Inference with Multi-Outcome Generalized Random Forest

Maria Nareklishvili*
Graduate School of Business, Stanford University

May 21, 2025

Abstract

Wald test statistic produces elliptical confidence regions for multi-dimensional parameters. When parameters, such as treatment effects, are correlated, failing to account for these dependencies can lead to misspecified confidence regions, and increased Type 1 and Type 2 error rates. This paper proposes the multi-outcome generalized random forest, an extension of the generalized random forest that enables inference on correlated treatment effects. The method is designed for settings with multiple correlated outcomes or treatments and explicitly incorporates these dependencies into the estimation process. Simulation results and empirical analysis show that modeling these correlations increases the statistical power of joint hypothesis tests both within and across subsets of covariates.

Keywords: generalized random forest, inference, joint hypothesis testing, multiple outcomes

*This paper benefited from constructive feedback of Susan Athey, Dmitry Arkhangelsky, Stéphane Bonhomme, Alexander Torgovitsky, Thomas Wiemann, participants at the California Econometrics Conference, seminar participants at Monash University, the University of Melbourne, the University of Glasgow, and the Federal Reserve Bank of Philadelphia, as well as the Computational Workshop at the University of Chicago. Financial support from the Research Council of Norway (project no. 280350/GE) is gratefully acknowledged. There are no competing interests to declare. Replication code is available at https://github.com/marnare/inferenceMGRF.

1 Introduction

A central challenge in estimating treatment effects across multiple outcomes is the correlation among those effects. Ignoring these dependencies can invalidate statistical inference, inflating both Type I and Type II error rates. This paper investigates how misspecification of the covariance structure in covariate-dependent treatment effects influences inference, and extends the generalized random forest framework to account for such correlations.

Policymakers frequently evaluate programs that influence multiple, potentially dependent outcomes. For example, an educational intervention may affect performance in mathematics, reading, and science, with improvements in one domain carrying over to others. In these settings, individual-level heterogeneity is often quantified by vector-valued parameters θ_i, estimators $\hat{\theta}_i$, and covariance matrices Σ_i for each individual i. When Σ_i is misspecified—for instance, by assuming independence of correlated effects—the joint confidence regions, typically represented as Wald ellipsoids, may be substantially distorted, affecting both their shape and size. These distortions worsen with the dimensionality of outcomes, as the determinant of Σ_i may differ significantly from that of a diagonal approximation, ultimately inflating Type I or Type II error rates (Gleser et al., 2009; Becker, 2000; Gleser and Olkin, 2000; Riley, 2009; Kim and Becker, 2010; Van den Noortgate et al., 2015).[1]

Athey and Wager (2018) and Athey et al. (2019) extend the classical random forest framework (Breiman, 2001) to develop the generalized random forest algorithm, enabling estimation and inference on parameters of interest, including heterogeneous treatment effects. Building on this framework, we demonstrate that the generalized random forest es-

[1] Becker (2000) concludes that "No reviewer should ever ignore dependence among study outcomes. Even the simplest ad hoc options are better than pretending such dependence does not exist."

timator can be expressed as an orthogonal projection onto a space of additive components, with an asymptotically vanishing residual term. This result holds even in the presence of multiple correlated treatment effects.

Inference on correlated treatment effects relies on several key assumptions. First, honesty requires using separate subsamples for partitioning and treatment effect estimation. Second, the Lipschitz continuity of the outcome moments ensures smooth variations in outcomes with respect to covariates. Third, random split trees maintain the randomness of partitions rather than purely optimizing split quality. Fourth, $\alpha - k$ regularity condition guarantees sufficient observations in each covariate region, preventing degenerate partitions. Finally, overlap ensures that all treatments are observed across regions, avoiding near-exclusion of any treatment group.

Simulations and an empirical application demonstrate that failing to account for strong correlations among treatment effects can elevate Type II error to 52% aand Type I error to 13%. As the number of outcomes exceeds ten, Type II error nears 100%, highlighting the vulnerability of joint confidence intervals to covariance misspecification.

2 Related Literature

Early theoretical research on random forests established consistency under simplified versions of the algorithm and stylized assumptions (Breiman, 2001, 2004; Jeon and Lin, 2006; Meinshausen and Ridgeway, 2006; Biau, 2012; Denil et al., 2014; Wager, 2014; Scornet et al., 2015). Subsequent extensions target causal inference, showing asymptotic normality for univariate treatment effects (Athey and Imbens, 2016; Athey and Wager, 2018; Korolyuk and Borovskich, 2013), and further adapt the random forest estimator to accommodate correlated parameters (Athey et al., 2019; Nekipelov et al., 2018; Li, 2020; Bühlmann et al., 2020; Wang et al., 2022).

Our work advances this line of literature by enabling joint estimation and inference on multiple treatment effects within a generalized random forest framework. Building on the local method of moments, the multi-outcome generalized random forest estimator minimizes the squared deviation between individual treatment effects and their local averages. It explicitly accounts for treatment effect covariance and improves inference in settings where effects cannot be assumed independent.

3 The Effect of Covariance on Confidence Regions

Suppose a researcher is studying the effect of an educational intervention on student performance across multiple subjects. In practice, a true relationship between the intervention effects across fields is unknown. However, if prior evidence suggests strong correlations between improvements in mathematics and physics scores, the researcher may prefer a structured covariance matrix Σ_i that accounts for this relationship rather than assuming an identity matrix $\Sigma_{i,0} = \sigma_i^2 I_d$, which treats all subject scores as independent. Consider a setting where each individual i has a parameter vector θ_i representing their specific effect. A corresponding estimator, $\hat{\theta}_i$, varies across individuals (such as generalized random forest, hierarchical or mixed effects models). Each individual estimate $\hat{\theta}_i$ has an associated covariance matrix Σ_i, leading to the standard Wald ellipsoid:

$$(\hat{\theta}_i - \theta_i)^\top \Sigma_i^{-1} (\hat{\theta}_i - \theta_i) \leq \chi^2_{d,1-\alpha}. \tag{1}$$

In contrast, the population-level parameter θ summarizes effects across individuals. A standard estimator $\hat{\theta}$ with covariance matrix Σ satisfies the quadratic form:

$$(\hat{\theta}-\theta)^\top \Sigma^{-1}(\hat{\theta}-\theta) \leq \chi^2_{d,1-\alpha}, \tag{2}$$

which defines a $(1-\alpha)$ confidence region under standard regularity conditions [2].

Here, $\chi^2_{d,\,1-\alpha}$ denotes the $(1-\alpha)$ quantile of the chi-square distribution with d degrees of freedom. The exact volume of this d-dimensional confidence ellipsoid is given by

$$V_{\text{cov}} := V_{\text{ellipsoid}} = \underbrace{\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}}_{\text{base unit circle volume}} \times \underbrace{(\chi^2_{d,1-\alpha})^{d/2}}_{\text{scaling factor}} \times \underbrace{\sqrt{\det(\Sigma_i)}}_{\text{shape transformation}},$$

where $\Gamma(\cdot)$ is the Gamma function, and $\chi^2_{d,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2_d distribution. The volume consists of three components: the volume of the unit d-dimensional sphere, a scaling factor determined by the quantile of the chi-square distribution, and a shape transformation defined by Σ_i. The chi-square quantile defines the squared Mahalanobis radius of the ellipsoid and scales the unit sphere. The covariance matrix Σ_i determines the shape of the sphere and transforms the sphere into an ellipsoid. Misspecification of Σ_i distorts the ellipsoid's shape, potentially inflating Type I or Type II errors. Specifically, if the off-diagonal elements of Σ_i are ignored and it is instead approximated by $\Sigma_{i,0} = \sigma_i^2 I_d$, the confidence region simplifies to a sphere of radius

$$R = \sqrt{\sigma_i^2\,\chi^2_{d,\,1-\alpha}},$$

with volume

$$V_{\text{nocov}} := V_{\text{sphere}} = \frac{\pi^{d/2}}{\Gamma\big(\frac{d}{2}+1\big)} R^d = \frac{\pi^{d/2}}{\Gamma\big(\frac{d}{2}+1\big)} (\sigma_i^2 \chi^2_{d,\,1-\alpha})^{d/2}.$$

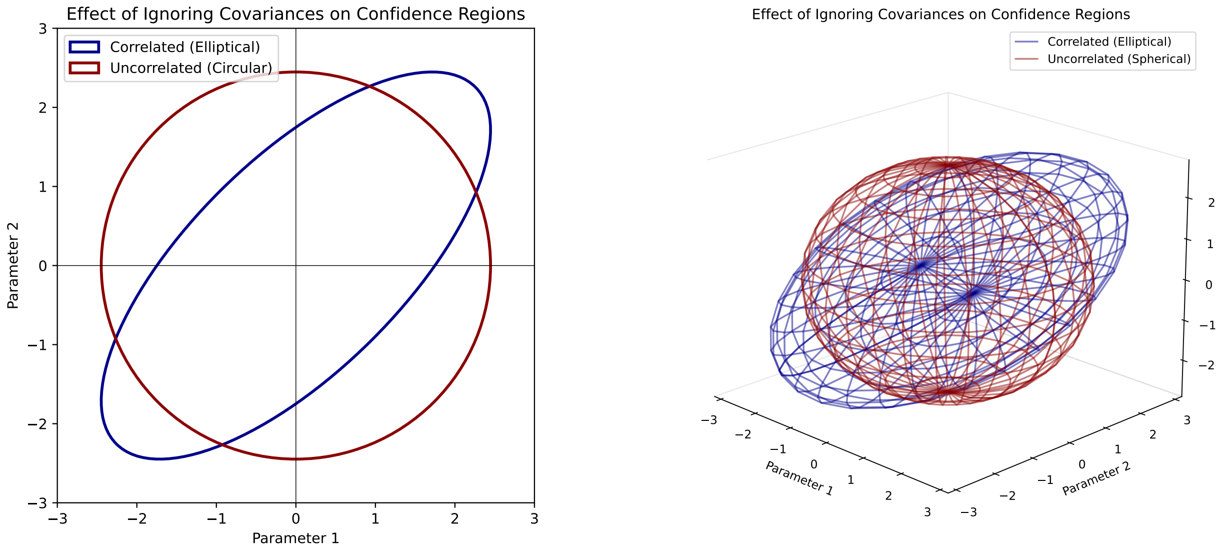
The ratio of the exact ellipsoidal volume to this naive spherical volume is

$$\frac{V_{\text{ellipsoid}}}{V_{\text{sphere}}} = \frac{\sqrt{\det(\Sigma_i)}}{\sqrt{\det(\Sigma_{i,0})}} = \frac{\sqrt{\det(\Sigma_i)}}{(\sigma_i^2)^{d/2}}.$$

[2] We investigate both but focus on personalized treatment effects.

