the number of training data points used to build each tree. Let* $a_n \to \infty$, $t_n \to \infty$, *and* $t_n(\log a_n)^9/a_n \to 0$. *This assumption, as a regularity condition, controls the rate at which the trees in the random forest grow. More specifically, the tree grows slower than the amount of training data.*

Theorem 1. *Under Assumptions 1 and 2, for any two trees in the random forest, i and j* $(i \neq j)$,

$$\lim_{n\to\infty} Cov(\widehat{X}^{(j)}, e^{(i)}) = 0.$$

This result, the proof of which can be found in Online Appendix A, implies that the covariance between one tree's prediction and another tree's prediction error goes to zero as the amount of training data goes to infinity, which establishes an asymptotic guarantee for instrument validity, and therefore, the theoretical foundation for considering trees within a random forest as instruments for one another. This theorem can also be understood intuitively. As the sample size and leaf nodes of a decision tree go to infinity (according to the rates specified in Assumption 2), predictions of the tree converge to the "predictable" part of the ground truth (i.e., the sum of additive functions over all features, as specified in Assumption 1), whereas the prediction errors of the tree converge to the "unpredictable" part of the ground truth (i.e., ζ). As a result, the prediction errors are asymptotically uncorrelated with the predictions, which provides support for the validity of instrumental variables.

More generally, Theorem 1 tells us that, asymptotically, and under mild assumptions, in a random forest of M trees, we can use predictions from any individual tree as the endogenous covariate in the regression model and use predictions from the *other* $M-1$ individual trees in the random forest as valid instruments. In finite samples, the previous asymptotic results with respect to random forest are in fact also reflected in empirical evidence. For example, Bernard et al. (2010) show that the performance of random forest statistically improves with increases in the accuracy of individual trees and decreases in the *correlation* between their prediction errors. In other words, a well-performing random forest should consist of individual trees that are relatively accurate (high *strength*) and have only weakly correlated errors (low *correlation*) (Breiman 2001). We make the key observation that high strength and low correlation are closely related to the requirements of a valid instrumental variable.

This highlights an interesting and perhaps counterintuitive characteristic of ForestIV. Because the predictive performance of an individual tree is typically worse than that of the entire random forest, we likely induce larger estimation biases by drawing on predictions from an individual tree as the endogenous covariate (rather than the aggregate prediction from the overall forest). However, this initial sacrifice is accompanied by the opportunity to leverage predictions from other trees as instruments, to address the estimation bias. In other words, for the purposes of *causal inference*, a more nuanced use of the entire random forest ensemble (rather than a traditional use of its aggregate predictions) allows the mitigation of the measurement error that is inevitably present in the model's predictions.

Finally, although this result provides a valuable theoretical foundation for the random forest algorithm as a generator of valid instruments, we note that in a *finite* sample, the practical objective of ForestIV is to improve econometric estimations rather than to completely eliminate estimation biases. As will be shown later, under a common scenario where the labeled data set is relatively small compared with the unlabeled data set, ForestIV can produce less biased estimates (than a naive regression without accounting for measurement error) and more precise estimates (than an unbiased regression directly using the labeled data).

3.4. Selecting Instruments for Correction

In practice, when using the predictions from a focal individual tree in a random forest as the endogenous covariate, it is necessary to select only a *subset* of the other trees' predictions for use as instruments, for several reasons. First, when there is only a finite set of training data, it is unlikely that all other trees have reached

their (asymptotic) state of valid instrumentation for the focal tree. Second, due to randomness in constructing the random forest with finite data, even valid instruments can be invalid by chance, that is, empirically $Cov(e^{(i)}, \widehat{X}^{(j)})$ could be large. Third, using an excessive number of instruments can lead to overfitting of the endogenous variables, including the endogenous components a researcher seeks to eliminate, and create estimation challenges (Roodman 2009). Finally, again due to randomness in the construction of a random forest, some instruments may be only weakly correlated with the endogenous covariate, despite meeting the exclusion requirement. Including these *weak instruments* in an instrumental variable regression can be counterproductive, yielding biased and inconsistent estimations (Hausman 2001).

In our setting, because we have access to a set of labeled data, we can assess instrument validity empirically, to a degree. Thus, after obtaining the predictions from each individual tree in the random forest, we focus on identifying a "desirable" subset of trees: one to be used as the endogenous covariate, and the remaining to serve as valid and strong instruments that can mitigate estimation biases. We decompose this task into three distinct steps, as follows:

- **Step 1: Removal of Invalid Instruments**: Given $i \in \{1, \ldots, M\}$, use $\widehat{X}^{(i)}$ as the endogenous covariate, then select a subset of other trees, $\boldsymbol{V_i} \subseteq \{\widehat{X}^{(1)}, \ldots, \widehat{X}^{(M)}\} \backslash \widehat{X}^{(i)}$, which omit *invalid* instruments for $\widehat{X}^{(i)}$. This step is conducted using D_{test}.
- **Step 2: Selection of Strong Instruments**: Given $i \in \{1, \ldots, M\}$, use $\widehat{X}^{(i)}$ as the endogenous covariate, then select a subset of other trees, $\boldsymbol{S_i} \subseteq \boldsymbol{V_i}$, that consists of *strong* instruments for $\widehat{X}^{(i)}$. This step is conducted using $D_{test} \cup D_{unlabel}$. Steps 1 and 2 are iterative (discussed in detail later).
- **Step 3: Estimation**: Based on selected instruments $\boldsymbol{S_i}$ for covariate $\widehat{X}^{(i)}$, obtain the 2SLS regression estimates. We carry out additional checks to gauge the validity of 2SLS regression estimates and retain the 2SLS estimates that meet the specific checks to produce the final corrected coefficient estimates. This step is conducted using $D_{label} \cup D_{unlabel}$.

The additional "validity checks" in Step 3 are necessary because, even though Steps 1 and 2 are designed to remove invalid instruments and select strong ones, there is limited theoretical guarantee that all selected instruments are always valid and strong in *finite samples*. The validity checks in the third step thus attempt to identify for which subset of trees the asymptotic properties appear to be present, reducing the likelihood that our approach produces erroneous results when predictions from all individual trees in the random forest are not suitable instruments. Moreover, with a finite sample, it is unlikely that instrument validity, particularly the exclusion restriction, will be satisfied *exactly*. Therefore, our procedure is consistent with the practice of using "plausibly exogenous" instruments for estimation in finite samples (Conley et al. 2012).

3.4.1. Step 1: Removal of Invalid Instruments.

To rule out invalid instruments for a given endogenous covariate, we rely on information from the labeled data. Recall that the random forest model is built on D_{train}, and its performance is then evaluated on D_{test}. As a result, for D_{test}, we observe the ground truth, the model-predicted values, and thus the prediction errors (the difference between ground truth and prediction). Using this information on D_{test}, we can gauge the validity of using individual trees' predictions as instruments and empirically rule out the invalid instruments that are strongly correlated with the measurement errors.

Grounded in prior work on instrument selection (Belloni et al. 2012), we adapt a lasso-based heuristic procedure to identify and discard instruments violating the exclusion requirement. Without loss of generality, suppose predictions from the first individual tree on D_{test}, $\widehat{X}^{(1)}$, serve as the endogenous covariate, and the corresponding prediction error is denoted as $e^{(1)}$. We estimate a lasso regression of $e^{(1)}$ on the predictions of other individual trees on D_{test}, that is, $e^{(1)} \sim \{\widehat{X}^{(2)}, \ldots, \widehat{X}^{(M)}\}$. The set of regressors for which the lasso yields nonzero coefficients are then dropped, because their linear combination is determined to be a strong predictor of $e^{(1)}$, which implies that those regressors violate the exclusion restriction with respect to $\widehat{X}^{(1)}$. Conversely, the set of regressors with zero coefficients are retained, denoted by $\boldsymbol{V_1}$, because the lasso fails to provide evidence to suggest that they violate the exclusion restriction.

3.4.2. Step 2: Selection of Strong Instruments.

To select a set of sufficiently strong instruments from the remaining set of candidate instruments, for a given endogenous covariate, we adopt another heuristic approach motivated by Belloni et al. (2012). Consider $\widehat{X}^{(1)}$ as the endogenous covariate, and $\boldsymbol{V_1}$ as the set of instruments obtained in Step 1. We estimate a lasso regression in which the endogenous covariate is regressed on all available instruments (similar to the first stage of a 2SLS estimation), that is, $\widehat{X}^{(1)} \sim \boldsymbol{V_1}$. The lasso attempts to shrink the coefficients of instruments that are conditionally uncorrelated with the endogenous covariate (i.e., weak in the presence of the other

instruments) to zero, with the nonzero coefficients indicating a set of regressors predictive of $\widehat{X}^{(1)}$, which intuitively correspond to *strong* instruments. Because this step requires *only the predictions* from individual trees, it is carried out on $D_{test} \cup D_{unlabel}$. In both lasso-based selection steps, the penalty level (λ) is automatically chosen in a data-driven manner (see Belloni et al. 2012, appendix A for implementation details).

