Internet Appendix for "Bots Synchronize Stock Returns"

Andreas Park* Jinhua Wang†

March 2, 2022

*University of Toronto andreas.park@rotman.utoronto.ca.
†University of Cambridge, Judge Business School, jw983@jbs.cam.ac.uk.

Appendix: Machine Learning Tools: Causal and Instrumental Random Forests

A. Critical Assumptions of the Causal Forest Methodology

We require two assumptions for causal forests to estimate unbiased treatment effects: conditional unconfoundedness assumption and overlap assumption (Wager and Athey 2018). The first requirement that we need to make for the methodology to hold is the *conditional unconfoundedness assumption* which formally says, for the propensity score $e(X_i)$, that

$$\Big\{Y_i|W_i=0, Y_i|W_i=1\Big\} \perp\!\!\!\perp W_i|e(X_i). \tag{1}$$

Under the unconfoundedness assumption, the securities are randomly assigned as controls based on a potentially high-dimensional form of covariates.

The second requirement that we need to make for the methodology to hold is the *overlap assumption* which formally says, for $\varepsilon > 0$, that

$$\varepsilon < P[W=1|X=x] < 1-\varepsilon. \tag{2}$$

The overlap assumption essentially ensures that, for every security in the treatment group, there are enough securities in the control group with similar characteristics to match on, and vice versa. Intuitively, if two securities, one in the treatment group and one in the control group, are classified into the same leaf of the forest frequently, the the causal forest will use the security in the treatment group as the counterfactual for the security in the control group, and vice versa. Taking a step back from the formalism, in the traditional matching approach that we used above, there was no shortage of candidate matches for the securities that go in and out of the S&P500. The mechanism of the causal forest makes further corrections for possible concerns. For instance, poor overlap could occur when the data is imbalanced. The causal forest would then not be able to properly extrapolate the counterfactual set of of securities for the control group, as there are not enough securities in the treatment group that are similar to the securities in the control group. When there are more securities in the treatment group than the securities in the control group, the available correction is to use an average treatment effect for the treated securities and an average treatment effect for the overlap-weighted securities to

mitigate the bias in the average treatment effects.

Under the conditional unconfoundedness and overlap assumptions, let X_i be a vector of characteristics of security i. The conditional average treatment effect τ can then be written as:

$$\tau = E(Y_i|X_i, W = 1) - E(Y_i|X_i, W = 0). \tag{3}$$

A key advantage of causal forests relative to causal trees is its resistance to overfitting due to aggregating the estimations among a large number of trees. To obtain more robust inference on the confidence intervals of our treatment effects, we planted 3000 trees in our causal forests based on the universe of all stocks traded in the market as the set of possible matches.

B. Challenges in using causal forests

Until recently, the random forest model has not seen much use in economic analysis. Aside from it not being part of the canon of graduate econometrics, there were three shortcomings. First, the traditional random forests approach is based on adaptive learning, which means that the same sample is used for both training and estimation. This approach can introduce biases because the algorithm may simply provide an estimation that fits the best in the current dataset but that does not generalize and that is therefore of little use for inference. The second concern is that there is a lack of statistical inference in the traditional random forests algorithm. Specifically, in our case we are interested in estimating a treatment effect. A standard random forest model can provide us with insights into the magnitude of the treatment effect but it actually does not allow us to assess whether the effect is significant. The third concern is the bias due to regularization in Machine Learning algorithms.

Athey and Imbens (2015) proposed a methodology called Honest Estimation in Causal Trees that addresses the first concern. In addition to the training, cross-validation, and hold-out samples, Honest Estimation further splits out an estimation sample. By separating the training samples and test samples from the estimation sample, a causal tree model essentially constructs the model in one sample, and then performs the estimation in another sample so as to provide unbiased estimates. A causal forest is essentially an ensemble of these causal trees, where the training sets, cross-validation sets, and estimation sets are randomly shuffled and re-sampled in every tree in the forest. The shuffling and re-sampling alleviate the shrinking training dataset problem by adding the estimation set.