In higher dimensions, the determinant $\det(\Sigma_i)$ can diverge significantly from $(\sigma_i^2)^{d/2}$, leading to substantial deviations in coverage. As the dimension d increases, these volume discrepancies amplify rapidly, potentially resulting in test statistics with false-positive or false-negative rates that deviate considerably from the nominal significance level α.

If the confidence region is too large (i.e., an over-expanded sphere), it includes more points than it should. This means we fail to reject more often, increasing Type II error (false negatives)—we fail to detect true deviations from the null. Figure 1 shows that the volume of the sphere excluding the intersection with the ellipse is the Type II error region. If the confidence region is too small, it excludes points that should be inside, leading to more rejections than necessary. This inflates Type I error (false positives)—we falsely detect deviations when they do not exist. Figure 1 shows that Type 1 error region is given by the volume of the ellipse excluding the intersection with the sphere. Both errors become more pronounced as dimensionality d increases, as mismatches between Σ_i and the naive $\sigma_i^2 I_d$ are magnified in a high-dimensional covariate space. Figure 1 illustrates the distortion in confidence regions when covariance is either accounted for or ignored, shown in both two- and three-dimensional coordinate systems.



(a) Confidence Regions in 2D

(b) Confidence Regions in 3D

Figure 1: Comparison of 2D and 3D confidence regions with and without parameter covariances.

In practice, we estimate volumes of elliptical (V_{cov}) and spherical (V_{nocov}) confidence regions using a high number of Monte Carlo simulations (100,000 here) to measure the error rates. Specifically, we generate N independent samples from a d-dimensional standard normal distribution:

$$\theta_i \sim \mathcal{N}(0, I_d), \quad i = 1, \ldots, N.$$

For now, assume Σ_i^{-1} is known, and non-diagonal terms are not zero. Then the misspecified covariance can be obtained by multiplying the variance-covariance matrix to the identity matrix $\Sigma_{i,0}^{-1} = \Sigma_i^{-1} \times I_d$. Then, for each sample, we compute the Mahalanobis distance under both covariance structures:

$$D_{\text{cov},i} = \theta_i^T \Sigma_i^{-1} \theta_i,$$

$$D_{\text{nocov},i} = \theta_i^T \Sigma_{i,0}^{-1} \theta_i.$$

A sample θ_i lies within the respective confidence region if:

$$D_{\text{cov},i} \leq \chi^2_{d,1-\alpha} \quad \text{(ellipsoid, accounting for covariance)}$$

$$D_{\text{nocov},i} \leq \chi^2_{d,1-\alpha} \quad \text{(sphere, assuming isotropic covariance).}$$

The intersection and exclusive volumes of the elliptical and spherical confidence regions are estimated by counting the proportion of Monte Carlo samples satisfying the respective inclusion criteria:

$$P_{\cap} = \frac{1}{N} \sum_{i=1}^{N} 1\{D_{\text{cov},i} \leq \chi^2_{d,1-\alpha} \text{ and } D_{\text{nocov},i} \leq \chi^2_{d,1-\alpha}\}$$

$$P_{\text{only cov}} = \frac{1}{N} \sum_{i=1}^{N} 1\{D_{\text{cov},i} \leq \chi^2_{d,1-\alpha}, D_{\text{nocov},i} > \chi^2_{d,1-\alpha}\}$$

$$P_{\text{only nocov}} = \frac{1}{N} \sum_{i=1}^{N} 1\{D_{\text{nocov},i} \leq \chi^2_{d,1-\alpha}, D_{\text{cov},i} > \chi^2_{d,1-\alpha}\}$$

The estimated volumes are then scaled to match the theoretical volume:

$$V_{\cap} = P_{\cap} \cdot \frac{V_{\text{cov}}}{P_{\text{cov}}}, \quad V_{\text{only cov}} = P_{\text{only cov}} \cdot \frac{V_{\text{cov}}}{P_{\text{cov}}}, \quad V_{\text{only nocov}} = P_{\text{only nocov}} \cdot \frac{V_{\text{cov}}}{P_{\text{cov}}}$$

Assuming elliptical confidence regions are correct (Table 1), the Type I and Type II error rates are given by:

$$\text{Type I Error} = \frac{P_{\text{only cov}}}{P_{\text{cov}}}, \tag{3}$$

$$\text{Type II Error} = \frac{P_{\text{only nocov}}}{P_{\text{nocov}}}. \tag{4}$$

Table 1: Type I and Type II errors when either the elliptical or spherical confidence regions are correct.

Perspective	only cov (in ellipse, out of sphere)	only nocov (in sphere, out of ellipse)
Ellipse is correct	Type I error of the sphere. *(The sphere rejects a point that should not be rejected.)*	Type II error of the sphere. *(The sphere accepts a point that should be rejected.)*
Sphere is correct	Type II error of the ellipse. *(The ellipse accepts a point that should be rejected.)*	Type I error of the ellipse. *(The ellipse rejects a point that should not be rejected.)*

The standard error of the Monte Carlo estimate is given by $1/\sqrt{N}$. This quantifies the sampling variability in the volume estimates and serves as a diagnostic for the accuracy of the Monte Carlo simulations.

3.1 High-Dimensional Inference and Misspecified Covariance Structures

With increasing number of outcomes or treatments, the discrepancy between a true covariance matrix Σ_i and a misspecified covariance matrix $\Sigma_{i,0}$ can lead to significant distortions in statistical inference. Consider a d-dimensional normal distribution $N(\mu_i, \Sigma_i)$. When Mahalanobis distances are computed using $\Sigma_{i,0}$ instead of Σ_i, they follow:

$$D_i^2 = (\hat{\theta}_i - \theta_i)^T \Sigma_{i,0}^{-1} (\hat{\theta}_i - \theta_i) \sim \chi_d^2 \cdot \lambda_i,$$

where λ_i is approximately the average eigenvalue of $\Sigma_{i,0}^{-1}\Sigma_i$. The number of covariance terms that are misspecified grows quadratically with d, as $\frac{d(d-1)}{2}$, causing λ_i to deviate from 1.

Proposition 1 (Expected Mahalanobis Distance)**.** *The expected Mahalanobis distances under the true and misspecified covariance matrices are given as:*

$$\mathbb{E}[D^2_{correct,i}] = d, \quad \mathbb{E}[D^2_{misspecified,i}] = d\lambda_i.$$

Proof in Appendix B.

The discrepancy in Proposition 1 scales as $O(d\lambda_i)$, becoming more pronounced as d increases. Moreover, for a random vector θ_i drawn from $N(0, I_d)$, the squared Euclidean norm $\|\theta_i\|_2^2$ follows a χ_d^2 distribution with mean d and variance $2d$. As d grows, $\|\theta_i\|_2^2$ becomes tightly concentrated around d, as described by the inequality:

$$P\left(\left|\frac{\|\theta_i\|_2}{\sqrt{d}} - 1\right| \geq \epsilon\right) \leq 2e^{-c\epsilon^2 d},$$

where $c > 0$ is a constant. This exponential concentration implies that high-dimensional random vectors lie near a spherical shell of radius $\sqrt{d}$, leaving little room for deviations.

When the covariance structure is misspecified, the resulting Mahalanobis distances systematically diverge from their true values. In high dimensions, this divergence causes confidence regions derived under the misspecified model to become nearly disjoint from those under the true model. As a result, the probability of failing to reject a false null hypothesis (Type II error) approaches 1.

4 Generalized Random forest for Multiple Outcomes

In this section, I introduce the generalized random forest estimator for correlated parameters and establish theoretical properties.

4.1 Setup and Identification Assumptions

Consider a dataset with N observations, where each observation i has a set of outcomes represented by the $N \times d$ matrix Y_i. The treatment variable, W_i, is binary ($W_i \in \{0, 1\}$) but may also extend to multiple dimensions. For example, W_i could indicate whether an individual participated in a job training program ($W_i = 1$) or not ($W_i = 0$), while the outcomes $Y_i \in \mathcal{Y} \in \mathbb{R}^d$ could include employment status, earnings, and job stability. In a multi-dimensional treatment setting, W_i could represent different levels or types of an intervention, such as access to different healthcare plans, where the outcomes Y_i could include medical expenses, health status, and frequency of doctor visits. An individual is also characterized by observable traits $X_i \in \mathcal{X} \in \mathbb{R}^p$. $D_i = (Y_i, X_i)$ collectively represents the data of observation i. The estimand for the conditional average treatment effect is defined as

$$\theta(x) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x], \tag{5}$$

where $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control, respectively. The unconditional average treatment effect (ATE) follows as

$$\mathbb{E}(\theta(x)) = \mathbb{E}\big[\mathbb{E}\big(Y_i(1) - Y_i(0) | X_i = x\big)\big] = \int_{\mathcal{X}} \mathbb{E}\big[Y_i(1) - Y_i(0) | X_i = x\big]\, f_X(x)\, \mathrm{d}x. \tag{6}$$

A key assumption for identifying (conditional) average treatment effects is unconfoundedness:

Assumption 1 (Unconfoundedness)**.** *The treatment assignment W_i is independent of potential outcomes, conditional on covariates:*

$$Y_i(1), Y_i(0) \perp W_i \mid X_i. \tag{7}$$

Assumption 1 (Rosenbaum and Rubin, 1983; Rubin, 1990) ensures that, after conditioning on X_i, treatment assignment is as good as random. Consequently, the observed outcomes of treated and control units can be used to estimate corresponding potential outcomes.

Assumption 2 (No Interference)**.** *For any two units i and j, the potential outcome for unit i does not depend on the treatment assigned to unit j. Formally, for any treatment assignment vectors w and w' such that $w_i = w'_i$, we have*

$$Y_i(w) = Y_i(w'),$$

where $Y_i(w)$ denotes the potential outcome of unit i when the treatment assignment for all units is given by w.

Assumption 2 states that a unit's outcome depends only on its own treatment status and is unaffected by the treatment assignments of other units.