Importantly, if the set of instruments changes during the two-step selection process (i.e., certain instruments are determined to violate exclusion or relevance requirements, and are thus dropped), we will *repeat* both of the lasso selection steps (1 and 2) with the remaining instruments, until the selection ceases to change. This iterative approach increases the likelihood that our selected instruments are simultaneously valid and strong. Moreover, we expect this procedure should work well with sufficient data, because when $\boldsymbol{V_i}$ contains only excluded instruments (satisfied asymptotically based on Theorem 1), Belloni et al. (2012) shows that asymptotically $\boldsymbol{S_i}$ will be the linearly optimal set of instruments. The instrumental variable selection procedure in Steps 1 and 2 is summarized in Algorithm 1.

In general, for each $\widehat{X}^{(i)}, i \in \{1, \ldots, M\}$, we can use this procedure to select a set of strong, excluded instruments, $\boldsymbol{S_i}$.

Algorithm 1 (Instrumental Variables Selection Procedure)

Data: Individual trees' predictions $\boldsymbol{P} = \{\widehat{X}^{(1)}, \ldots, \widehat{X}^{(M)}\}$ on D_{test} and $D_{unlabel}$ and ground truth X on D_{test}

Notation: denote $\|.\|_1$ as the L1-norm, $\|.\|_2$ as the L2-norm, λ as the lasso penalty level, $\boldsymbol{\Delta} = \{\delta_j\}$ and $\boldsymbol{\Gamma} = \{\gamma_j\}$ as coefficients of lasso regressions;

Set $\widehat{X}^{(i)}$ as the endogenous covariate;

Set $\boldsymbol{P_{-i}} \leftarrow \boldsymbol{P} \backslash \widehat{X}^{(i)}$ as the pool of candidate instruments;

Set $CurrIVs \leftarrow \boldsymbol{P_{-i}}$;

while *True* **do**

  //Step 1: removal of invalid instruments

  Obtain $e^{(i)} = \widehat{X}^{(i)} - X$ on D_{test};

  Estimate lasso regression $\min_{\boldsymbol{\Delta}} \left\{ \frac{1}{|D_{test}|} \| e^{(i)} - \sum_{\widehat{X}^{(j)} \in CurrIVs} \delta_j \widehat{X}^{(j)} \|_2^2 + \lambda \|\boldsymbol{\Delta}\|_1 \right\}$;

  Get $\boldsymbol{V_i} \leftarrow \{\widehat{X}^{(j)} \in CurrIVs | \delta_j = 0\}$ as the set of instruments with *zero* coefficients;

  //Step 2: selection of strong instruments

  On $D_{test} \cup D_{unlabel}$, estimate lasso regression $\min_{\boldsymbol{\Gamma}} \left\{ \frac{1}{|D_{test} \cup D_{unlabel}|} \left\| \widehat{X}^{(i)} - \sum_{\widehat{X}^{(j)} \in \boldsymbol{V_i}} \gamma_j \widehat{X}^{(j)} \right\|_2^2 + \lambda \|\boldsymbol{\Gamma}\|_1 \right\}$;

  Get $\boldsymbol{S_i} \leftarrow \{\widehat{X}^{(j)} \in \boldsymbol{V_i} | \gamma_j \neq 0\}$ as the set of instruments with *nonzero* coefficients;

  if $\boldsymbol{S_i} == CurrIVs$ **then**

    Break;//remaining instruments are valid and strong

  end

  Set $CurrIVs \leftarrow \boldsymbol{S_i}$;//repeat selection

end

Output: $\boldsymbol{S_i}$, the set of valid and strong instruments for $\widehat{X}^{(i)}$.

3.4.3. Step 3: Estimation. Consider the econometric model of interest, $Y = X\beta_X + \boldsymbol{Z}\boldsymbol{\beta_Z} + \varepsilon$. If all variables in the econometric model, i.e., $\{Y, X, \boldsymbol{Z}\}$, are directly observable in D_{label}, then one can obtain unbiased estimates of $\{\beta_X, \boldsymbol{\beta_Z}\}$, denoted as $\widehat{\boldsymbol{\beta}}_{label}$, by estimating the econometric model on the error-free D_{label}. In practice, the size of D_{label} is often limited, due to the cost of acquiring labels. As a result, $\widehat{\boldsymbol{\beta}}_{label}$ may exhibit a large standard error and would not be particularly suitable for drawing causal inferences. Nonetheless, the estimates are unbiased and can be used as a useful baseline for determining the quality of instrument estimations.[4] In this section, we discuss the final selection procedure, based on a comparison of 2SLS estimates and the unbiased estimation using precisely measured covariates in D_{label}, that is, $\widehat{\boldsymbol{\beta}}_{label}$.

Specifically, following Steps 1 and 2, we obtain a set of 2SLS estimates for each pair of $(\widehat{X}^{(i)}, \boldsymbol{S_i})$, using $\widehat{X}^{(i)}$ as the endogenous covariate and $\boldsymbol{S_i}$ as its instruments, denoted as $\widehat{\boldsymbol{\beta}}^i_{IV}$. Denote the variance-covariance matrices of $\widehat{\boldsymbol{\beta}}^i_{IV}$ and $\widehat{\boldsymbol{\beta}}_{label}$ as $\widehat{\boldsymbol{\Sigma}}^i_{IV}$ and $\widehat{\boldsymbol{\Sigma}}_{label}$, respectively. To compare $\widehat{\boldsymbol{\beta}}^i_{IV}$ with $\widehat{\boldsymbol{\beta}}_{label}$, we use the Hotelling's T^2 test with unequal variance (Seber 2009). This test is a multivariate generalization of the T test, designed to evaluate a null hypothesis of mean equality between two vectors of random variables, whose joint distributions have unequal variances. The test statistic is $H_i = (\widehat{\boldsymbol{\beta}}^i_{IV} - \widehat{\boldsymbol{\beta}}_{label})^T (\widehat{\boldsymbol{\Sigma}}^i_{IV} + \widehat{\boldsymbol{\Sigma}}_{label})^{-1} (\widehat{\boldsymbol{\beta}}^i_{IV} - \widehat{\boldsymbol{\beta}}_{label})$. H_i is asymptotically distributed as $\chi^2(K)$, where K represents the total number of covariates (Seber 2009). Therefore, if H_i is larger than the critical value of $\chi^2(K)$ at a user-chosen significance level (we pick $\alpha = 0.05$ for illustration purposes), and then $\widehat{\boldsymbol{\beta}}^i_{IV}$ is significantly different from the unbiased $\widehat{\boldsymbol{\beta}}_{label}$, indicating that one or more instruments likely violate the relevance and/or exclusion requirement. Such $\widehat{\boldsymbol{\beta}}^i_{IV}$ is then discarded.

Meanwhile, suppose more than one $(\widehat{X}^{(i)}, \boldsymbol{S_i})$ pair has an associated Hotelling statistic below the critical value, such that each pair empirically yields estimates that are not significantly different from $\widehat{\boldsymbol{\beta}}_{label}$. We then measure the empirical bias and variance of each set of

estimates, using *empirical MSE*, as follows:

$$MSE_i = tr\Big((\widehat{\boldsymbol{\beta}}^i_{IV} - \widehat{\boldsymbol{\beta}}_{label})(\widehat{\boldsymbol{\beta}}^i_{IV} - \widehat{\boldsymbol{\beta}}_{label})^T + \widehat{\boldsymbol{\Sigma}}^i_{IV}\Big).$$

Finally, we select the estimates that have the smallest empirical MSE, that is, the estimates that have the smallest sum of empirical bias and variance. Testing these multiple hypotheses, without control, will potentially inflate our type I error: the probability of incorrectly rejecting a truly valid tuple. Given that we likely have multiple tuples to choose from, our preference is to be conservative (over-reject tuples) and we prefer to make *no* inference rather than make an erroneous inference.

The 2SLS variances (i.e., diagonal elements of $\widehat{\boldsymbol{\Sigma}}^i_{IV}$) associated with ForestIV estimates do not reflect the estimator's true variability, because we intentionally select estimates with low variances. To quantify the variability of the ForestIV estimator, we propose to bootstrap $\{D_{label}, D_{unlabel}\}$ and apply the entire ForestIV procedure (including all the instrument selection and estimation steps) on each bootstrap sample, in order to obtain the empirical sampling distributions of ForestIV estimates. The standard deviations of the sampling distributions can serve as the standard errors of ForestIV estimates. Meanwhile, statistical inference following variable/model selection is a challenging problem (Berk et al. 2013, Taylor and Tibshirani 2015, Lee et al. 2016), and because the two instrument selection steps in ForestIV are carried out iteratively, establishing theoretical guarantees of postselection inference for ForestIV may be even more complex. Therefore, we believe that developing the theoretical guarantees of the bootstrapping procedure (and inference in general) for ForestIV is an interesting future research direction. We summarize the pseudocode for ForestIV in Algorithm 2 and the pseudocode for the bootstrapping procedure in Algorithm 3.