Wager (2014) addresses the second concern and identifies the asymptotic normality of random forests so that we can estimate the variance of the prediction results using the Infinitesimal Jackknife approach. Wager and Athey (2018) further show that the heterogeneous treatment effects estimated with causal forest is consistent and asymptotically both Gaussian and centered under the so-called "unconfoundedness" and "overlap" assumptions. Our estimation approach makes use of all of these recent insights.

To correct for the bias in estimators introduced by regularization, the third concern, we use Robinson (1988)'s residuals-on-residuals estimation approach, which is also referred to as called Double/Debiased Machine Learning.[1]

We finally emphasize that this procedure is computationally intensive and requires the use of high-performance computers.[2]

C. Heterogeneous Treatment Effects

Standard econometric models, such as equation (**??**), asserts a strong assumption that treatment effects are constant across securities. However, this is often an over-simplification of the real-world, the impacts of bot trading on R-squared should be different for securities with different characteristics.

We relax the economic assumption asserted in section **??** that the treatment effect is constant across all securities. Instead, we allow the treatment effect of bot trading on the R^2 to be heterogeneous conditional on the characteristics X_i of securities. Let $\mu(X_i)$ be the intercept, $\tau(X_i)$ be the heterogeneous treatment effect on R^2 and W_i be the treatment variable. Then the semi-parametric equation we are estimating with causal forest can be written as:

$$Y_i = \mu(X_i) + \tau(X_i)W_i + \epsilon_i. \tag{4}$$

Parameter $\tau(X_i)$ is the non-parametric heterogeneous treatment effect that the causal forest method estimates with moment matching conditions.

Causal Forests Heterogeneous Treatment Effects Estimators and Mean-squared Error Estimators. Let $b \in [1, B]$ be the b-th tree in the causal forest, and let $L_b(X_i)$ be the leaf of the b-th tree in which the test point X_i falls in. Let $\alpha_j(X_i)$ be a measure of how often the

[1] See Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

[2] We are most grateful to Cambridge HPC and Compute Canada to let us allow to use their supercomputers.

j-th training example x_j falls in the same leaf as the i-th test point X_i:

$$\alpha_j(X_i) = \frac{1}{B}\sum_{b=1}^{B}\frac{1(\{x_j \in L_b(X_i), j \in S_b\})}{|\{j : x_j \in L_b(X_i), j \in S_b\}|}. \tag{5}$$

Let S_b be the sub-sample on which tree b is grown. Subscript $-j$ denotes the so-called out-of-bag estimators which only use the trees b for which $j \notin S_b$. The propensity score of treatment is denoted as $e(X_i) = P[W_i|X_i]$ and the expected outcome of marginalizing over the treatment is denoted as $m(x) = E[Y_i|X_i]$. The heterogeneous treatment effect $\hat{\tau}(X_i)$ is an orthogonalized[3] and double/debiased estimator[4]

$$\hat{\tau}(X_i) = \frac{\sum_{j=1}^{n}\alpha_j(X_i)(y_j - \hat{m}^{-j}(x_j))(W_j - \hat{e}^{-j}(x_j))}{\sum_{j=1}^{n}\alpha_j(X_i)(W_j - \hat{e}^{-j}(x_j))^2}. \tag{6}$$

Another challenge when using random forests is the estimation of a treatment effect because it is difficult to directly measure the in-sample goodness of fit because the "ground truth" treatment effect is not observed. Let S^{tr} represent the training sample, S^{te} the test sample, and N^{tr} the number of training samples. We follow Athey and Imbens (2016) and Wager and Athey (2018) who show that the in-sample mean-squared-error can be unbiasedly estimated as:

$$-\hat{\text{MSE}}_\tau(S^{tr}, S^{tr}, \Pi) = \frac{1}{N^{tr}}\sum_{i \in S^{tr}}\hat{\tau}^2(X_i; S^{tr}, \Pi). \tag{7}$$