4.2 A Single Outcome

Athey and Wager (2018) extend the standard random forest framework (Breiman, 2001) to treatment effect estimation. In a generalized random forest, the covariate space $\mathcal{X} \subset \mathbb{R}^p$ is recursively partitioned to estimate treatment effects within each resulting subset (or *leaf*). A split is defined by selecting a covariate j and a threshold $c \in \mathbb{R}$, dividing the current node $\mathcal{P}^{(t,m)}$ into two disjoint subsets:

$$\mathcal{P}^{(t+1,1)} = \{x \in \mathcal{P}^{(t,m)} \mid x_j \leq c\},$$
$$\mathcal{P}^{(t+1,2)} = \{x \in \mathcal{P}^{(t,m)} \mid x_j > c\}.$$

The optimal split is chosen to maximize the between-node variance in the estimated treatment effects, given by:

$$\arg\max_{(j,c)} \Big[\mathrm{Var}\big(\theta_i \mid X_i \in \mathcal{P}^{(t+1,1)}\big) + \mathrm{Var}\big(\theta_i \mid X_i \in \mathcal{P}^{(t+1,2)}\big) \Big], \tag{8}$$

where θ_i is the individual treatment effect. Since θ_i is unobserved, in practice, we estimate $\hat{\theta}_i$ as the average treatment effect within each subset and maximize its variance across nodes:

$$\hat{\theta}_i = \frac{1}{|\{i : W_i = 1,\, X_i \in \ell\}|} \sum_{\substack{i:\, W_i=1 \\ X_i \in \ell}} Y_i - \frac{1}{|\{i : W_i = 0,\, X_i \in \ell\}|} \sum_{\substack{i:\, W_i=0 \\ X_i \in \ell}} Y_i.$$

$|\{i : W_i = w, X_i \in \ell_n\}|$ is the number of treated ($w = 1$) or control ($w = 0$) units in leaf ℓ_n. Athey and Imbens (2016) demonstrate that maximizing the in-sample variance of treatment effects, as shown in (8), is equivalent to minimizing the mean squared error up to a constant. The mean squared error criterion penalizes splits that result in similar treatment effect estimates across subgroups. If a split fails to create distinct subgroups with heterogeneous treatment effects, the squared treatment effect estimates will be small, increasing the mean squared error. Therefore, minimizing the mean squared error encourages the algorithm to prioritize splits that lead to greater variation in treatment effects between subgroups. This procedure continues recursively, generating a hierarchy of nested subsets $\{\mathcal{P}^{(t,m)}\}$ at each depth t, until a stopping criterion (e.g., a minimum node size) is met. This algorithm is known as a generalized tree.

To construct a generalized random forest, this procedure is repeated across multiple subsamples. Given N observations $\{(X_i, W_i, Y_i)\}_{i=1}^{N}$, we consider all possible subsets of size $s < N$, typically chosen as $s = N^{\beta}$ for some $\beta < 1$. For each subset, a tree is grown using additional randomness ξ to allow for random covariate selection. The conditional average treatment effect (CATE) estimate in each leaf is given by:

$$\hat{\theta}(x, D_1, \ldots, D_N) \ = \ \frac{1}{\binom{N}{s}} \sum_{1 \le i_1 < \cdots < i_s \le N} \mathbb{E}_{\xi}\big[\theta(x, \xi, D_{i_1}, \ldots, D_{i_s})\big], \tag{9}$$

where $\theta(x, \Pi, D_1, \ldots, D_N)$ represents estimate in each sample. Π denotes a partition of the covariate space into disjoint subsets $\{\ell_1, \ldots, \ell_{|\Pi|}\}$. $\theta(x, \Pi, D_1, \ldots, D_N)$ denotes conditional average treatment effect for a covariate vector x, estimated by a single tree:

$$\theta(x, \xi, D_1, \ldots, D_N) = \sum_{n=1}^{|\Pi|} 1\{x \in \ell_n\} \hat{\tau}(\ell_n),$$

where $\hat{\tau}(\ell_n)$ represents the difference in mean outcomes between treated and control units within a leaf:

$$\hat{\tau}(\ell_n) = \frac{1}{|\{i : X_i \in \ell_n, W_i = 1\}|} \sum_{\substack{i: X_i \in \ell_n \\ W_i = 1}} Y_i - \frac{1}{|\{i : X_i \in \ell_n, W_i = 0\}|} \sum_{\substack{i: X_i \in \ell_n \\ W_i = 0}} Y_i.$$

The resulting tree partitions Π consist of disjoint terminal nodes of the covariate space, with $|\Pi|$ denoting their number:

$$\Pi = \{\ell_1, \ell_2, \ldots, \ell_{|\Pi|}\}, \quad \bigcup_{n=1}^{|\Pi|} \ell_n = \mathcal{X}.$$

The splitting process ensures that each element of $\mathcal{X}$ belongs to exactly one partition. In practice, the expectation in (9) is approximated by Monte Carlo simulations, drawing B subsamples without replacement and averaging over their corresponding trees:

$$\hat{\theta}(x) \approx \frac{1}{B} \sum_{b=1}^{B} \theta\big(x, \xi^{*(b)}, D_{i_1}^{*(b)}, \ldots, D_{i_s}^{*(b)}\big),$$

where $D_{i_1}^{*(b)}, \ldots, D_{i_s}^{*(b)}$ represent b-th data set sampled without replacement from $\{D_1, \ldots, D_N\}$.

4.3 Multiple Outcomes

We now extend the random forest estimator to the multi-outcome setting by applying a generalized method of moments (GMM) approach. Let each observation i have a vector-valued outcome Y_i and a treatment indicator W_i; denote by $D_i = (Y_i, W_i, X_i)$ the full data for observation i. We assume there is a vector-valued parameter θ^ℓ associated with each leaf (or partition) ℓ of the covariate space $\mathcal{X}$. The population conditional moment function is then written:

$$\mathbb{E}\big[\rho\big(D_i, \theta^\ell\big) \,\big|\, (X_i \in \ell)\big] \;=\; 0,$$

where $\rho(\cdot)$ is a GMM moment function.

Assumption 3 (Existence of the solution)**.** *For all $x \in \mathcal{X}$, the conditional expectation $\mathbb{E}[\rho(D_i, \theta^\ell) \mid X_i = x]$ is bounded, and its supremum norm* [3] *converges to zero as the sample size increases:*

$$\sup_{x \in \mathcal{X}} \big|\big|\mathbb{E}[\rho(D_i, \theta^\ell) \mid X_i = x]\big|\big| = o(1).$$

[3] The Frobenius norm can be used as norm for matrices and is defined as $||A||_F = \sqrt{\mathrm{tr}(AA^T)}$, where A^T denotes the transpose of A.

Assumption 3 ensures that within each partition, the estimation error remains controlled and diminishes as the sample size grows, preventing unbounded divergence of the estimated treatment effects from their population counterparts.

Assumption 4 (Invertible covariance)**.** *For each partition ℓ, there exists a weighting matrix $\Omega(X_i) \in \mathbb{R}^{p\times p}$ such that its eigenvalues are uniformly bounded by a constant λ, and the following matrix is strictly positive definite:*

$$\mathbb{E}\left[\Omega(X_i)\frac{\partial\rho(D_i,\theta^\ell)}{\partial\theta^\ell}\right]. \tag{10}$$

Assumption 4 ensures the relevant covariance/weighting matrix in the GMM formulation is well-conditioned, so leaf-level parameters θ^ℓ are identifiable.

Define the random forest estimator of treatment effects by $\hat{\theta}(X_i, S^{est}, \Pi)$, where $\Pi = \{\ell_1, \ldots, \ell_{|\Pi|}\}$ is a partition of the covariate space obtained using a training sample S^{tr}, and S^{est} is an independent estimation sample. Within each leaf ℓ, the estimator $\hat{\theta}^\ell$ is predicted using data $\{D_i : X_i \in \ell\} \subset S^{est}$. The objective is to minimize the expected quadratic deviation between the estimated and population treatment effects:

$$\mathbb{E}_{S^{tr},S^{est}}\left[(\theta_i - \hat{\theta}(X_i, S^{est}, \Pi))^T\, \Sigma_i^{-1}\, (\theta_i - \hat{\theta}(X_i, S^{est}, \Pi))\right], \tag{11}$$

where θ_i is the (true) vector of individual-level treatment effects, and $\Sigma_i \equiv \Sigma(X_i)$ is a (potentially) observation-specific covariance or weighting matrix. By weighting squared errors with the inverse covariance Σ_i^{-1}, this criterion adjusts for heteroskedasticity and correlation in treatment effect estimates during splitting. Σ_i may be set to the identity matrix during splitting. The random forest aggregates predictions across multiple trees, reducing the influence of any individual split. While misspecification of Σ_i does not affect partitioning asymptotically [4], it can substantially impact inference.

[4] See, for example, Ishak et al. (2008), who find that heterogeneous treatment effect estimates remain largely robust to inaccuracies in the covariance matrix approximation.

Proposition 2 (Method of moments estimator)**.** *Let Assumptions 3 and 4 hold. Then minimizing the mean squared error in* (11) *is equivalent to maximizing the variance of the vector-valued treatment effects up to a constant:*

$$\hat{\theta}(x, S^{est}, \Pi) \;=\; \arg\max_{\{\theta_\ell\}} \sum_{\ell=1}^{L} \frac{N_\ell^{tr}}{N^{tr}}\, \theta_\ell^T\, \widehat{\Sigma}_\ell^{-1}\, \theta_\ell \quad \textit{subject to } x \in \ell(\Pi), \tag{12}$$

where Π *denotes the union of all terminal leaves that constitute the covariate space. The covariance matrix can be estimated as* $\hat{\Sigma}_\ell = \hat{\Sigma}_\ell\big(\hat{\theta}(x, S^{tr}, \Pi)|N^{est}\big)$*. In this article, training and estimation samples have an equal number of observations,* $N^{tr} = N^{est}$*.*

Proof in Appendix C.