Algorithm 2 (Pseudocode for ForestIV Approach)

Data: Individual trees' predictions $\boldsymbol{P} = \{\widehat{X}^{(1)}, \ldots, \widehat{X}^{(M)}\}$ on D_{test} and $D_{unlabel}$ and ground truth X on D_{test}
Estimate $\widehat{\boldsymbol{\beta}}_{label}$ as unbiased coefficients on D_{label};
foreach $i \in \{1, \ldots, M\}$ **do**
  Set $\widehat{X}^{(i)}$ as the endogenous covariate in the econometric model of interest;
  Select $\boldsymbol{S_i} \subseteq \boldsymbol{P} \backslash \widehat{X}^{(i)}$ using Algorithm 1;
  //Step 3: estimation
  if $\boldsymbol{S_i} \neq \emptyset$ **then**
    Use $\boldsymbol{S_i}$ as instrumental variables, estimate $\widehat{\boldsymbol{\beta}}^i_{IV}$ on $D_{unlabel}$, with variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}^i_{IV}$;
    Calculate Hotelling's T^2 statistic, H_i, between $\widehat{\boldsymbol{\beta}}^i_{IV}$ and $\widehat{\boldsymbol{\beta}}_{label}$;
    if $H_i < CriticalValue$ **then**
      Retain $\widehat{\boldsymbol{\beta}}^i_{IV}$;
      Calculate $MSE_i = tr\Big((\widehat{\boldsymbol{\beta}}^i_{IV} - \widehat{\boldsymbol{\beta}}_{label})(\widehat{\boldsymbol{\beta}}^i_{IV} - \widehat{\boldsymbol{\beta}}_{label})^T + \widehat{\boldsymbol{\Sigma}}^i_{IV}\Big)$;
    end
  end
end
Output: The retained $\widehat{\boldsymbol{\beta}}^i_{IV}$ with the smallest MSE_i.

Algorithm 3 (Pseudocode for the Bootstrapping Procedure)

Data: Labeled data D_{label} and unlabeled data $D_{unlabel}$; the total number of bootstrapping rounds B
foreach $k \in \{1, \ldots, B\}$ **do**
  Draw $D^{(k)}_{label}$ from D_{label} with replacement; draw $D^{(k)}_{unlabel}$ from $D_{unlabel}$ with replacement;
  Partition $D^{(k)}_{label}$ into $D^{(k)}_{train}$ and $D^{(k)}_{test}$;
  Build a random forest model based on $D^{(k)}_{train}$ and make predictions for $D^{(k)}_{test}$ and $D^{(k)}_{unlabel}$;
  Apply the ForestIV approach (Algorithm 2) based on $D^{(k)}_{test}$ and $D^{(k)}_{unlabel}$ to obtain the estimates, denoted as $\widehat{\boldsymbol{\beta}}^{(k)}_{IV}$
end
Output: The standard deviations of $\{\widehat{\boldsymbol{\beta}}^{(1)}_{IV}, \ldots, \widehat{\boldsymbol{\beta}}^{(B)}_{IV}\}$ as standard error estimates.

4. Simulation Experiments

We now turn to the systematic evaluation of ForestIV's bias correction performance. Our evaluation is based on a semisynthetic strategy and leverages a combination of real data for machine learning tasks and simulated data for econometric estimation. Simulating the data for econometric models means that the true coefficients are known in advance, which allows us to objectively evaluate the correction performance of ForestIV.

4.1. Basic Simulation Setup

Our first demonstration uses the Bike Sharing Data (Fanaee-T and Gama 2014) for the machine learning task, which contains 17,379 instances of hourly bike rental activities. It is a commonly used data set for benchmarking and evaluating machine learning models (Giot and Cherrier 2014).

We randomly partition the data set, such that 1,000 observations serve as D_{train}, 200 observations serve as D_{test}, and the remaining 16,179 observations serve as $D_{unlabel}$. This represents a realistic scenario where $D_{unlabel}$ is much larger than D_{label}. Using D_{train}, we build a random forest model of 100 trees to predict the log-transformed count of total hourly bike rentals (denoted as *lnCnt*, log-transformed to reduce skewness), based on

12 features, including the time of the rental as well as weather and seasonal information. Importantly, the random forest generates aggregate (ensemble) predictions, as well as predictions from each of its individual trees. We denote $\widehat{lnCnt}$ as the aggregate predictions, and $\widehat{lnCnt}_i$ as the predictions from individual tree $i \in \{1, \ldots, 100\}$.

Next, we simulate an econometric model with *lnCnt* as an independent covariate. The model specification is $Y = 1 + 0.5lnCnt + 2Z_1 + Z_2 + \varepsilon$, where $Z_1 \sim Uniform[-10, 10]$, $Z_2 \sim N(0, 100)$, and $\varepsilon \sim N(0, 4)$. In the bike sharing data set, $\sigma_{lnCnt} = 1.5$, which is smaller than the standard deviation of the regression error term. Therefore, this simulation setup represents a realistic scenario where the variation in "noise" is comparable to that of the "signal". We keep the true regression coefficients fixed, that is, $\beta_0 = 1, \beta_{lnCnt} = 0.5, \beta_{Z_1} = 2, \beta_{Z_2} = 1$, to quantify the degree of estimation bias and the effectiveness of any corrections.

The previous simulation procedure is repeated for 100 rounds. Within each simulation round, a random forest model is built based on a random split of the data (as described earlier), and an artificial data set is generated for the econometric estimations. Specifically, we first estimate the *biased regression*, $Y \sim \beta_0 + \beta_1 \widehat{lnCnt} + \beta_2 Z_1 + \beta_3 Z_2$, where $\widehat{lnCnt}$ is the aggregate predictions from that round's random forest. This is precisely what would be done if one were to use the machine-learning-predicted covariate directly in the econometric model, without considering measurement error, as is commonly the case. We then apply ForestIV to obtain the corrected coefficients. Because an independent set of $\{D_{train}, D_{test}, D_{unlabel}\}$ is drawn in each simulation round, the ForestIV estimates across 100 simulation rounds naturally form the empirical sampling distributions. We report the standard deviations of the sampling distributions as the standard errors of ForestIV estimates.[5] Finally, besides the biased and ForestIV estimates, we also report the *unbiased estimates* from directly running the regression on D_{label}.

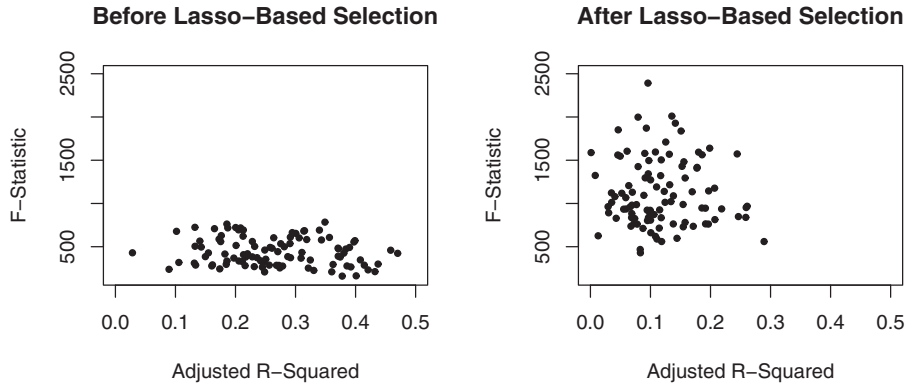
4.2. Basic Simulation Results

We first provide some descriptive evidence regarding the effectiveness of our two-step lasso regression procedure to select valid and strong instruments from candidates. On the bike sharing data, we calculate two metrics using D_{test}, where both the ground truth *lnCnt* and the random forest's predictions are observed. Before the two-step lasso-based selection procedure, for a given individual tree *i*, we treat *all* other trees' predictions as instruments. We calculate (1) the *F* statistic associated with a 2SLS regression, which serves as an illustrative measure of instrument strength and (2) the adjusted R^2 associated with an OLS regression of tree *i*'s prediction error on the instruments, which serves as an illustrative measure of instrument exclusion. Observing small *F* statistic and large R^2 indicate having weak and invalid instruments. After the selection, we again calculate these two metrics, using only *selected* instruments. We plot the *F* statistics against the adjusted R^2, both before and after the lasso-based selections, in Figure 1.

We can see that, after the lasso-based selections, the *F* statistics become much larger ($p < 0.001$), and the adjusted R^2 become even smaller ($p < 0.001$). This provides descriptive evidence that our lasso-based procedures are useful in further selecting strong and valid instruments.

We next present the main results. In Table 1, we report the average coefficients and standard errors (in parentheses) across all simulation rounds, both for the biased regression and the unbiased regression (obtained on the labeled data). For ForestIV, we report the average coefficients and the standard deviations of the sampling distributions across 100 simulation runs as standard errors. For each of the three estimators, we also compute

Figure 1. Plot of *F* Statistic Against Adjusted R^2 Based on a Single Simulation Run



Note. The left plot corresponds to 2SLS estimations using all candidate instruments (i.e., without selection), and the right plot corresponds to 2SLS estimations using only the selected instruments.

Table 1. ForestIV Results on Bike Sharing Data

	True	Biased	Unbiased	ForestIV
Intercept	1.0	0.708	0.999	0.958
		(0.093)	(0.162)	(0.118)
lnCnt	0.5	0.565	0.500	0.511
		(0.019)	(0.034)	(0.024)
Z_1	2.0	2.000	1.999	2.000
		(0.003)	(0.010)	(0.003)
Z_2	1.0	1.000	1.001	1.000
		(0.002)	(0.005)	(0.002)
RMSE		0.314	0.166	0.128

Notes. Standard errors in parentheses. RMSE contains the empirical RMSE associated with each set of estimates, averaged across 100 simulation runs.