Minimising the in-sample mean-squared error is equivalent to maximising the variance of $\hat{\tau}(X_i)$ in the training set. The cross-validation mean-squared-error can also be unbiasedly estimated as:

$$\hat{MSE}_\tau(S^{te}, S^{tr}, \Pi) \equiv -\frac{2}{N^{tr}}\sum_{i \in S^{te}}\hat{\tau}(X_i; S^{te}, \Pi) \times \hat{\tau}(X_i; S^{tr}, \Pi) + \frac{1}{N^{tr}}\sum_{i \in S^{te}}\hat{\tau}^2(X_i; S^{tr}, \Pi). \tag{8}$$

[3] See Athey and Wager (2019).

[4] See Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

D. *Average Treatment Effect*

Although we allow the treatment effect $\tau(X_i)$ to be heterogeneous across securities conditional on X_i, it would still be important to understand, on average, how does bot trading affects R^2. Therefore, we estimate three τ estimators of average treatment effects.

When the treatment W_i is the event indicator, which is binary a variable, we provide three doubly-robust augmented inverse-propensity weighted average treatment effect estimators:[5] (1) the average treatment effect for all securities, (2) the average treatment effect for the treated securities, and (3) the average treatment effect of the overlap-weighted securities (GRF-Labs 2019).

The average treatment effect for all securities is the average treatment effect τ_{all} over *all* securities.

$$\tau_{all} = \sum_{i=1}^{n} \frac{E[Y(1) - Y(0)|X = X_i]}{n}. \tag{9}$$

The average treatment effect for the treated securities ($\tau_{treated}$) is the average treatment effect conditional on the *treated* population.

$$\tau_{treated} = \sum_{W_i=1} \frac{E[Y(1) - Y(0)|X_i]}{|i : W_i = 1|}. \tag{10}$$

The average treatment effect for the overlap-weighted securities is computed for treatment propensity scores that are very close to 0 or 1.[6]

$$\tau_{overlap} = \sum_{i=1}^{n} \frac{e(X_i)(1 - e(X_i))E[Y(1) - Y(0)|X_i]}{\sum_{i=1}^{n} e(X_i)(1 - e(X_i))}, \text{ where } e(X_i) = P[W_i = 1|X_i]. \tag{11}$$

The estimators we used for equations above can be found in Appendix D.. The average treatment effect for the treated securities is our main estimator of interest, as our data has fewer treated securities than control securities, and the average treatment effect for the treated estimator only needs to recover the counterfactual for the treated securities. The average treatment for all securities estimator, in contrast, also needs to recover the counterfactual for the control population, which might have a poor match or overlap in the treated population

[5] See Robins, Rotnitzky, and Zhao (1994).

[6] See Li, Morgan, and Zaslavsky (2018).

and could lead to a violation of our overlap assumption. We therefore, focus on the average treatment effect for the treated securities.

When the treatment W_i is an bot trading proxy (which is a continuous variable), we provide two average treatment effect estimators: the average treatment effect for the overlap-weighted securities, and the average partial treatment effect. The average treatment effect for the overlap-weighted securities estimator is the same as described in equation (14), as the estimator for binary treatment generalizes to the continuous treatment case. The average partial effect for continuous treatment is estimated as[7]

$$\tau_{partial} = \frac{1}{n}\sum_{i=1}^{n}\frac{Cov[W_i, Y_i|X_i]}{Var[W_i|X_i]}. \tag{12}$$