We use bootstrapping to estimate the variance-covariance matrix of parameters. Let $b = 1, \ldots, B$ index bootstrap samples. For each bootstrap iteration b, we estimate the parameter vector $\hat{\theta}(x, S_b^{est}, \Pi_b)$ for an individual with a covariate value x, using sample S_b^{est} and a signle tree Π_b. The bootstrap average is then given by

$$\bar{\theta}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}(x, S_b^{est}, \Pi_b), \tag{13}$$

where $\bar{\theta}(x)$ is an d-dimensional parameter vector. The variance-covariance matrix $\hat{\Sigma}(x)$ for the d-dimensional parameter vector $\hat{\theta}(x)$ is given by

$$\hat{\Sigma}(x) = \begin{bmatrix} \mathrm{Var}(\hat{\theta}_1(x)) & \mathrm{Cov}(\hat{\theta}_1(x), \hat{\theta}_2(x)) & \ldots & \mathrm{Cov}(\hat{\theta}_1(x), \hat{\theta}_d(x)) \\ \mathrm{Cov}(\hat{\theta}_2(x), \hat{\theta}_1(x)) & \mathrm{Var}(\hat{\theta}_2(x)) & \ldots & \mathrm{Cov}(\hat{\theta}_2(x), \hat{\theta}_d(x)) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{\theta}_d(x), \hat{\theta}_1(x)) & \mathrm{Cov}(\hat{\theta}_d(x), \hat{\theta}_2(x)) & \ldots & \mathrm{Var}(\hat{\theta}_d(x)) \end{bmatrix},$$

where the variance and covariance of parameters is defined pair-wise

$$\mathrm{Var}(\hat{\theta}_m(x)) = \frac{1}{B} \sum_{b=1}^{B} \Big(\hat{\theta}_m(x, S_b^{est}, \Pi_b) - \bar{\theta}_m(x)\Big)^2,$$

$$\mathrm{Cov}(\hat{\theta}_m(x), \hat{\theta}_{m'}(x)) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_m(x, S_b^{est}, \Pi_b) - \bar{\theta}_m(x)\right)\left(\hat{\theta}_{m'}(x, S_b^{est}, \Pi_b) - \bar{\theta}_{m'}(x)\right),$$

for $m, m' = 1, \ldots, d$-th parameters where $m \neq m'$. The confidence ellipse can be constructed as:

$$(\hat{\theta}(x, S_b^{est}, \Pi_b) - \bar{\theta}(x))^\top \hat{\Sigma}(x)^{-1}(\hat{\theta}(x, S_b^{est}, \Pi_b) - \bar{\theta}(x)) = r^2, \tag{14}$$

where $\hat{\Sigma}(x)$ denotes the estimated bootstrap variance-covariance matrix, and r^2 is the critical value determined from a chi-square distribution with d degrees of freedom.

5 Large Sample Properties

A common technique for investigating the asymptotic properties of tree-based estimators, including the random forest, is to view them as predictors of outcomes rather than as methods for estimating treatment effects (Wager, 2014; Athey and Wager, 2018). Although we use the outcome-prediction perspective, the resulting insights carry over to the generalized random forest framework. Building on the canonical approach introduced by Breiman (2001), we define the random forest estimator as the average of predictions produced by all possible size-s subsamples of the original N observations, in conjunction with auxiliary noise ξ. Specifically, for an individual characterized by covariates x, the random forest estimator is given by

$$\mathcal{F}(x, D_1, \ldots, D_N) = \frac{1}{\binom{N}{s}} \sum_{1 \leq i_1 \leq \cdots \leq i_s \leq N} \mathbb{E}_\xi\left[T(x, \xi, D_{i_1}, \ldots, D_{i_s})\right], \tag{15}$$

where $\binom{N}{s}$ is the number of all size-s subsamples, and T denotes the prediction of a single tree trained on a given subset. Each tree partitions the covariate space into disjoint subsets $\ell_1, \ldots, \ell_{|\Pi|}$. For a point x, the prediction of a tree takes the form

$$T(x,\xi,D_1,\ldots,D_s)=\sum_{n=1}^{|\Pi|}1(x\in\ell_n)\frac{1}{N_{\ell_n}}\sum_{\{i:X_i\in\ell_n\}}Y_i,$$

where N_{ℓ_n} is the number of training observations in leaf ℓ_n. $\mathcal{F}(x,D_1,\ldots,D_N)$ and $T(x,\xi,D_{i_1},\ldots,D_{i_s})$ are d-dimensional vectors (in $\mathbb{R}^d$), with all arithmetic operations applied coordinate-wise. We impose a set of assumptions to obtain consistency and asymptotic normality of the generalized random forest estimator for multiple correlated parameters.

Assumption 5 (Honesty)**.** *The outcome vector Y_i is statistically independent of the splitting parameters (j,x), which define the splitting coordinates and thresholds. Specifically, for each individual i contributing to the final prediction:*

$$F(Y_i|X_i,(j,x))=F(Y_i|X_i),$$

where F denotes the joint probability distribution of the $d-$dimensional outcome vector.

Assumption 5 is satisfied by partitioning the dataset into separate subsets: a training set (S^{tr}) and an estimation set (S^{est}). The training set S^{tr} is used to determine the tree structure, including the optimal splitting rules, while the estimation set S^{est} is used exclusively for outcome predictions. This ensures that prediction errors are not systematically influenced by the tree-building process (Athey and Imbens, 2016).

Assumption 6 (Data Generating Process)**.** *The covariates X_i are supported on the unit cube $X_i\in[0,1]^p$ with a density bounded away from zero and infinity. The conditional first and second moments, $\mathbb{E}(Y_i|X_i=x)$ and $\mathbb{E}(Y_i^2|X_i=x)$, are Lipschitz-continuous. Additionally, the conditional variance is strictly positive:*

$$\inf_{x\in[0,1]^G} Var(Y_i|X_i=x)>0.$$

These regularity conditions in Assumption 6 are common in the literature (Athey and Wager, 2018; Biau, 2012). Strictly positive variance ensures that the response variable has inherent randomness and is not fully determined by covariates X_i. While our results do not depend on specific distributional assumptions for X_i, the uniform support assumption influences constants carried throughout the analysis (see Lemma 2 and Theorem 3 in Athey and Wager, 2018).

Assumption 7 (Random Split Trees)**.** *At each recursive step, the probability of selecting the j-th covariate for splitting is at least π/p for some $\pi \in (0,1]$ and all $j = 1, \ldots, p$.*

Following Meinshausen and Ridgeway (2006) and Athey and Wager (2018), Assumption 7 ensures that every covariate has a strictly positive probability of being chosen at each step of the tree construction. This prevents the algorithm from systematically ignoring certain covariates.

Assumption 8 ((α, k)-Regular Splitting)**.** *There exists a constant $\alpha > 0$ such that each split retains at least an α fraction of the available training observations on each side. Moreover, splitting ceases when a node contains fewer than k observations for some integer $k \geq 1$.*

Assumption 8 ensures that partitions maintain a sufficient sample size, preventing excessively small leaf nodes. As shown in Wager and Walther (2015), under this condition, the resulting partitions are large in Euclidean volume. Additionally, the constraint on terminal node size imposes an upper bound on the variance of tree-based predictions.

Assumption 9 (Overlap)**.** *For some $\epsilon > 0$ and all $x \in [0,1]^p$,*

$$\epsilon < \mathbb{P}(W_i = 1 | X_i = x) < 1 - \epsilon$$

in each subset (leaf).

Assumption 9 guarantees that, for sufficiently large N, there exist both treated and untreated individuals across the covariate space. This condition can be satisfied by design.

A random forest estimator can be formulated as a U-statistic, a concept introduced by Hoeffding (1961) and further developed in statistical theory (Korolyuk and Borovskich, 2013). The Hoeffding decomposition, which provides a structured expansion of U-statistics, has been studied in the univariate case by Hájek (1968) and der Vaart (1998). To extend the large-sample theory of the random forest to a multivariate setting, I generalize the Hoeffding decomposition to multiple outcomes. Then I investigate the asymptotic properties of the random forest estimator by showing that the distance (squared L^2 norm) between the estimator and its Hoeffding decomposition converges to zero.

Consider a vector-valued function denoted as $T \in \mathbb{R}^d$. Assume, this function is measurable and permutation symmetric, where the latter implies that $T(\pi x) = T(x)$ holds true for all $\pi \in \Pi$ (a tree in this context). The Hajek projection of this function is defined as follows:

$$\mathring{T} = \mathbb{E}(T) + \sum_{i=1}^{N} \big[\mathbb{E}(T|X_i) - \mathbb{E}(T)\big] = \sum_{i=1}^{N} \mathbb{E}(T|X_i) - (N-1)\mathbb{E}(T). \tag{16}$$

Intuitively, (16) represents a projection of the vector-valued function T onto the linear subspace encompassing all random variables of the form $\sum_{i=1}^{N} f_i(X_i)$, where $f_i : \mathbb{R}^p \mapsto \mathbb{R}$ are arbitrary measurable functions satisfying $\mathbb{E}(f_i^2(X_i)) < \infty$ for $i = 1, \dots, N$. This projection ensures that the conditional expectation of $\mathring{T}$ in (16) coincides with the conditional expectation of T, denoted as:

$$\begin{aligned} \mathbb{E}(\mathring{T}|f_i(X_i)) &= \mathbb{E}(T|f_i(X_i)), \text{ and} \\ \mathbb{E}(\mathring{T}) &= \mathbb{E}(T). \end{aligned} \tag{17}$$

Now consider a vector-valued random forest estimator defined as $\mathcal{F}(x, D_1, \dots, D_N) \in \mathbb{R}^d$, with a corresponding vector of means μ. Let $\mathring{\mathcal{F}}(x, D_1, \dots, D_N)$ represent the Hajek projection of this multivariate random forest estimator, and let Σ denote the covariance matrix of the Hajek projection. It is important to note that the trees in $\mathring{\mathcal{F}}(x, D_1, \ldots, D_N)$ are symmetric, and the observations are independently and identically distributed (i.i.d) as before. Under these conditions, Lemma 1 holds.

Lemma 1. *The Hajek projection, denoted as $\mathring{\mathcal{F}}(x, D_1, \ldots, D_N)$, is given by the expression:*

$$\mathring{\mathcal{F}}(x, D_1, \ldots, D_N) - \mu = \frac{s}{N} \sum_{i=1}^{N} \big(T_1(D_i) - \mu\big),$$

where $\mathring{T} = \sum_{i=1}^{s} T_1(D_i)$ with $T_1(a) = \mathbb{E}_{\xi, D_2, \ldots, D_N} T(x, \xi, a, D_2, \ldots, D_N)$ represents the Hajek projection of a tree $T(x, D_1, \ldots, D_N) = \mathbb{E}_{\xi} T(x, \xi, D_1, \ldots, D_N) \in \mathbb{R}^d$. The parameter $s = N^{\beta}$ as before, and d represents the dimension of the outcome variables in each terminal node.

The covariance matrix Σ of the Hajek projection is given by:

$$\Sigma = \frac{s}{N} \mathbb{V}(\mathring{T}) \in \mathbb{R}^{d \times d},$$

where $\mathbb{V}$ denotes the covariance matrix of the projected elements of the tree.

Proof. See Appendix D. □

The projection meets the required conditions for the Lindeberg central limit theorem (Billingsley, 2013; DiCiccio and Romano, 2022), therefore, the Hajek projection of the multivariate random forest estimator is asymptotically normally distributed:

$$\Sigma^{-1/2} \big(\mathring{\mathcal{F}}(x, D_1, \ldots, D_N) - \mu\big) \xrightarrow{d} \mathcal{N}(0, I_{d \times d}),$$

where 0 is a $\mathbb{R}^d$ vector of zeros and $I_{d \times d}$ is an identity matrix.