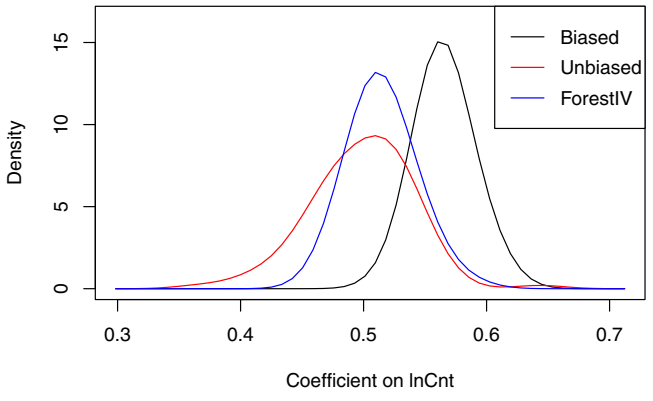
their RMSE, defined as $\sqrt{tr\left((\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{true})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{true})^T+\widehat{\boldsymbol{\Sigma}}\right)}$, which measures how close they are to the true coefficient values. In addition, we plot the sampling distributions of biased, unbiased, and ForestIV estimation on *lnCnt* (i.e., the covariate generated by machine learning) across all 100 simulations in Figure 2.

We make several observations of the results. First, directly using the $\widehat{lnCnt}$ predicted by random forest in the regression model clearly results in biases. The coefficient on *lnCnt* is *overestimated* on average by 13.0%, and the intercept term is underestimated by 29.2%. Second, compared with the biased regression, ForestIV effectively mitigates estimation bias on *lnCnt*. Third, compared with the unbiased estimation, ForestIV estimation has a "narrower" distribution and smaller standard errors, indicating sizable increase in estimation precision. In fact, based on the RMSE metric, ForestIV estimates are statistically closest to the true coefficients.

4.3. Sensitivity Analyses

We conduct three sets of sensitivity analyses to understand performance of ForestIV with respect to several parameters. First, we examine how ForestIV estimations change as the size of the *unlabeled* data set increases. This helps illustrate the asymptotic properties of ForestIV estimations. Second, we compare the ForestIV estimates with the unbiased estimates as we increase the size of the *labeled* data set, to understand when ForestIV can yield preferable estimates than the unbiased regression on the labeled data set. Third, we vary the total number of trees in the random forest model to explore the relationship between random forest's predictive performance and ForestIV effectiveness.

Figure 2. (Color online) Distributions of Biased, Unbiased, and ForestIV Estimation on *lnCnt* Across 100 Simulation Runs



4.3.1. Size of Unlabeled Data.

We repeat the main simulation with eight different sizes of unlabeled data, respectively, 100, 500, 1,000, 5,000, 7,500, 10,000, 12,500, and 16,179 (i.e., all remaining instances). Other parameters, for example, the size of the labeled data set and the econometric model specifications, are kept unchanged across these simulations. We plot the RMSE of biased, unbiased, and ForestIV estimates under different sizes of unlabeled data in Figure 3.

We observe that the biased estimates always have the highest RMSE regardless of increases in the volume of unlabeled data, which is expected given the fact that measurement error results in inconsistent estimates. On the other hand, with sufficient unlabeled data (more than 1,000 in this case), ForestIV estimates achieve lower RMSE, and hence more precise estimation, than the unbiased estimates.

4.3.2. Size of Labeled Data.

Next, we repeat the main simulation with five different sizes of labeled data, respectively, 1,200, 2,400, 3,500, 4,800, and 6,000, while keeping the same 5 : 1 training/testing split ratio as in the basic simulation setup. Other parameters are also kept the same. We plot the RMSE of biased, unbiased, and ForestIV estimates under different sizes of labeled data in Figure 4.

We find that ForestIV estimates achieve smaller RMSE than the unbiased estimates when the labeled

Figure 3. (Color online) RMSE of Biased, Unbiased, and ForestIV Estimates Under Different Sizes of Unlabeled Data

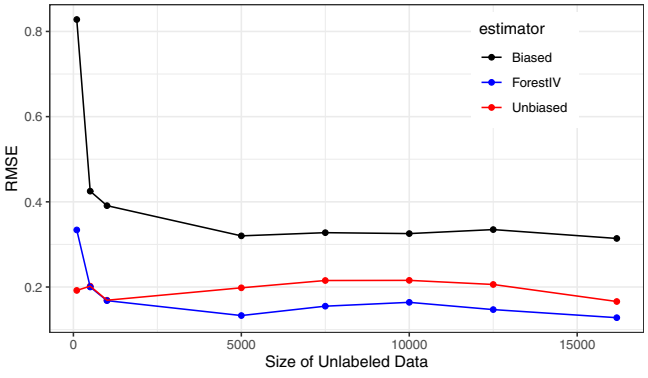
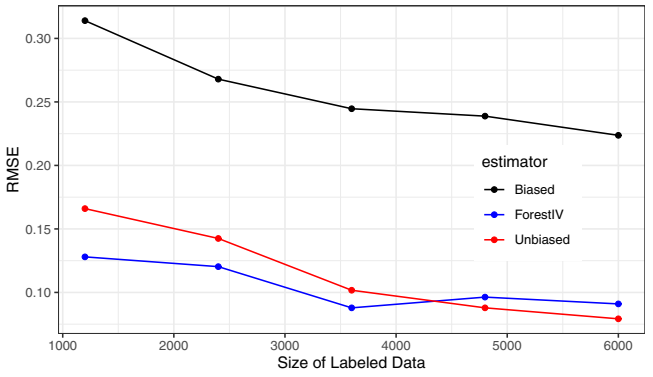


Figure 4. (Color online) RMSE of Biased, Unbiased, and ForestIV Estimates Under Different Sizes of Labeled Data



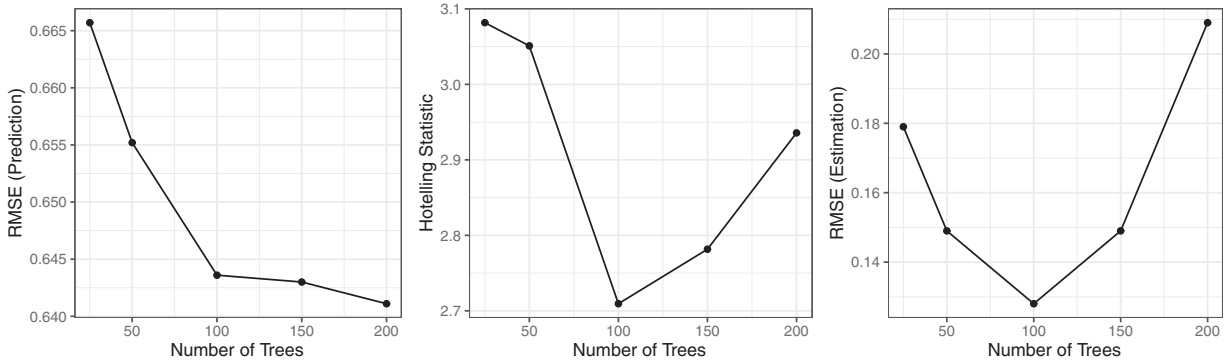
data set is relatively small, but the opposite is true when the size of labeled data set grows larger than 4,800 (or roughly 28% of all data). These two sensitivity analyses collectively suggest that, in practice, the benefit of combining machine learning with econometric modeling is most salient when the unlabeled data are much larger than the labeled data (e.g., when acquiring a large amount of labeled data are prohibitively expensive). ForestIV can offer substantial utility in such a scenario. As the volume of unlabeled data grows larger, ForestIV brings gains in estimation precision relative to the unbiased estimates obtained using the (relatively small) labeled data set. If a sufficiently large labeled data set can be acquired, then it is more advantageous to rely on the unbiased regression, and there is less need to build a machine learning model (or to use ForestIV on the unlabeled data set).

4.3.3. Predictive Performance of Random Forest. Thus far, we have treated the random forest model as fixed. However, the standard procedure of building a predictive machine learning model typically involves tuning various parameters to achieve the best predictive performance on the hold-out labeled data. Therefore, in the next set of sensitivity analyses, we examine the relationship between the predictive performance of random forest (measured using root mean square error (RMSE), a common performance metric in machine learning) and the correction performance of ForestIV. Given that many parameters of a random forest can be fine-tuned, we choose to focus on the total number of trees (i.e., M), because this parameter is directly related to the number of candidate instruments. We repeat the simulation with five different values of M: 25, 50, 100, 150, and 200. Across these simulations, the other parameters remain as described in the main simulation. In Figure 5, we plot (1) the prediction RMSE of the forest, (2) the Hotelling T^2 statistic of ForestIV estimates (averaged over 100 simulation runs), and (3) the RMSE of ForestIV estimates, across different number of trees.

We see that having too few trees (e.g., $M = 25$ or $M = 50$) result in higher prediction RMSE *and* higher estimation RMSE (indicative of poor correction performance). Meanwhile, as the number of trees increases, we observe marginal improvements on RMSE, but this does not lead to better correction performance. Specifically, when $M = 150$ and $M = 200$, we observe slight improvements in RMSE scores compared with when $M = 100$, but the estimation RMSE increases again. Importantly, the average Hotelling T^2 statistic that compares $\widehat{\boldsymbol{\beta}}_{label}$ with ForestIV estimates aligns with ForestIV's correction performance. Smaller T^2 statistic also corresponds to lower estimation RMSE.