Causal Forests Average Treatment Effects Estimators. Because we do not observe the counterfactual in equation (9), (10), and (11), we use the causal forest to construct estimators that are asymptotically unbiased. The average treatment effect estimator is the average of the doubly-robust corrected out-of-bag prediction of treatment effects.[8] Let $\hat{\tau}$ be the average treatment effect estimator and $\hat{\sigma}$ be the standard error of the treatment effect estimate. To take into consideration that stocks included/excluded from an index in the same event might be similar in characteristics, we cluster our standard error by index-change event. Let $E_i \in \{1, 2, ..., K-1, K\}$ denote the event where stock i was excluded from an index or the event when stock i was included into an index. Let n_k denote the total number of securities in our data during event K. Athey and Wager (2019) show that:

$$\begin{aligned}
&\hat{\tau} = \frac{1}{K}\sum_{k=1}^{K}\hat{\tau}_k, \text{where } \hat{\tau}_k = \frac{1}{n_k}\sum_{i:E_i=k}\hat{\Gamma}_i \text{ and } \hat{\sigma}^2 = \frac{1}{K(K-1)}\sum_{k=1}^{K}(\hat{\tau}_i - \hat{\tau})^2, \\
&\hat{\Gamma}_i = \hat{\tau}^{-i}(X_i) + \frac{W_i - \hat{e}^{-i}(X_i)}{\hat{e}^{-i}(X_i)(1-\hat{e}^{-i}(X_i))}(Y_i - \hat{m}^{-i}(X_i) - (W_i - \hat{e}^{(-i)}(X_i))\hat{\tau}^{-i}(X_i)).
\end{aligned} \tag{13}$$

The average treatment effect of the overlap-weighted securities is[9]

$$\hat{\tau} = \frac{\sum(Wi - e(Xi))(Yi - m(Xi))}{\sum(Wi - e(Xi))^2}. \tag{14}$$

[7] See Wager (2019a).
[8] See Athey and Wager (2019).
[9] See Wager (2019d).

E. Best-Pruned Tree in Causal Forests

An feature of traditional econometrics is that the estimation results are usually easy to interpret and that the procedure that delivers the results is easy to follow. The same often does not hold true for machine learning tools such as the random forest that we use here. There are thousands of trees in random forests and the tree structures are unstable. Therefore, it is infeasible, although important, to understand every tree in the forest. One work-around to provide some intuition is the the *best-pruned tree representation* (Wager (2019c)), which we use here to show the average treatment effects. We outline the the best-pruned tree algorithm below in Algorithm 1.

The general idea is that because causal forests use many trees to avoid overfitting, the large number of trees in the forest obfuscates the decision rules of the individual trees. The best-pruned tree algorithm searches for a representative tree in the forest, which has the minimum cross-validated loss after pruning in the forest, where we use pruning to reduce the risk of overfitting. The output is a representative graph of decision rules made by causal forests (Wang and Weeks 2019).

F. Causal Random Forest Estimation of the Change in R^2 and bot trading

We construct a difference-in-difference estimator by employing the above described "causal forest method." Our task is to assess the effect of the treatment on the the dependent variables DV_i, the R^2 and the bot trading measures. In contrast to the panel estimation of the preceding section, here the dependent variables are computed as the difference of the 20-day average before to after the event.

We use both binary and continuous treatment variables. When we estimate the causal impact on R^2, for 5-second, 5-minute and daily: $\Delta(R^2$ 5sec), $\Delta(R^2$ 5min), and $\Delta(R^2$ daily). When we estimate the causal impact on bot trading proxies, we use one of the following variables as the dependent variable: Δ(Number of Quotes) or Δ(Fragmentation). As co-variates we use the difference of the 20-day average before to after the event of the following variables: Δ(Price), Δ(Market Cap), Δ(\$-volume), Δ(p-imp), Δ(VIX), Δ(qspread cents), and Δ(qspread bps). As binary treatment variables we use dummy variables that indicate if a security was included or excluded from S&P500 index on a particular day. As continuous treatment variables we use the difference of the 20-day average before to after the event of the

Algorithm 1
Best-Pruned Tree (Causal Forests)

1. Train a causal forest F with N trees. Denote the $i_{th.}$ tree in the forest as f_i. Denote node m of the i_{th} tree by L_{im}.