To establish the asymptotic normality of the multivariate random forest estimator, we introduce an insightful relationship between the estimator and its projection by adding and subtracting $\Sigma^{-1/2}\mathring{\mathcal{F}}(x, D_1, \ldots, D_N)$ into the expression for $\Sigma^{-1/2}\big(\mathcal{F}(x, D_1, \ldots, D_N) - \mu\big)$. This leads to the following decomposition:

$$\Sigma^{-1/2}\big(\mathcal{F}(x, D_1, \ldots, D_N) - \mu\big) = \Sigma^{-1/2}\big(\mathcal{F}(x, D_1, \ldots, D_N) - \mathring{\mathcal{F}}(x, D_1, \ldots, D_N)\big) + \\ \Sigma^{-1/2}\big(\mathring{\mathcal{F}}(x, D_1, \ldots, D_N) - \mu\big).$$

Formally, the objective of this article is to show that:

$$\Sigma^{-1/2}\big(\mathcal{F}(x, D_1, \ldots, D_N) - \mathring{\mathcal{F}}(x, D_1, \ldots, D_N)\big) \xrightarrow{p} 0.$$

Then by Slutsky's theorem, the multivariate random forest estimator is asymptotically normally distributed.

In line with Athey and Wager (2018), I derive the lower bound of the variance of a vector-valued forest $\mathring{\mathcal{F}}(x, D_1, \ldots, D_N)$ and demonstrate its' convergence to zero. The primary focus lies in proving the convergence in squared mean of the expression $\Sigma^{-1/2}\big(\mathcal{F}(x, D_1, \ldots, D_N) - \mathring{\mathcal{F}}(x, D_1, \ldots, D_N)\big)$. For the sake of clarity, we shall use the more concise notations $\mathcal{F}$ and $\mathring{\mathcal{F}}$ to represent $\mathcal{F}(x, D_1, \ldots, D_N)$ and $\mathring{\mathcal{F}}(x, D_1, \ldots, D_N)$, respectively. Lemma 2 obtains the upper bound of the squared deviation between the forest and its' Hajek projection.

Lemma 2. *The mean squared difference of $\mathcal{F}$ and $\mathring{\mathcal{F}}$ has the upper bound:*

$$\mathbb{E}\big(\mathcal{F} - \mathring{\mathcal{F}}\big)^T \Sigma^{-1}\big(\mathcal{F} - \mathring{\mathcal{F}}\big) \leq \frac{s}{N} tr\Big(\big(\mathbb{V}(\mathring{T})\big)^{-1}\mathbb{V}(T)\Big),$$

where tr is a trace operator, and $\mathbb{V}(T)$ and $\mathbb{V}(\mathring{T})$ denote the variance of a multivariate tree and its' Hajek projection, respectively.

Proof. See Appendix E. □

Under Assumptions 5-9, Theorem 1 shows that $\frac{s}{N}tr\Bigg(\big(\mathbb{V}(\mathring{T})\big)^{-1}\mathbb{V}(T)\Bigg)$ approaches zero in the limit.

Theorem 1. *The entries of* $\mathbb{V}(T)$ *are bounded and its diagonal elements are bounded away from zero. Moreover, the lower bound of the off-diagonal terms of* $\mathbb{V}(\mathring{T})$ *are on the order of* $o\big(\frac{1}{\log^p(s)}\big)$*. The upper bound in Lemma 2 converges to zero in the limit:*

$$\frac{s}{N}tr\Bigg(\big(\mathbb{V}(\mathring{T})\big)^{-1}\mathbb{V}(T)\Bigg) \to 0.$$

Proof. See Appendix F. □

Intuitively, Theorem 1 shows that $\mathbb{V}(\mathring{T})$ shrinks at least as fast as $1/\log^p(s)$, while $\mathbb{V}(T)$ remains bounded. That ensures $(\mathbb{V}(\mathring{T}))^{-1}$ grows on the order $\log^p(s)$. By Slutsky's theorem, Theorem 1 implies that the multivariate random forest estimator is asymptotically normally distributed.

6 Simulated Experiments

We investigate the consequences of misspecifying both constant and heterogeneous covariance structures by repeatedly simulating two designs. In both designs, we generate a sample of size N with n-dimensional outcomes 500 times. Covariates $X \in \mathbb{R}^3$ are drawn from a standard normal distribution, and treatment assignment follows a Bernoulli distribution:

$$P_i \sim \text{Bernoulli}(0.5). \tag{18}$$

The vector of true treatment effects, $\tau_i = (\tau_{i1}, \ldots, \tau_{in})^\top$, is drawn from a multivariate normal distribution:

$$\tau_i \sim \mathcal{N}(\mu, \Sigma), \tag{19}$$

where $\mu = (5, \ldots, 5)^\top$.

Constant Covariance. The covariance matrix is constant across individuals, defined as

$$\Sigma_{jl} = \rho\sqrt{\sigma_j^2 \sigma_l^2}, \quad j \neq l, \quad \Sigma_{jj} = \sigma_j^2, \tag{20}$$

where the variances follow a decreasing pattern:

$$\sigma_j^2 = 0.1 - 0.02(j-1). \tag{21}$$

Observed outcomes are generated as

$$Y_{ij} = 0.5 + P_i(\tau_{ij} + k) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, 1), \tag{22}$$

where $k = 5$. The dataset includes covariates X_i, treatment assignments P_i, observed matrix of outcomes Y_i, and the fixed variance-covariance structure.

Personalized Covariance. In the second design, the covariance matrix is heteroskedastic. The variance for outcome j follows a log-normal distribution:

$$\sigma_{ij}^2 \sim \text{Lognormal}(\log(1.0 - 0.2(j-1)), 0.3). \tag{23}$$

The covariance matrix is defined as

$$\Sigma_i[j, l] = \rho\sqrt{\sigma_{ij}^2 \sigma_{il}^2}, \quad j \neq l, \quad \Sigma_i[j, j] = \sigma_{ij}^2. \tag{24}$$

Observed outcomes incorporate covariate effects and are given by

$$Y_{ij} = 0.5 + P_i(\tau_{ij} + k) + 0.2X_{i1} + 0.3X_{i2} + 0.4X_{i3} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, 0.5). \tag{25}$$

The dataset includes covariates X_i, treatment assignments P_i, observed outcomes Y_i, true treatment effects $\tau_i + k$ with $k = 5$, and a heteroskedastic covariance matrix.

Error rates for varying sample sizes and correlations. Table 2 reports Type I error rates for varying sample sizes and treatment effect correlations under constant and heteroskedastic covariance. When the covariance is misspecified, higher correlations (ρ) increase Type I error rates, particularly when the covariance is constant. Increasing the sample size from 500 to 2000 reduces Type I error rates across all conditions. With the heteroskedastic covariance matrix, Type I error rates are consistently lower, especially at higher correlations, indicating improved statistical precision when accounting for individual-specific variance-covariance. The results are stable across simulated experiments.

Table 2: Type I error rates with standard errors (in parentheses) for varying sample sizes and treatment effect correlations, averaged over 500 simulations. Estimates of treatment effects and covariance matrices are estimated using a multi-outcome generalized random forest. The number of trees in each simulated experiment is set to 1000 and the number of outcomes equals two.

	Constant Covariance				Personalized Covariance			
	Correlation of treatment effects (ρ)							
N	0.1	0.3	0.5	0.9	0.1	0.3	0.5	0.9
500	0.070 (0.002)	0.086 (0.002)	0.108 (0.002)	0.129 (0.001)	0.057 (0.000)	0.063 (0.000)	0.069 (0.000)	0.087 (0.000)
1000	0.055 (0.001)	0.081 (0.001)	0.108 (0.001)	0.129 (0.000)	0.063 (0.000)	0.067 (0.000)	0.071 (0.000)	0.089 (0.000)
2000	0.044 (0.001)	0.078 (0.001)	0.107 (0.001)	0.130 (0.000)	0.070 (0.000)	0.072 (0.000)	0.076 (0.000)	0.091 (0.000)

Table 3 demonstrates that higher correlation among treatment effects substantially increases Type II error rates under misspecified covariance. When the covariance is constant and the sample size is 2000, the error rate increases by up to a factor of 7.3 as correlation rises from 0.1 to 0.9.

Table 3: Type II error rates with standard errors (in parentheses) for varying sample sizes and treatment effect correlations, averaged over 500 simulations. Estimates of treatment effects and covariance matrices are estimated using a multi-outcome generalized random forest. The number of trees in each simulated experiment is set to 1000 and the number of outcomes equals two.

	Constant Covariance				Personalized Covariance			
	Correlation of treatment effects (ρ)							
N	0.1	0.3	0.5	0.9	0.1	0.3	0.5	0.9
500	0.144 (0.005)	0.207 (0.007)	0.306 (0.007)	0.487 (0.007)	0.157 (0.001)	0.193 (0.001)	0.241 (0.001)	0.419 (0.001)
1000	0.096 (0.004)	0.176 (0.005)	0.292 (0.006)	0.497 (0.005)	0.173 (0.001)	0.197 (0.001)	0.232 (0.001)	0.388 (0.001)
2000	0.071 (0.003)	0.158 (0.004)	0.278 (0.004)	0.518 (0.003)	0.199 (0.000)	0.215 (0.000)	0.240 (0.000)	0.363 (0.001)

Figure 2 illustrates the relationship between the number of outcomes and Type I and Type II error rates. As the number of outcomes increases, the Type II error rate rises sharply, approaching 100%. This indicates that as the dimensionality of the outcomes grows, the probability of failing to reject a false null hypothesis becomes nearly certain, leading to substantial losses in statistical power.