This set of simulations implies that ForestIV cannot replace efforts to tune the hyper-parameters associated with the random forest prediction model. That is, having a better-performing random forest tends to improve the performance of ForestIV on average. At the same time, in this particular demonstration, we observe that using an excessive number of trees can potentially hurt ForestIV's correction performance, likely because instrument selection can be challenging with a large number of candidate instruments. In practice, researchers can

Figure 5. Prediction RMSE, Hotelling T^2 Statistic, and Estimation RMSE Under Different Number of Trees



rely on the Hotelling T^2 statistic as a signal and select the number of trees that minimizes T^2 statistic.

4.4. Evaluations of Alternative Designs

Taking the proposed ForestIV procedure as a baseline, we also consider three possible alternative design choices, related to instrument discovery, selection, and estimation. First, a key feature of ForestIV is that we "break up" a random forest and exploit individual trees to discover instruments for bias correction. As an alternative, one might instead consider a *sample splitting* approach, where D_{train} is split into two independent subsets and a random forest is built on each subset. Then, predictions from one forest might serve as an instrument for predictions from the other. This approach is intuitively appealing because the predictive performance of a forest is arguably better than that of a single tree, thereby reducing the measurement error problem to start with. In the same vein, a second alternative design of the instrument selection procedure in ForestIV is to use the aggregate predictions from a *subset of trees* (rather than predictions from a single tree) as the endogenous covariate and instrumental variables. Third, in the estimation step of ForestIV, whereas we propose to select the estimates that minimize empirical MSE, an alternative design is to *average* across all estimates that are not rejected by the Hotelling's T^2 test.

We implement and empirically evaluate the relative benefit of the three alternative design choices. We refer readers to Online Appendix C for a detailed discussion of each evaluation. On the whole, we find that none of the previous three alternative designs outperforms our proposed ForestIV method (at least in the context of our bike sharing data and associated simulation setup). That said, the possibility remains that one or more of these alternatives could yield improvements under certain conditions, and each holds promise as an avenue of further exploration in future work.

5. ForestIV with Binary Endogenous Covariates

In this section, we turn to the case of a binary misclassified (and thus endogenous) covariate, as would be generated by a machine-learning classifier model. It turns out that ForestIV still exhibits an ability to generate (and select) instrumental variables that produce improved estimates, although the underlying mechanisms are somewhat different from the case of a continuous endogenous covariate.

5.1. Theoretical Results

Suppose the binary outcome label can have values of zero (negative class) or one (positive class). Consider a random forest *classifier* with M decision trees. We again use notations X, $\widehat{X}$, and $\widehat{X}^{(i)}$ to represent the ground truth, prediction of the forest, and prediction of individual tree i. The prediction error of individual tree i is defined as $e^{(i)} = \widehat{X}^{(i)} - X$. For any given data point, $e^{(i)}$ can take three possible values: 0 (correct prediction), 1 (false positive), and –1 (false negative).

Meanwhile, based on the econometric literature (Angrist and Pischke 2008), instrumental variables with binary endogenous covariates can be applied in the same manner as in the continuous case; that is, treat the variable as continuous, and use a 2SLS estimation. The first stage of this 2SLS estimation amounts to a linear probability model. The predicted values for the endogenous covariate recovered from the first-stage regression thus reflect exogenous variation in continuous class probabilities. These values would then be used in the second stage regression. Intuitively, to evaluate whether predictions from one tree, j, can serve as a valid instrument for predictions from another tree, i, we need to assess (1) $Cov(\widehat{X}^{(i)}, \widehat{X}^{(j)})$ (the relevance condition) and (2) $Cov(e^{(i)}, \widehat{X}^{(j)})$ (the exclusion restriction).

The relevance condition is typically satisfied, because two individual trees from a reasonably well-performing random forest are both somewhat predictive of the outcome, that is, $Cov(\widehat{X}^{(i)}, \widehat{X}^{(j)}) \neq 0$. The exclusion restriction can be written as $Cov(e^{(i)}, e^{(j)} + X)$. Accordingly, we next provide several theoretical results to characterize $Cov(e^{(i)}, e^{(j)})$ and $Cov(e^{(i)}, X)$, respectively.

Theorem 2. *The error rate of a random forest binary classifier decreases with* $\mathbb{E}_j\mathbb{E}_i Corr(|e^{(i)}|, |e^{(j)}|)$, *where* $e^{(i)}$ *and* $e^{(j)}$ *are prediction errors of tree* i *and tree* j $(i \neq j)$.

All proofs are included in Online Appendix A. This theorem suggests that a well-performing random forest would have relatively small $Corr(|e^{(i)}|, |e^{(j)}|)$. Typically, because $Corr(|e^{(i)}|, |e^{(j)}|) \neq Corr(e^{(i)}, e^{(j)})$, this result seems to indicate that the exclusion restriction of instrument validity, $Cov(e^{(i)}, e^{(j)} + X) = 0$, is not satisfied. However, the following two theorems show that $Cov(e^{(i)}, X)$ is always nonzero (i.e., the classical measurement error assumption is never true for binary misclassification), and can be offset, or cancelled out, by $Cov(e^{(i)}, e^{(j)})$, thereby making the exclusion condition plausible.

Theorem 3. *For all* $i \in \{1, \ldots, M\}$, $Cov(e^{(i)}, X) < 0$.

For notation simplicity, denote the probability that $X = \alpha, \widehat{X}^{(i)} = \beta, \widehat{X}^{(j)} = \gamma$ as $p_{\alpha\beta\gamma}$ $(\alpha, \beta, \gamma \in \{0, 1\})$, and denote the probability that $X = \alpha$ as $p_{\alpha\bullet\bullet}$.

Theorem 4. *For all* $i, j \in \{1, \ldots, M\}$ *and* $i \neq j$, $Cov(e^{(i)}, e^{(j)}) > 0$ *if and only if* $(p_{000} + p_{111})(p_{011} + p_{100}) + 2(p_{0\bullet\bullet} - p_{000})p_{100} + 2(p_{1\bullet\bullet} - p_{111})p_{011} + (p_{010} - p_{101})(p_{110} - p_{001}) > 0$.

Theorem 3 shows that the prediction error of an individual tree is always negatively correlated with the ground truth. Theorem 4 suggests that, a few corner cases aside, the correlation between prediction errors of two individual trees is positive.[6] As a result, $Cov(e^{(i)}, X)$ is likely to offset $Cov(e^{(i)}, e^{(j)})$, leading to a relatively small value of $Cov(e^{(i)}, e^{(j)} + X)$.[7] In other words, in the case of a binary endogenous covariate generated by a random forest classifier, other trees' predictions can still plausibly serve as instrumental variables.

5.2. Simulation Experiments

We demonstrate the performance of ForestIV in the case of binary classification using the bank marketing data (Moro et al. 2014) as an example data set, which contains 45,211 records related to a bank's telemarketing efforts. We randomly partitioned the data into 1,500 observations to serve as D_{train}, 500 observations to serve as D_{test}, and the remaining 43,211 observations to serve as $D_{unlabel}$. Using the training data, we build a random forest classifier comprised of 100 trees to predict a binary outcome, *Deposit*, representing whether a client subscribed to a term deposit as a result of the phone call, based on 16 attributes describing the client and the marketing campaign.

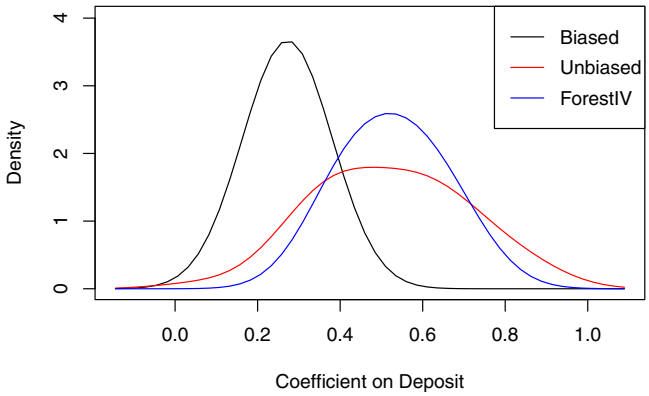
We next simulate an econometric model: $Y = 1 + 0.5Deposit + 2Z_1 + Z_2 + \varepsilon$, where $Z_1 \sim Uniform[-1, 1]$, $Z_2 \sim N(0, 1)$, and $\varepsilon \sim N(0, 4)$. As before, we repeat the simulation for 100 rounds. Within each round, we estimate the biased regression (directly using the random forest predictions, $\widehat{Deposit}$, in the regression), the unbiased regression obtained on D_{label}, and the corrected coefficient obtained from the ForestIV procedure. In Table 2, we report the average coefficients and standard errors (in parentheses) across all simulation rounds for the biased regression and the unbiased regression. For ForestIV, we report the average coefficients and standard deviations of the sampling distributions across 100 simulation rounds as standard errors. We also plot the distributions of biased, unbiased, and ForestIV estimation on *Deposit* in Figure 6.