2. For each node L_{im}, define the following:

   (a) The **raw** loss estimated using the cross-validated error on the honest estimation set is R_{im}. It is calculated as the loss based on the R-Learning criteria (Nie and Wager 2017):

   $$R_{im} = \sum_{i=1}^{N} \left(Y_i - \hat{m}(X_i) - \Big(W_i - \hat{e}(X_i)\Big) \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}(X_i) \right)^2. \tag{15}$$

   (b) Define P_{im} as the **pruned** loss.

   (c) Define L_{im}^{left} as the left child node and L_{im}^{right} as the right child node.

3. For each tree i in forest F:

   (a) Denote the root node pruned loss as R_{i1}.

   (b) Recursively calculate P_{im} for each node as:

      i. if node L_{im} is a leaf, then $P_{im} = R_{im}$.

      ii. if node L_{im} is not a leaf, then $P_{im} = minimum(R_{im}, c + P_{im}^{left} + P_{im}^{right})$, where $c > 0$.

4. Find b so that $R_{bi} = min(R_{i1})$, for $\forall i \in (1, N)$.

5. Tree f_b is the *best-pruned tree*.

bot trading proxies: Δ(Number of Quotes) or Δ(Fragmentation). The result are discussed in the main text.

G. Instrumental Forests

Similar to causal forests, instrumental forests estimate the heterogeneous treatment effect under the potential outcome framework. Suppose Z is the instrument, W is the treatment variable, Y is the dependent variable and X_i are the covariates, we assume the following structural model:

$$Y_i = \mu(X_i) + \tau(X_i)W_i + \varepsilon_i. \tag{16}$$

$\tau(X_i)$ is our non-parametric estimator of the heterogeneous treatment effects. In addition to the conditional unconfoundedness and overlap assumptions, we also require the following two moment conditions for instrumental forests to estimate unbiased treatment effects:

$$E\Big[Y_i - W_i\tau(X_i) - \mu(X_i)|X_i\Big] = 0, \tag{17}$$

$$E\Big[Z_i\Big(Y_i - W_i\tau(X_i) - \mu(X_i)\Big)|X_i\Big] = 0. \tag{18}$$

That is, we require our instruments Z to influence the outcome variable Y only through the treatment W. As discussed in part **??** of section **??**, the assumptions underlying instrumental forest is that an index inclusion or exclusion does not affect the relation of a stock's return with the market directly and that instead such an event affects algorithmic/high frequency trading directly. We use the index inclusion and exclusion events as the instruments, the change in algorithmic proxies as the instrumented variable, and the change R^2 as the dependent variable.

Local Average Treatment Effect (LATE) We use index inclusion or index exclusion as the instrumental variables and bot trading proxies as treatment variables (instrumented variables). Wager (2019b) proposed doubly robust estimator of the local average treatment effect (LATE) $\tau_{LATE} = E[\tau(X_i)]$, where

$$\tau(X_i) = \frac{Cov[Y, Z|X_i]}{Cov[W, Z|X_i]}. \tag{19}$$

Our LATE estimator is a generalization of the Inverse Compliance Score Weighted (ICSW) average treatment estimator in Aronow and Carnegie (2013) for the case of a binary instrument

and continuous treatments.

Instrumental Forest Estimation of the Change in R^2 and bot trading. As the instrument we use a dummy variable that indicates if a security was included or excluded from S&P500 index on a particular day. As treatment (instrumented) variables we use the difference of the 20-day average before to after the event of the bot trading proxies: $\Delta(\text{Number of Quotes})$ or $\Delta(\text{Fragmentation})$. As co-variates we use the difference of the 20-day average before to after the event of the following variables: $\Delta(\text{Price})$, $\Delta(\text{Market Cap})$, $\Delta(\$\text{-volume})$, $\Delta(\text{p-imp})$, $\Delta(\text{VIX})$, $\Delta(\text{qspread cents})$, and $\Delta(\text{qspread bps})$. We outline the the best-pruned tree algorithm, based on Wang and Weeks (2019), for an Instrumental Forest in below in Algorithm 2.