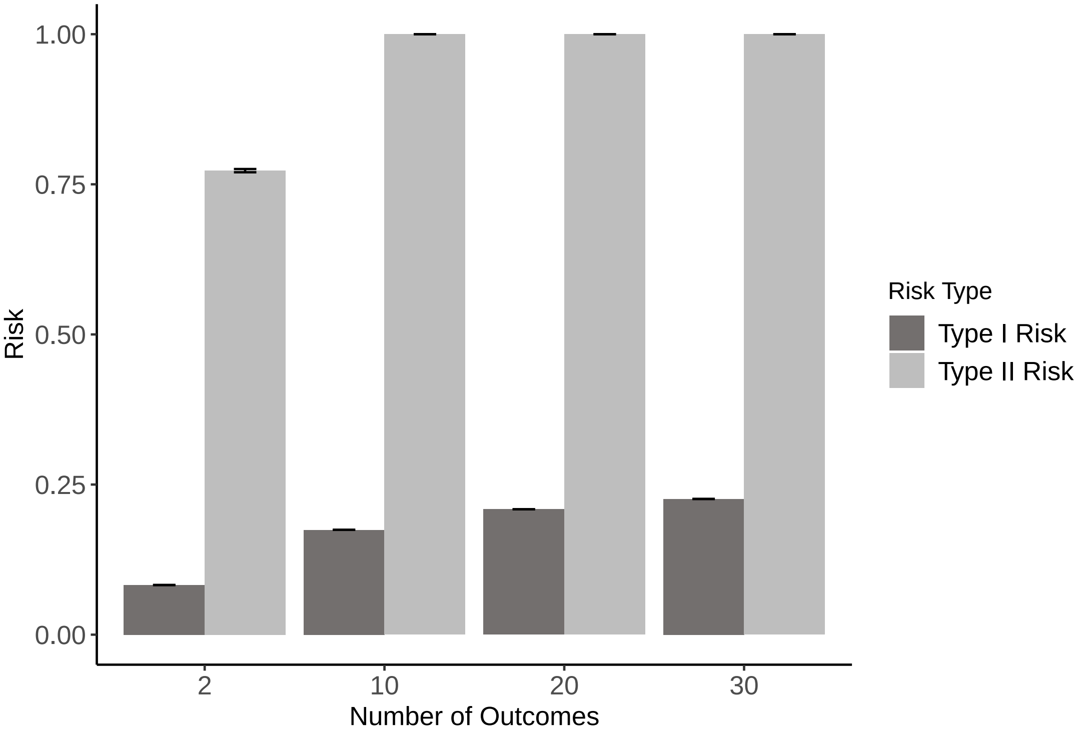


Figure 2: Average Type I and Type II error rates by the number of outcomes, with a fixed covariance of 0.5 and a sample size of 500. Results are averaged over 500 Monte Carlo simulations. The number of trees in each experiment is 1000.

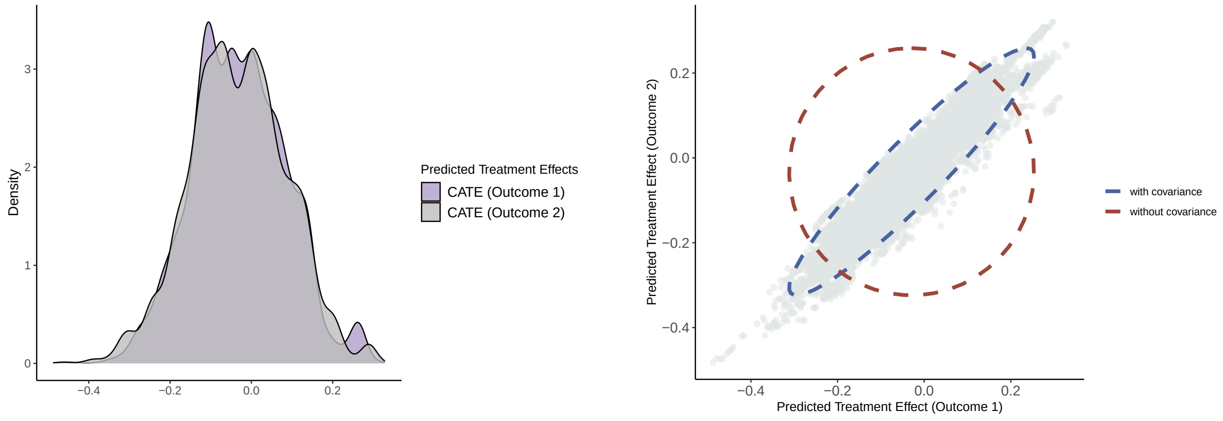
7 Application

The Pennsylvania Reemployment Bonus Experiment was a randomized experiment conducted in 1988–89 to study the impact of financial incentives on unemployment duration. Participants were assigned to a control group receiving standard unemployment benefits or to one of six treatment groups, which were offered cash bonuses contingent on securing full-time employment (at least 32 hours per week) within a designated qualification period and maintaining it for at least 16 weeks. The experiment tested two bonus levels—500 (low) and 997 (high)—and two qualification periods of 6 and 12 weeks. Additionally, some treatment groups were offered job-search assistance, including workshops and individualized assessment sessions. The analysis in this article focuses on treatment Group 4, which received a high bonus and a long qualification period. We consider two measures of unemployment duration as outcomes: the logarithm of length of the unemployment spell and logarithm of the time to stable reemployment. These outcomes have 92% correlation (Bilias, 2000). [5]

Figure 3 illustrates the distribution and covariance structure of the estimated treatment effects. Panel (a) presents the density of treatment effects for both unemployment measures, showing a central concentration around zero with substantial variation, including both positive and negative effects. The treatment effects are highly correlated, with a correlation coefficient of 0.94.

Panel (b) illustrates the joint confidence regions of treatment effects. The confidence ellipses show that omitting correlation (dashed red) inflates uncertainty, resulting in a Type I error rate of 16.7% and a Type II error rate of 71.0%. In contrast, incorporating covariance (dashed blue) provides a more precise representation of joint treatment effect variability.

[5] Variables are described here. Raw data can be downloaded using this link.



(a) Density of Treatment Effects

(b) Covariance Structure of Treatment Effects

Figure 3: Joint confidence regions. Panel (a) shows the density of estimated treatment effects for both outcomes. Panel (b) displays the joint distribution with confidence ellipses with and without the covariance of treatment effects. Treatment effects are estimated using a multi-outcome generalized random forest with 1,000 trees. The confidence ellipse and confidence circle are constructed based on the mean of the predicted treatment effects and their corresponding variance-covariance matrix.

8 Conclusion

This paper introduces a multivariate extension of the generalized random forest framework for estimating and inferring correlated treatment effects across multiple outcomes. The primary theoretical contribution is to show that the deviation between the generalized random forest estimator and its' orthogonal projection onto a space of additive components vanishes asymptotically, even when treatment effects are correlated. This ensures valid statistical inference under structural assumptions, including honesty, Lipschitz continuity, random splits, α-k regularity, and overlap. Simulation results and an empirical application highlight the importance of accounting for parameter correlations in the design. The proposed framework explicitly accounts for the covariance of heterogeneous treatment effects in the design and improves the power of the joint hypothesis testing. A potential extension of this approach is to integrate covariates with measurement error.

References

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360. doi: 10.1073/pnas.1510489113.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2). doi: 10.1214/18-AOS1709.

Athey, S. and Wager, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242. doi: 10.1080/01621459.2017.1319839.

Becker, B. J. (2000). Multivariate meta-analysis. *Handbook of applied multivariate statistics and mathematical modeling*, pages 499–525.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13:1063–1095.

Bilias, Y. (2000). Sequential testing of duration data: the case of the pennsylvania 'reemployment bonus' experiment. *Journal of Applied Econometrics*, 15(6):575–594. doi: https://www.jstor.org/stable/2678561.

Billingsley, P. (2013). *Convergence of probability measures.* John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Breiman, L. (2004). Consistency for a simple model of random forests. *University of California at Berkeley. Technical Report*, 670.

Bühlmann, P., Ćevid, D., Michel, L., Näf, J., and Meinshausen, N. (2020). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *arXiv preprint arXiv:2005.14458*.

Denil, M., Matheson, D., and De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *International conference on machine learning*, pages 665–673. PMLR.

der Vaart, Van, A. A. (1998). Cambridge university press: New york. *NY, USA*.

DiCiccio, C. and Romano, J. (2022). Clt for u-statistics with growing dimension. *Statistica Sinica*, 32(1). doi:https://www.jstor.org/stable/27108526.

Gleser, L. J. and Olkin, I. (2000). Meta-analysis for 2× 2 tables with multiple treatment groups. In *Meta-analysis in medicine and health policy*, pages 165–176. CRC Press.

Gleser, L. J., Olkin, I., et al. (2009). Stochastically dependent effect sizes. *The handbook of research synthesis and meta-analysis*, 2:357–376.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, pages 325–346. https://www.jstor.org/stable/2239025.

Hoeffding, W. (1961). The strong law of large numbers for u-statistics. Technical report, North Carolina State University. Dept. of Statistics.

Ishak, K. J., Platt, R. W., Joseph, L., and Hanley, J. A. (2008). Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in medicine*, 27(5):670–686. doi: 10.1002/sim.2913.

Jeon, Y. and Lin, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590. doi: https://doi.org/10.1198/016214505000001230.

Kim, R.-S. and Becker, B. J. (2010). The degree of dependence between multiple-treatment effect sizes. *Multivariate Behavioral Research*, 45(2):213–238. doi: https://doi.org/10.1080/00273171003680104.

Korolyuk, V. S. and Borovskich, Y. V. (2013). *Theory of U-statistics*, volume 273. Springer Science & Business Media.

Li, K. (2020). Asymptotic normality for multivariate random forest estimators. *arXiv preprint arXiv:2012.03486*. doi: https://doi.org/10.1111/j.1467-985X.2008.00593.x.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of machine learning research*, 7(6).

Nekipelov, D., Novosad, P., and Ryan, S. P. (2018). Moment forests. *National Bureau of Economic Research, working paper*.

Peccati, G. (2004). Hoeffding-anova decompositions for symmetric statistics of exchangeable observations. *The Annals of Probability*. doi: 10.1214/009117904000000405.

Riley, R. D. (2009). Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 172(4):789–811. doi: https://doi.org/10.1111/j.1467-985X.2008.00593.x.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55. doi: https://doi.org/10.1093/biomet/70.1.41.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292. https://doi.org/10.1016/0378-3758(90)90077-8.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*. doi: 10.1214/15-AOS1321.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., and Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior research methods*, 47:1274–1294. doi: https://doi.org/10.3758/s13428-014-0527-2.

Wager, S. (2014). Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*.

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Wang, G., Li, J., and Hopp, W. J. (2022). An instrumental variable forest approach for detecting heterogeneous treatment effects in observational studies. *Management Science*, 68(5):3399–3418. doi: . https:// doi.org/10.1287/mnsc.2021.4084.