Table 2. ForestIV Results on Bank Marketing Data

	True	Biased	Unbiased	ForestIV
Intercept	1.0	1.041	0.992	0.994
		(0.010)	(0.040)	(0.012)
Deposit	0.5	0.276	0.523	0.521
		(0.041)	(0.140)	(0.102)
Z_1	2.0	1.995	1.996	1.995
		(0.017)	(0.071)	(0.017)
Z_2	1.0	1.000	1.005	1.000
		(0.009)	(0.041)	(0.009)
RMSE		0.233	0.169	0.107

Notes. Standard errors in parentheses. RMSE contains the empirical RMSE associated with each set of estimates, averaged across 100 simulation runs.

Figure 6. (Color online) Distribution of Biased, Unbiased, and ForestIV Estimation on *Deposit* Across 100 Simulation Runs



Consistent with previous simulations, we observe that directly using the random forest's predictions in our regression leads to 44.8% underestimation on *Deposit* on average. ForestIV is again effective in mitigating the biases. Finally, compared with unbiased estimation, ForestIV achieves gains in estimation precision, as indicated by its "narrower" distribution. In general, this set of simulations confirms the effectiveness of ForestIV for a binary endogenous covariate. We repeat all the sensitivity analyses that were conducted previously in the continuous covariate scenario and observe consistent insights. All results are described in Online Appendix D.

6. Benchmarking ForestIV with Three Alternative Approaches

In the econometrics literature, researchers have proposed several methods to correct for estimation biases caused by measurement error (we refer to Grace 2016 for a review). Here, we benchmark ForestIV against three alternative bias correction approaches: (i) simulation-extrapolation (SIMEX), (ii) regression adjustment for nonparametrically generated regressors, and (iii) latent instrumental variables (LatentIV). We provide a brief description of each method.

First, SIMEX (Cook and Stefanski 1994) is a general simulation-based approach that can be applied to address measurement error in any econometric model. It directly leverages error magnitude information (e.g., error variance, based on the performance of machine learning predictions) to create a set of bootstrap samples of the observed data, artificially introducing larger measurement error with each subsequent resampling. The algorithm then estimates a corresponding set of coefficients with respect to different degrees of measurement errors, fits a parametric function to the pairs of coefficient-error observations, and finally extrapolates the coefficient estimates to the case where measurement error is zero. Although the original SIMEX method was proposed to address measurement error in a continuous covariate, researchers later developed

Table 3. Benchmarking Results (Bike Sharing Data)

	True	Biased	Unbiased	ForestIV	SIMEX	Reg. Adj.	LatentIV
Intercept	1.0	0.708	0.999	0.958	1.079	N.A.	0.040
		(0.093)	(0.162)	(0.118)	(0.055)		(0.018)
lnCnt	0.5	0.565	0.500	0.511	0.483	0.564	0.584
		(0.019)	(0.034)	(0.024)	(0.012)	(0.147)	(0.022)
Z_1	2.0	2.000	1.999	2.000	2.000	2.000	N.A.
		(0.003)	(0.010)	(0.003)	(0.003)	(0.003)	
Z_2	1.0	1.000	1.001	1.000	1.000	1.000	N.A.
		(0.002)	(0.005)	(0.002)	(0.002)	(0.002)	
RMSE		0.314	0.166	0.128	0.126	0.344	2.242

Notes. Standard errors in parentheses. RMSE contains the empirical RMSE associated with each set of estimates, averaged across 100 simulation runs. N.A., not applicable; Reg. Adj., regression adjustment.

a variation named Misclassification-SIMEX (MC-SIMEX) to deal with misclassification in a discrete covariate (Küchenhoff et al. 2006, 2007). Flexibility is a key advantage of SIMEX. It only requires aggregated information about the measurement error, for example, error variance for a continuous covariate or recall rates for a binary covariate, which can be estimated from the testing data. Moreover, SIMEX can be applied to a large set of econometric models with a standard procedure and does not need to be explicitly reformulated for each econometric model specification. The effectiveness of SIMEX for correcting biases due to measurement error has been comprehensively documented by Yang et al. (2018) for a variety of estimators and econometric specifications. For the sake of brevity, we refer to Yang et al. (2018) for technical details on applying the SIMEX approach.

Second, we consider a particular regression adjustment approach proposed in the generated regressors literature. Meng et al. (2016) studies the estimation of a linear regression where one covariate is not directly observed but can be approximated based on a sample of relevant data (e.g., a nation's income inequality measure is unobserved but can be estimated from a sample of individual incomes in the nation). Because sample-based approximation of the unobserved covariate can be noisy, the linear regression suffers from a measurement error issue. The authors assume that the functional relationship underlying the generated regressor is inconsistent across observations, and thus the generated regressor can be viewed as nonparametric. Meng et al. (2016) derive an explicit formula for the magnitude of bias, as a function of the first two *moments* of the measurement error (e.g., mean and variance), which allows the biased estimates to be adjusted accordingly. In our setting, the moment statistics of measurement/prediction error can be readily estimated using the testing data, where prediction errors are directly observed. We are therefore able to implement the proposed adjustment approach of Meng et al. (2016) in our case.

Finally, the LatentIV approach is proposed by Ebbes et al. (2005, 2009) to address endogeneity in linear regression models without the use of observable instruments. LatentIV achieves identification by modeling a latent (i.e., unobserved) discrete instrument to account for the correlation between the endogenous covariate and the regression model's error term. We adopt the implementation of LatentIV in the R package "REndo."

We benchmark ForestIV against all three alternative approaches, both with a continuous endogenous covariate (using the bike sharing data) and with a binary endogenous covariate (using the bank marketing data). The setups are the same as in the main simulations. We report the estimation results in Tables 3 and 4. Meng et al. (2016) do not explicitly discuss how to adjust for

Table 4. Benchmarking Results (Bank Marketing Data)

	True	Biased	Unbiased	ForestIV	MC-SIMEX	Reg. Adj.	LatentIV
Intercept	1.0	1.041	0.992	0.994	1.041	N.A.	−0.435
		(0.010)	(0.040)	(0.012)	(0.010)		(0.010)
Deposit	0.5	0.276	0.523	0.521	0.735	0.262	−1.435
		(0.041)	(0.140)	(0.102)	(0.110)	(0.994)	(1.624)
Z_1	2.0	1.995	1.996	1.995	2.000	1.995	N.A.
		(0.017)	(0.071)	(0.017)	(0.017)	(0.017)	
Z_2	1.0	1.000	1.005	1.000	1.001	1.000	N.A.
		(0.009)	(0.041)	(0.009)	(0.010)	(0.009)	
RMSE		0.233	0.169	0.107	0.270	1.030	3.935

Notes. Standard errors in parentheses. RMSE contains the empirical RMSE associated with each set of estimates, averaged across 100 simulation runs.

bias in the intercept term, and we therefore omit the intercept term estimation for the regression adjustment approach. Additionally, the REndo package does not support the estimation of other exogenous control variables in the model—they are also omitted.[8]

In the case of a continuous endogenous covariate, we find that ForestIV outperforms the regression adjustment and LatentIV approaches. ForestIV achieves comparable correction performance as SIMEX in terms of estimation RMSE—although ForestIV produces point estimates that are closer to the true values, SIMEX estimates have smaller standard errors. In the case of a binary endogenous covariate, ForestIV clearly outperforms all three alternative approaches.

As part of our benchmarking exercise, we have also identified a systematic "blindspot" in the SIMEX correction procedure that occurs when measurement error is correlated with other precisely measured covariates in an econometric model. Although SIMEX will produce provably erroneous correction results in the presence of such correlation (see Online Appendix E for the proof), empirical results suggest that our ForestIV approach does not suffer from this limitation. We demonstrate this on the bike sharing data using the same simulation setup as before, except that we generate a control variable, $Z_2 \sim N(0,1)$, to be correlated with the prediction error in *lnCnt* with a correlation coefficient of 0.3. The estimation results, summarized in Table 5, show that the coefficient on Z_2 is underestimated due to the (correlated) measurement error. SIMEX fails to mitigate the estimation bias in Z_2 (and in fact exacerbates such bias). In contrast, ForestIV is able to not only mitigate the bias on *lnCnt*, but also notably correct the coefficient on Z_2 in the right direction. Overall, ForestIV represents a more robust bias correction approach than SIMEX.

Meanwhile, we acknowledge that alternative methodologies which have seen recent theoretical development in the econometrics literature may achieve better bias correction results than ForestIV in certain situations. As such, we believe this to be a potentially quite fruitful future research direction by which our method could be improved upon.

Table 5. Benchmarking Results (Bike Sharing Data) with $\rho_{Z_2,e} = 0.3$

	True	Biased	Unbiased	ForestIV	SIMEX
Intercept	1.0	0.729	0.993	0.950	1.082
		(0.086)	(0.175)	(0.125)	(0.062)
lnCnt	0.5	0.559	0.501	0.511	0.482
		(0.017)	(0.037)	(0.025)	(0.013)
Z_1	2.0	2.000	1.999	2.000	2.000
		(0.003)	(0.010)	(0.003)	(0.003)
Z_2	1.0	0.911	0.983	0.921	−0.085
		(0.016)	(0.051)	(0.019)	(0.018)
RMSE		0.304	0.188	0.160	1.095

Notes. Standard errors in parentheses. RMSE contains the empirical RMSE associated with each set of estimates, averaged across 100 simulation runs.