Algorithm 2
Best-Pruned Tree (Instrumental Forests)

---

1. Train a instrumental forest F with N trees. Denote the $i_{th.}$ tree in the forest as f_i. Denote node m of the i_{th} tree by L_{im}.

2. For each node L_{im}, define the following:

   (a) The **raw** loss estimated using the cross-validated error on the honest estimation set is R_{im}. We derive the loss function from the loss based on the R-Learning criteria (Nie and Wager 2017):

   $$R_{im} = \sum_{i=1}^{N} \left[\left(Z_i - \hat{Z}(X_i) \right) \times \left(Y_i - \hat{m}(X_i) - \left(W_i - \hat{e}(X_i) \right) \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}(X_i) \right) \right]^2 . \quad (20)$$

   (b) Define P_{im} as the **pruned** loss.

   (c) Define L_{im}^{left} as the left child node and L_{im}^{right} as the right child node.

3. For each tree i in forest F:

   (a) Denote the root node pruned loss as R_{i1}.

   (b) Recursively calculate P_{im} for each node as:

      i. if node L_{im} is a leaf, then $P_{im} = R_{im}$.

      ii. if node L_{im} is not a leaf, then $P_{im} = minimum(R_{im}, c + P_{im}^{left} + P_{im}^{right})$, where $c > 0$.

4. Find b so that $R_{bi} = min(R_{i1})$, for $\forall i \in (1, N)$.

5. Tree f_b is the *best-pruned tree*.

REFERENCES

Aronow, Peter M, and Allison Carnegie, 2013, Beyond late: Estimation of the average treatment effect with an instrumental variable, *Political Analysis* 21, 492–506.

Athey, Susan, and Guido Imbens, 2015, Recursive partitioning for heterogeneous causal effects, Working paper Stanford University.

——— , 2016, Recursive partitioning for heterogeneous causal effects, *Proceedings of the National Academy of Sciences* 113, 7353–7360.

Athey, Susan, and Stefan Wager, 2019, Estimating treatment effects with causal forests: An application, *arXiv preprint arXiv:1902.07409*.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, 2018, Double/debiased machine learning for treatment and structural parameters, .

GRF-Labs, 2019, The grf algorithm, `https://github.com/grf-labs/grf/blob/master/REFERENCE.md`.

Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky, 2018, Balancing covariates via propensity score weighting, *Journal of the American Statistical Association* 113, 390–400.

Nie, Xinkun, and Stefan Wager, 2017, Quasi-oracle estimation of heterogeneous treatment effects, *arXiv preprint arXiv:1712.04912*.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao, 1994, Estimation of regression coefficients when some regressors are not always observed, *Journal of the American statistical Association* 89, 846–866.

Robinson, Peter M, 1988, Root-n-consistent semiparametric regression, *Econometrica: Journal of the Econometric Society* pp. 931–954.

Wager, Stefan, 2014, Asymptotic theory for random forests, Working paper Stanford University.

——— , 2019a, Add function for estimating average partial effects, `https://github.com/grf-labs/grf/pull/175`.

——— , 2019b, Estimate average effects using instrumental forests, `https://github.com/grf-labs/grf/pull/490`.

——— , 2019c, Find the best tree in the random forest, https://github.com/grf-labs/grf/issues/281.

——— , 2019d, "overlap" in average treatment effects, https://github.com/grf-labs/grf/issues/375.

——— , and Susan Athey, 2018, Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association* 113, 1228–1242.

Wang, Jinhua, and Melvyn Weeks, 2019, Policy prediction in job training programs, *Working Paper*.