A Supplementary Material

B Expected Mahalanobis Distance under Misspecified Covariance

Proof. For a d-dimensional random vector $\hat{\theta}_i \sim N(\theta_i, \Sigma_i)$, we derive the expected Mahalanobis distances under correct and misspecified covariance matrices. When using the correct covariance matrix Σ_i, the Mahalanobis distance is:

$$D^2_{\text{correct},i} = (\hat{\theta}_i - \theta_i)^T \Sigma_i^{-1} (\hat{\theta}_i - \theta_i).$$

If $\hat{\theta}_i \sim N(\theta_i, \Sigma_i)$, then $D^2_{\text{correct},i} \sim \chi^2_d$. Since $E[\chi^2_d] = d$, we have:

$$\mathbb{E}[D^2_{\text{correct},i}] = d.$$

If we instead use an incorrect covariance matrix $\Sigma_{0,i}$, the Mahalanobis distance becomes:

$$D^2_{\text{misspecified},i} = (\hat{\theta}_i - \theta_i)^T \Sigma_{0,i}^{-1} (\hat{\theta}_i - \theta_i).$$

Since $\hat{\theta}_i \sim N(\theta_i, \Sigma_i)$, we can substitute $\hat{\theta}_i - \theta_i = \Sigma_i^{1/2} Z_i$:

$$D^2_{\text{misspecified},i} = Z_i^T (\Sigma_i^{1/2} \Sigma_{0,i}^{-1} \Sigma_i^{1/2}) Z_i.$$

Now, for the expectation:

$$\begin{aligned}
\mathbb{E}[D^2_{\text{misspecified},i}] &= \mathbb{E}[Z_i^T (\Sigma_i^{1/2} \Sigma_{0,i}^{-1} \Sigma_i^{1/2}) Z_i] \\
&= \mathbb{E}[\text{tr}(Z_i^T (\Sigma_i^{1/2} \Sigma_{0,i}^{-1} \Sigma_i^{1/2}) Z_i)] \quad (\text{scalar} = \text{trace of } 1 \times 1 \text{ matrix}) \\
&= \text{tr}(\mathbb{E}[Z_i Z_i^T] (\Sigma_i^{1/2} \Sigma_{0,i}^{-1} \Sigma_i^{1/2})) \quad (\text{linearity of trace}) \\
&= \text{tr}(I_d (\Sigma_i^{1/2} \Sigma_{0,i}^{-1} \Sigma_i^{1/2})) \quad (\text{since } \mathbb{E}[Z_i Z_i^T] = I_d \text{ for } Z_i \sim N(0, I_d)) \\
&= \text{tr}(\Sigma_i^{1/2} \Sigma_{0,i}^{-1} \Sigma_i^{1/2}).
\end{aligned}$$

Finally, define λ_i as the average eigenvalue:

$$\lambda_i = \frac{1}{d}\mathrm{tr}(\Sigma_i^{1/2}\Sigma_{0,i}^{-1}\Sigma_i^{1/2}).$$

That implies:

$$\mathbb{E}[D^2_{\text{misspecified},i}] = \mathrm{tr}(\Sigma_i^{1/2}\Sigma_{0,i}^{-1}\Sigma_i^{1/2}) = d\lambda_i.$$

□

C Method of Moments Estimator

Let Π be a partition of the covariate space into leaves $\ell = 1, \ldots, L$. For each leaf ℓ, define:

$$\theta^\star_\ell \;=\; \mathbb{E}\big[\theta_i \,\big|\, X_i \in \ell(\Pi)\big], \quad \text{and} \quad \hat{\theta}_\ell \;=\; \hat{\theta}(x, S^{est}, \Pi) \text{ for all } x \in \ell.$$

That is, $\theta^\star_\ell$ is the *population* (true) treatment-effect vector in leaf ℓ, and $\hat{\theta}_\ell$ is the corresponding *estimator*, trained on a subsample S^{tr} and evaluated using an independent subsample S^{est}.

Let Σ_ℓ be the (true) covariance matrix of θ_i given $X_i \in \ell$, and let $\widehat{\Sigma}_\ell$ be an empirical covariance estimate obtained from the training sample S^{tr}.

We study the *scaled mean squared error*

$$\mathbb{E}_{S^{tr},S^{est}}\Big[\big(\theta_i - \hat{\theta}(X_i, S^{est}, \Pi)\big)^T \,\Sigma_i^{-1}\, \big(\theta_i - \hat{\theta}(X_i, S^{est}, \Pi)\big) \;-\; \theta_i^T\,\Sigma_i^{-1}\,\theta_i\Big],$$

where $\Sigma_i = \Sigma_\ell$ whenever $X_i \in \ell$. To rewrite this expression, define

$$A \;=\; \theta_i \;-\; \theta^\star_\ell, \quad B \;=\; \theta^\star_\ell \;-\; \hat{\theta}_\ell.$$

Note that A depends on the random variables θ_i (and thus on the training sample S^{tr}) but not on $\hat{\theta}_\ell$, whereas B depends on $\hat{\theta}_\ell$ (and hence on S^{est}) but not on θ_i.

Using A and B, one obtains the following decomposition:

$$\mathbb{E}_{S^{tr},S^{est}}\Big[\big(\theta_i - \hat{\theta}(X_i, S^{est}, \Pi)\big)^T \Sigma_i^{-1}\big(\theta_i - \hat{\theta}(X_i, S^{est}, \Pi)\big) \;-\; \theta_i^T\,\Sigma_i^{-1}\,\theta_i\Big]$$

$$= \mathbb{E}_{S^{tr},S^{est}}\big[(A+B)^T\,\Sigma_i^{-1}\,(A+B) \;-\; \theta_i^T\,\Sigma_i^{-1}\,\theta_i\big].$$

Since $A = \theta_i - \theta_\ell^\star$ and $B = \theta_\ell^\star - \hat{\theta}_\ell$ are constructed so that $\mathrm{Cov}(A, B) = 0$ (because A depends on S^{tr} and B depends on S^{est}), one obtains the decomposition

$$\begin{aligned}
&\mathbb{E}_{S^{tr},S^{est}}\Big[\big(\theta_i - \hat{\theta}(X_i, S^{est}, \Pi)\big)^T \Sigma_i^{-1}\big(\theta_i - \hat{\theta}(X_i, S^{est}, \Pi)\big) - \;\theta_i^T\,\Sigma_i^{-1}\,\theta_i\Big] && (26)\\
&\quad= \;\mathbb{E}_{S^{tr}}\big[\theta_i^T\,\Sigma_i^{-1}\,\theta_i - \;2\,\theta_i^T\,\Sigma_i^{-1}\,\theta_\ell^\star + \;(\theta_\ell^\star)^T\,\Sigma_i^{-1}\,\theta_\ell^\star - \;\theta_i^T\,\Sigma_i^{-1}\,\theta_i\big] && (27)\\
&\quad+ \;\mathbb{E}_{X_i,S^{est}}\Big[\big(\theta_\ell^\star - \hat{\theta}_\ell\big)^T\,\Sigma_\ell^{-1}\,\big(\theta_\ell^\star - \hat{\theta}_\ell\big)\Big]\\
&\quad= \;-\,\mathbb{E}_{X_i}\big[(\theta_\ell^\star)^T\,\Sigma_\ell^{-1}\,\theta_\ell^\star\big] \;+\; \mathbb{E}\big[\mathrm{tr}(I)\big].
\end{aligned}$$

Here we have used:

- $\mathbb{E}[\hat{\theta}_\ell \mid X_i \in \ell] = \theta_\ell^\star$;
- The cross-term $\mathrm{Cov}(A, B) = 0$ because A depends on S^{tr} but not S^{est}, while B depends on S^{est} but not S^{tr};
- $\mathbb{E}[(\theta_\ell^\star - \hat{\theta}_\ell)^T\Sigma_\ell^{-1}(\theta_\ell^\star - \hat{\theta}_\ell)] = \mathrm{tr}(\Sigma_\ell^{-1}\Sigma_\ell) = \mathrm{tr}(I)$.

Hence, up to an additive constant (that does not depend on $\hat{\theta}_\ell$), the quantity we want to *minimize* is $\mathbb{E}[(\theta_i - \hat{\theta}_i)^T\Sigma_i^{-1}(\theta_i - \hat{\theta}_i)]$. Equivalently, one can *maximize* $\mathbb{E}[(\theta_\ell^\star)^T\,\Sigma_\ell^{-1}\,\theta_\ell^\star]$.

In practice, we obtain an estimator $\widehat{\Sigma}_\ell$ of Σ_ℓ (for each leaf ℓ) from a training sample S^{tr}. Then, using an estimation sample S^{est} of size N^{est}, one solves

$$\hat{\theta}(x, S^{est}, \Pi) \;=\; \arg\max_{\{\theta_\ell\}} \sum_{\ell=1}^{L} \frac{N_\ell^{tr}}{N^{tr}}\;\theta_\ell^T\,\widehat{\Sigma}_\ell^{-1}\,\theta_\ell \quad \text{subject to } x \in \ell(\Pi),$$

where N_ℓ^{tr} is the number of training-sample observations in leaf ℓ, and $N^{tr} = \sum_\ell N_\ell^{tr}$. Throughout this article, we assume $N^{tr} = N^{est}$ for simplicity, but one could use any two independent samples for training the algorithm and parameter estimation, respectively.

D Lemma 1

Proof. Define the Hajek projection of the multivariate random forest estimator:

$$\mathring{\mathcal{F}}(x, D_1, \ldots, D_N) - \mu = \sum_{i=1}^{N} \mathbb{E}\big(\mathcal{F}(x, D_1, \ldots, D_N) - \mu | D_i\big) = \tag{28}$$

$$\frac{1}{\binom{N}{s}} \sum_{i=1}^{N} \mathbb{E}\Bigg(\sum_{1 \leq i_1 \leq \cdots \leq i_s \leq N} \mathbb{E}_{\xi} T(x, \xi, D_{i_1}, \ldots, D_{i_s}) - \mu | D_i \Bigg),$$

where $\binom{N}{s}$ is the number of $i_1 \leq \cdots \leq i_s$ size-s subsets from $1, \ldots, N$ observations. When the observation i is not in samples, then the conditional expectation of the tree (aggregated over the randomization) is the same as the unconditional one. Therefore:

$$\mathbb{E}\big(\mathbb{E}_{\xi} T(\xi, D_{i_1}, \ldots, D_{i_s}) | A_i\big) = \mathbb{E}_{\xi, D_{i_1}, \ldots, D_{i_s}} T(x, \xi, D_{i_1}, \ldots D_{i_s}) = \mu.$$

Overall, there are $\binom{N-1}{s-1}$ samples that contain observation i. Moreover, the sequence of observations is *i.i.d.* and the trees are permutation symmetric. Therefore, for each sample,

$$\mathbb{E}\big(\mathbb{E}_{\xi} T(x, \xi, D_{i_1}, \ldots, D_{i_s}) - \mu | D_i\big) = T_1(D_i) - \mu, \tag{29}$$

where $T_1(a) = \mathbb{E}_{\xi, D_2, \ldots, D_N} T(x, \xi, a, D_2, \ldots, D_N)$.