7. Conclusion and Future Work

To summarize, we introduce a new approach, ForestIV, which addresses bias in regression estimates that is attributable to (predictive) measurement error in data-mined covariates. With a continuous endogenous covariate, the intuition behind ForestIV is that a high-performing random forest will be comprised of (i) trees that are individually accurate in their predictions and thus overlap, offering "repeated measures" of true, exogenous variation in the data-mined variable, and (ii) trees that exhibit low correlations in their prediction errors, which in tandem with the former point implies that trees make "different" mistakes, and thus embed orthogonal measurement errors. Our approach is closely connected to the idea of using multiple error-prone measures as instrumental variables (Blackburn and Neumark 1992, Hausman et al. 1995, Lewbel 2019). With a binary endogenous covariate, although trees no longer necessarily have low error correlations, we show that instrument validity is nonetheless plausible.

The application of our approach in empirical contexts has the potential to improve the precision and robustness of estimations and thus subsequent decision making. Meanwhile, our approach shows the possibility of automatically generating candidate instruments based on an ensemble learning technique, which complements the emerging literature on the use of machine learning methods for causal inference (Athey and Imbens 2016, McFowland et al. 2018). At the core of ForestIV is the fundamental tradeoff between an estimator's bias and variance, which together describe its statistical risk. ForestIV attempts to provide an estimate with reduced overall risk, one with substantially lower bias than the biased regression and lower variance than the unbiased regression.

In practice, because true coefficients are not known a priori, it is useful to have a few guidelines to gauge the effectiveness of ForestIV in a particular finite sample.

1. Using the hold-out data set (e.g., D_{test}), researchers can empirically evaluate instrument validity and strength, both before and after the proposed two-step lasso-based selections.

2. The Hotelling T^2 test statistic can also be a useful signal. The p value associated with the Hotelling T^2 test comparing $\widehat{\boldsymbol{\beta}}_{label}$ and ForestIV estimates indicates the probability of observing their empirical differences under the null of equality. Researchers can define the threshold of evidence they require before accepting ForestIV estimates by adjusting the significance level for this test.

3. Researchers can also examine whether the asymptotic properties of ForestIV have yet to "kick-in" by examining empirical convergence in resulting coefficient estimates as the procedure is exposed to more unlabeled data. If a convergence plot indicates that the coefficient estimates have not yet plateaued, this may be a sign that ForestIV estimates have not yet converged, and that more unlabeled data are perhaps needed.

4. To guide the parameterization of ForestIV (e.g., the choice of M and other tuning parameters of the random forest), we also recommend a diagnostic procedure that leverages the *labeled data set* to run two analyses: (1) an *unbiased* regression (using the observed true labels); and (2) a *corrected* regression based on ForestIV, with a particular set of parameters. By comparing the coefficient estimates between (1) and (2), one can gauge the effectiveness of bias correction under a particular ForestIV configuration. Repeating this procedure with different sets of parameter configurations, one can then select the configuration that produces the most effective correction results.

5. To assess whether ForestIV estimates or the estimates from unbiased regression should be used for inference, researchers could potentially consider applying specification tests, such as the Durbin-Wu-Hausman test (Hausman 1978). However, the validity and properties of this test remain to be established by future work.

Finally, to better delineate the use case of ForestIV, we reiterate that if the size of labeled data are large enough that $\widehat{\boldsymbol{\beta}}_{label}$ can be estimated reliably and precisely enough using only the available labeled data, then there is no need to mine variables in the first place. One should simply take $\widehat{\boldsymbol{\beta}}_{label}$ and proceed with inferences and decision making. Accordingly, statistical power analyses are likely to be useful when determining whether "big data" and machine learning methods are needed for a particular inference problem (Ellis 2010).

Several future research directions are worth pursuing. For example, although this paper focuses on selecting valid and strong instruments from a given random forest, future work might look to leverage a specialized random forest algorithm that explicitly aims to minimize individual trees' prediction error correlations. The rotation forest algorithm (Blaser and Fryzlewicz 2016) and the dynamic random forest algorithm (Bernard et al. 2012) represent two such attempts. Another direction to pursue is to generalize ForestIV to bagging-based machine learning models, more broadly. Intuitively, because individual learners from a bagging model are trained on different bootstrap samples of the training data, they are likely to generate correlated predictions and weakly correlated prediction errors. Future work can investigate whether general bagging models can produce useful instruments, whether the types of individual learners (e.g., decision tree or other techniques) affect the validity of instruments, and the performance of those alternative ensembles relative to ForestIV.

Endnotes

[1] We implement ForestIV as an R package (https://github.com/mochenyang/ForestIV), and illustrate its use in a blog post (https://mochenyang.github.io/mochenyangblog/research/2022/01/10/ForestIV.html).

[2] As an example, purely randomly generated instruments are typically not useful, because they are not correlated with the endogenous covariates (i.e., they violate the relevance condition).

[3] For example, a binary response model, Probit or Logit, can be formulated as a latent linear model with a binary transformation of the dependent variable, with the measurement error specified for the latent linear model.

[4] If D_{label} is sufficiently large, implying that researchers can afford to acquire a large set of labeled data with sufficient statistical power to estimate the econometric model of interest, then it is unnecessary to use machine learning models to construct/mine variables in the first place; the researcher should simply estimate the econometric model on D_{label}.

[5] We also obtained standard errors from our bootstrap approach, and the estimates are highly similar.

[6] One such corner case that $Cov(e^{(i)}, e^{(j)}) < 0$ is when $p_{011} = p_{100} = 0$ and $(p_{010} - p_{101})(p_{110} - p_{001}) < 0$, which means that the predictions from the two trees are never wrong at the same time, and their misclassification patterns satisfy a stringent condition. Such corner case is not easily realized.

[7] This is empirically supported in our simulation experiments using the Bank Marketing data, discussed in the next section. In particular, the average covariance between prediction errors of any two trees, $\mathbb{E}_{i\neq j}Cov(e^{(i)}, e^{(j)})$, is 0.071 (correlation is 0.465). Meanwhile, the average covariance between a tree's prediction errors and the ground truth, $\mathbb{E}_i Cov(e^{(i)}, X)$, is −0.079 (correlation is −0.601). As a result, the average covariance between one tree's predictions and another tree's prediction errors, $\mathbb{E}_{i\neq j}Cov(e^{(i)}, \widehat{X}^{(j)})$, is only −0.008 (correlation is −0.061).

[8] Ignoring the exogenous control variables in LatentIV estimation leads to very large biases. We therefore take a conditional approach in the spirit of Frisch-Waugh-Lovell (Frisch and Waugh 1933): we separately regress the dependent variable and the endogenous covariate on the control variables, then obtain the residuals from these two regressions. Doing so "conditions out" the effects of control variables. We then apply the LatentIV approach based on the residuals.

[9] In Scornet et al. (2015, p. 1722), the authors state that "an easy adaptation of Theorem 1 shows that the CART algorithm is consistent under the same assumptions." This is because their proof of their theorem 1 is independent of, and therefore holds for any values of, forest-specific parameters (including number of trees, number of features attempted at each split, and the size of the subsample employed by each tree). Consequently, their result is also applicable to any given tree in the forest. Consistency of trees has also been shown in other work (Györfi et al. 2006, Biau et al. 2008) and discussed in Biau and Scornet (2016). We further note that tree consistency requires the number of training data points in a leaf node to go to infinity as $n \to \infty$, which is accommodated by Assumption 2. Instead, if the trees are fully grown (i.e., each leaf node contains only one training data point), then individual trees are inconsistent while the forest can remain consistent (Scornet et al. 2015, theorem 2).

[10] We acknowledge that, if the size of training data were significantly larger, performance of the sample splitting approach could potentially further improve. However, once again, we must reiterate that, in the presence of more training data, there is no need to "mine" covariates via machine-learning techniques in the first place.

[11] Relaxing this assumption makes the derivation more elaborate without changing the underlying mechanism. When $Cov(X_1, X_2) \neq 0$, our statement is still true under slightly more strict conditions

References

Aggarwal CC (2015) *Data Mining: The Textbook* (Springer, Berlin).
Aggarwal R, Gopal R, Gupta A, Singh H (2012) Putting money where the mouths are: The relation between venture financing and electronic word-of-mouth. *Inform. Systems Res.* 23(3-part-2): 976–992.
Angrist JD, Krueger AB (1995) Split-sample instrumental variables estimates of the return to schooling. *J. Bus. Econom. Statist.* 13(2): 225–235.
Angrist JD, Pischke JS (2008) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, Princeton, NJ).
Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. *ProPublica* May:23.
Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proc. National. Acad. Sci. USA* 113(27):7353–7360.
Athey S, Imbens GW (2017) The state of applied econometrics: Causality and policy evaluation. *J. Econom. Perspective* 31(2):3–32.
Belloni A, Chen D, Chernozhukov V, Hansen C (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6):2369–2429.
Berk R, Brown L, Buja A, Zhang K, Zhao L (2013) Valid post-selection inference. *Ann. Statist.* 41(2):802–837.
Bernard S, Adam S, Heutte L (2012) Dynamic random forests. *Pattern Recognition Lett.* 33(12):1580–1586.
Bernard S, Heutte L, Adam S (2010) A study of strength and correlation in random forests. *Proc. Internat. Conf. on Intelligent Comput.* (Springer, Berlin), 186–191.
Biau G, Scornet E (2016) A random forest guided tour. *TEST* 25(2): 197–227.
Biau G, Devroye L, Lugosi G (2008) Consistency of random forests and other averaging classifiers. *J. Machine Learn. Res.* 9(9).
Blackburn M, Neumark D (1992) Unobserved ability, efficiency wages, and interindustry wage differentials. *Quart. J. Econom.* 107(4):1421–1436.
Blaser R, Fryzlewicz P (2016) Random rotation ensembles. *J. Machine Learn. Res.* 17(1):126–151.
Blundell RW, Powell JL (2004) Endogeneity in semiparametric binary response models. *Rev. Econom. Stud.* 71(3):655–679.
Breiman L (1996) Bagging predictors. *Machine Learn.* 24(2):123–140.
Breiman L (2001) Random forests. *Machine Learn.* 45(1):5–32.
Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proc. Conf. on Fairness, Accountability and Transparency* (Association for Computing Machinery, New York), 77–91.
Buse A (1992) The bias of instrumental variable estimators. *Econometrica* 60(1):173–180.
Buzas JS, Stefanski LA (1996) Instrumental variable estimation in generalized linear measurement error models. *J. Amer. Statist. Assoc.* 91(435):999–1006.
Carroll RJ, Stefanski LA (1994) Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statist. Medicine* 13(12):1265–1282.
Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey WK (2017) Double/debiased/neyman machine learning of treatment effects. *Amer. Econom. Rev.* 107(5):261–265.
Conley TG, Hansen CB, Rossi PE (2012) Plausibly exogenous. *Rev. Econom. Statist.* 94(1):260–272.
Cook J, Stefanski L (1994) Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* 89(428):1314–1328.
Denisko D, Hoffman MM (2018) Classification and interaction in random forests. *Proc. National Acad. Sci. USA* 115(8):1690–1692.
Ebbes P, Wedel M, Böckenholt U (2009) Frugal iv alternatives to identify the parameter for an endogenous regressor. *J. Appl. Econometrics* 24(3):446–468.
Ebbes P, Wedel M, Böckenholt U, Steerneman T (2005) Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quant. Marketing Econom.* 3(4):365–392.
Ellis PD (2010) *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results* (Cambridge University Press, Cambridge, UK).
Fanaee-T H, Gama J (2014) Event labeling combining ensemble detectors and background knowledge. *Progress Artificial Intelligence* 2(2-3):113–127.
Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? *J. Machine Learn. Res.* 15(1):3133–3181.
Fong C, Tyler M (2021) Machine learning predictions as regression covariates. *Political Anal.* 29(4):467–484.
Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. *Proc. 13th Internat. Conf. Internat. Conf. Machine Learn.* (ACM, New York), 148–156.
Frisch R, Waugh FV (1933) Partial time regressions as compared with individual trends. *Econometrica* 1(4):387–401.
Gebru T, Krause J, Wang Y, Chen D, Deng J, Aiden EL, Fei-Fei L (2017) Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proc. National Acad. Sci. USA* 114(50): 13108–13113.
Ghose A, Ipeirotis PG (2010) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* 23(10):1498–1512.
Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.
Giot R, Cherrier R (2014) Predicting bikeshare system usage up to one day ahead. *Proc. IEEE Sympos. on Comput. Intelligence in Vehicles and Transportation Systems* (IEEE, New York), 22–29.
Goh KY, Heng CS, Lin Z (2013) Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Inform. Systems Res.* 24(1):88–107.
Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (MIT Press, Cambridge, MA).
Grace YY (2016) *Statistical Analysis with Measurement Error or Misclassification* (Springer, Berlin).
Greene WH (2003) *Econometric Analysis* (Pearson Education India).
Gu B, Konana P, Raghunathan R, Chen HM (2014) Research note-the allure of homophily in social media: Evidence from investor responses on virtual communities. *Inform. Systems Res.* 25(3): 604–617.
Gu B, Konana P, Rajagopalan B, Chen HWM (2007) Competition among virtual communities and user valuation: The case of investing-related communities. *Inform. Systems Res.* 18(1):68–85.
Gustafson P (2003) *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments* (CRC Press, Boca Raton, FL).
Györfi L, Kohler M, Krzyzak A, Walk H (2006) *A Distribution-Free Theory of Nonparametric Regression* (Springer Science & Business Media, New York).
Hausman JA (1978) Specification tests in econometrics. *Econometrica* 46(6):1251–1271.

Hausman J (2001) Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *J. Econom. Perspective* 15(4):57–67.

Hausman JA, Newey WK, Powell JL (1995) Nonlinear errors in variables estimation of some engel curves. *J. Econometrics* 65(1): 205–233.

Hu Y, Schennach SM (2008) Instrumental variable treatment of nonclassical measurement error models. *Econometrica* 76(1):195–216.

Jelveh Z, Kogut B, Naidu S (2015) Political language in economics. Working paper.

Küchenhoff H, Lederer W, Lesaffre E (2007) Asymptotic variance estimation for the misclassification SIMEX. *Comput. Statist. Data Anal.* 51(12):6197–6211.

Küchenhoff H, Mwalili SM, Lesaffre E (2006) A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics* 62(1):85–96.

Lee JD, Sun DL, Sun Y, Taylor JE (2016) Exact post-selection inference, with application to the lasso. *Ann. Statist.* 44(3):907–927.

Lewbel A (2019) Using instrumental variables to estimate models with mismeasured regressors. Working paper.

Liu Y, Chen R, Chen Y, Mei Q, Salib S (2012)" i loan because …" understanding motivations for pro-social lending. *Proc. 5th ACM Internat. Conf. on Web Search and Data Mining*, 503–512.

Loken E, Gelman A (2017) Measurement error and the replication crisis. *Science* 355(6325):584–585.

Lu Y, Jerath K, Singh PV (2013) The emergence of opinion leaders in a networked online community: A dyadic model with time dynamics and a heuristic for fast estimation. *Management Sci.* 59(8):1783–1799.

Mammen E, Rothe C, Schienle M (2016) Semiparametric estimation with generated covariates. *Econometric Theory* 32(5):1140–1177.

Mammen E, Rothe C, Schienle M (2012) Nonparametric regression with nonparametrically generated covariates. *Ann. Statist.* 40(2): 1132–1170.

McFowland III E, Somanchi S, Neill DB (2018) Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. Preprint, submitted March 24, https://arxiv.org/abs/1803.09159.

Meng L, Wu B, Zhan Z (2016) Linear regression with an estimated regressor: Applications to aggregate indicators of economic development. *Empirical Econom.* 50(2):299–316.

Moreno A, Terwiesch C (2014) Doing business with strangers: Reputation in online service marketplaces. *Inform. Systems Res.* 25(4): 865–886.

Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62: 22–31.

Murphy KM, Topel RH (1985) Estimation and inference in two-step econometric models. *J. Bus. Econom. Statist.* 20(1):88–97.

Murray MP (2006) Avoiding invalid instruments and coping with weak instruments. *J. Econom. Perspective* 20(4):111–132.

Nagar AL (1959) The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* 27(4):575–595.

Newey WK (1984) A method of moments interpretation of sequential estimators. *Econom. Lett.* 14(2-3):201–206.

Oxley L, McAleer M (1993) Econometric issues in macroeconomic models with generated regressors. *J. Econom. Survery* 7(1):1–40.

Pagan A (1984) Econometric issues in the analysis of regressions with generated regressors. *Internat. Econom. Rev.* 25(1):221–247.

Roodman D (2009) A note on the theme of too many instruments. *Oxf. Bull. Econom. Statist.* 71(1):135–158.

Ryu JY, Kim HU, Lee SY (2018) Deep learning improves prediction of drug–drug and drug–food interactions. *Proc. National Acad. Sci. USA* 115(18):E4304–E4311.

Schennach SM (2016) Recent advances in the measurement error literature. *Annu. Rev. Econom.* 8:341–377.

Scornet E, Biau G, Vert JP, et al (2015) Consistency of random forests. *Ann. Statist.* 43(4):1716–1741.

Seber GA (2009) *Multivariate Observations*, vol. 252 (John Wiley & Sons, Hoboken, NJ).

Singh PV, Sahoo N, Mukhopadhyay T (2014) How to attract and retain readers in enterprise blogging? *Inform. Systems Res.* 25(1): 35–52.

Sperlich S (2009) A note on non-parametric estimation with predicted variables. *Econom. J.* 12(2):382–395.

Taylor J, Tibshirani RJ (2015) Statistical learning and selective inference. *Proc. National Acad. Sci. USA* 112(25):7629–7634.

Tirunillai S, Tellis GJ (2012) Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Sci.* 31(2):198–215.

Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: A survey and results of new tests. *Pattern Recognition* 44(2):330–349.

Wang T, Kannan KN, Ulmer JR (2013) The association between the disclosure and the realization of information security risk factors. *Inform. Systems Res.* 24(2):201–218.

Wooldridge JM (2002) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).

Yang M, Adomavicius G, Burtch G, Ren Y (2018) Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Inform. Systems Res.* 29(1):4–24.

Zhu H, Kraut R, Kittur A (2012) Effectiveness of shared leadership in online communities. *Proc. ACM Conf. on Comput. Supported Cooperative Work* (Association for Computing Machinery, New York), 407–416.

Zhu H, Kraut RE, Wang YC, Kittur A (2011) Identifying shared leadership in wikipedia. *Proc. SIGCHI Conf. on Human Factors in Comput. Systems* (Association for Computing Machinery, New York), 3431–3434.