Incorporating (29) in (28) yields:

$$\mathring{\mathcal{F}}(x, D_1, \ldots, D_N) - \mu = \frac{\binom{N-1}{s-1}}{\binom{N}{s}} \sum_{i=1}^{N} \big(T_1(D_i) - \mu\big) = \frac{s}{N} \sum_{i=1}^{N} \big(T_1(D_i) - \mu\big). \tag{30}$$

Since the observations $D_1, \ldots, D_N$ are i.i.d, the same property holds for $T_1(D_i)$. By taking the expectation of both sides in (30), we can easily verify that $\mathbb{E}\big(\mathring{\mathcal{F}}(x)\big) = \mu$ where $\mathring{\mathcal{F}}(x) = \mathring{\mathcal{F}}(x, D_1, \ldots, D_N)$. Define Σ to be the covariance matrix of $\mathring{\mathcal{F}}(x, D_1, \ldots, D_N)$. Then:

$$\Sigma = \mathbb{V}\Bigg[\frac{s}{N} \sum_{i=1}^{N} \big(T_1(D_i) - \mu\big)\Bigg] = \frac{s^2}{N} \mathbb{V}\big(T_1(D_i)\big) = \frac{s}{N} \mathbb{V}\Big(\sum_{i=1}^{s} T_1(D_i)\Big) = \frac{s}{N} \mathbb{V}(\mathring{T}) \in \mathbb{R}^{d \times d}, \tag{31}$$

where $\mathring{T} = \sum_{i=1}^{s} T_1(D_i)$ is the Hajek projection of a tree

$$T(x, D_1, \ldots, D_N) = \mathbb{E}_\xi T(x, \xi, D_1, \ldots, D_N) \in \mathbb{R}^d.$$

Note that, a tree T is symmetric in its arguments, and observations $i = 1, \ldots, N$ are *i.i.d.* Therefore, the Hajek projection of a tree estimator reduces to $\sum_{i=1}^{s} T_1(D_i)$ (as in (30)). We disregard the second (constant) term, as it does not enter in the variance $\mathbb{V}$. Note that since the statistic $T_1(D_i)$ is a vector, the operation $\mathbb{V}$ applies coordinate-wise.

□

E Lemma 2

Proof. Define the mean squared deviation of the multivariate forest estimator and its projection:

$$\mathbb{E}(\mathcal{F} - \mathring{\mathcal{F}})^T \Sigma^{-1} (\mathcal{F} - \mathring{\mathcal{F}}) = \mathbb{E}\big[\mathrm{tr}\Sigma^{-1}(\mathcal{F} - \mathring{\mathcal{F}})(\mathcal{F} - \mathring{\mathcal{F}})^T\big] = \tag{32}$$

$$\mathrm{tr}\Sigma^{-1}\mathbb{E}(\mathcal{F} - \mathring{\mathcal{F}})(\mathcal{F} - \mathring{\mathcal{F}})^T = \mathrm{tr}\Sigma^{-1/2}\mathbb{V}\big(\mathcal{F} - \mathring{\mathcal{F}}\big)\Sigma^{-1/2}.$$

Assume there exist functions T_i, such that the following equality holds:

$$\mathbb{E}\big(T_i(X_i \in B) | X_i \notin B)\big) = 0, \tag{33}$$

where B is a generic measurable set in the covariate space. Equation (33) is the necessary condition for the weak independence of the exchangeable sequences of X_i. Assume, $T_i(X_i \in B)$ are symmetric, square-integrable, vector-valued functions. Then each T_i and $T_{i'}$ are pairwise independent. Since $i = 1, \ldots, N$ is an exchangeable ($i.i.d$) sequence, Theorem 6 of Peccati (2004) applies. In addition, Proposition 1 of Li (2020) applies to our case as well.

We define Höeffding decomposition of a multivariate U-statistic:

$$\mathcal{F} - \mathring{\mathcal{F}} = \frac{1}{\binom{N}{s}} \Bigg[\sum_{i<j} \binom{N-2}{s-2} \big(T_2(D_i, D_j) - \mu\big) + \sum_{i<j<m} \binom{N-3}{s-3} \big(T_3(D_i, D_j, A_m) - \mu\big) + \dots \tag{34}$$

where $T_2, T_3 \dots$ are second, third, and higher order projections of a tree T that meet the following conditions:

$$\mathbb{E}\big(T_i - \mu\big)^T \Sigma^{-1} \big(T_{i'} - \mu\big) = 0 \text{ for each } i \neq i', \text{ and} \tag{35}$$

$$\mathbb{E}\big(T_i - \mu\big)^T \Sigma^{-1} \big(T_i - \mu\big) \leq \mathbb{E}\big(T - \mu\big)^T \Sigma^{-1} \big(T - \mu\big), \tag{36}$$

where T_i and $T_{i'}$ are the i- th and i'-th projections of the tree, with $i = 1, \dots, N$.

We fix the variance (Σ) of the multivariate random forest estimator. Moreover, we notice that $\binom{N}{s} \geq \binom{N-1}{s-1} \geq \binom{N-2}{s-2} \geq \binom{N-3}{s-3} \geq \dots$. Therefore:

$$\mathcal{F} - \mathring{\mathcal{F}} \leq \frac{s}{N} \Bigg[\sum_{i<j} \big(T_2(D_i, D_j) - \mu\big) + \sum_{i<j<m} \big(T_3(D_i, D_j, A_m) - \mu\big) + \dots \Bigg], \tag{37}$$

where $\frac{s}{N} = \frac{\binom{N-1}{s-1}}{\binom{N}{s}}$. Based on Equation (36), the variance of $\mathcal{F} - \mathring{\mathcal{F}}$ has an upper bound:

$$\mathbb{V}\big(\mathcal{F} - \mathring{\mathcal{F}}\big) \leq \left(\frac{s}{N}\right)^2 \mathbb{V}(T). \tag{38}$$

In (31) we derived $\Sigma = \frac{s}{N}\mathbb{V}(\mathring{T})$. Plugging the value of Σ and (38) in (32) leads to the upper bound of the squared deviation:

$$\mathbb{E}(\mathcal{F} - \mathring{\mathcal{F}})^T \Sigma^{-1} (\mathcal{F} - \mathring{\mathcal{F}}) \leq tr\bigg(\big(\frac{s}{N}\mathbb{V}(\mathring{T})\big)^{-1/2} \left(\frac{s}{N}\right)^2 \mathbb{V}(T) \big(\frac{s}{N}\mathbb{V}(\mathring{T})\big)^{-1/2} \bigg) = \tag{39}$$
$$\frac{s}{N} tr\bigg(\big(\mathbb{V}(\mathring{T})\big)^{-1} \mathbb{V}(T) \bigg)$$

In the final equality, we use the cyclic property of the trace operator: $tr(XYZ) = tr(YZX) = tr(ZXY)$.

□

F Theorem 1

Proof. Bounded elements of $\mathbb{V}(T)$ directly follow from the proposed assumptions. According to Assumption 8, the number of observations in each terminal node is bounded above. This implies that the variance of the tree is bounded above by constant times $\mathbb{V}(Y_{im}|X_i = x)$ for $m-$ th outcome out of d outcomes. Moreover, Assumption 6 guarantees that $\mathbb{V}(Y_{im}|X_i = x)$ is bounded away from zero.

In the context at hand, we rely on the findings presented by Athey and Wager (2018) concerning the order of variance terms. Specifically, they show that:

$$\mathbb{V}(\mathring{T})_{ii} = \frac{C}{\log^p(s)}, \text{ for some constant } C. \tag{40}$$

$\mathbb{V}(\mathring{T})_{ii}$ denotes the diagonal terms of the variance of the projection of a tree estimator. We show that the off-diagonal terms $\mathbb{V}(\mathring{T})_{ij} = o\big(\frac{1}{\log^p(s)}\big)$ for all $i \neq j$.

Start with the definition of a Hajek projection of a tree:

$$\mathring{T} - \mu = \sum_{i=1}^{s} \mathbb{E}(T|D_i) \tag{41}$$

Since the observations are $i.i.d.$, then:

$$\mathbb{V}\big(\mathring{T}\big) = s\mathbb{V}\big(\mathbb{E}(T|D_1)\big). \tag{42}$$

Then it is clear to see that:

$$\mathbb{V}\big(\mathbb{E}(T|D_1)\big) = \mathbb{V}\big(\mathbb{E}(T|D_1) - \mathbb{E}(T|X_1)\big) + \mathbb{V}\big(\mathbb{E}(T|X_1)\big). \tag{43}$$

Consider m-th outcome variable, where $m = 1, \ldots, d$. Since the tree is honest, the diagonal terms in (43) simplify as follows (see the Proof of Theorem 5 in Athey and Wager, 2018):

$$\begin{aligned} \mathbb{V}\big(\mathbb{E}(T|D_1) - \mathbb{E}(T|X_1)\big)_{mm} = \mathbb{V}\big(\mathbb{E}(S_{\ell_n}|X_1)(Y_{1m} - \mathbb{E}(Y_{1m}|X_1)\big)_{mm} \approx \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]\mathbb{E}\big[\big(Y_{1m} - \mathbb{E}(Y_{1m}|X_1)\big)^2\big] = \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]Var(Y_m|X_1 = x), \end{aligned} \tag{44}$$

and

$$\mathbb{V}\big(\mathbb{E}(T|X_1)_{mm} = \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]Var\big(\mathbb{E}(Y_m|X_1 = x)\big). \tag{45}$$

where S_{ℓ_n} is the indicator function and equals one if $X_1 \in \ell_n(x, \Pi)$, and zero otherwise. Note that the first approximation stems from honesty (Assumption 5). Specifically, since we use two independent samples for finding optimal splitting variables and estimation, respectively, we can separate variance of S_{ℓ_n} and a given outcome.

The off-diagonal terms equal to:

$$\mathbb{V}\big(\mathbb{E}(T|D_i) - \mathbb{E}(T|X_1)\big)_{mm'} = \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]\mathbb{E}\big[(Y_{1m} - \mathbb{E}(X_1))(Y_{1m'} - \mathbb{E}(Y_{1m'}|X_1))\big]. \tag{46}$$

According to Assumption 6, the variance of each outcome variable is bounded away from zero. Cauchy-Schwarz inequality implies that $|Cov(Y_{im}, Y_{im'}|X_1)|$ is also bounded away from zero [6].

$$|Cov(Y_{1m}, Y_{1m'}|X_1 = \gamma)| \leq \sqrt{Var(Y_{1m}|X_1 = \gamma)Var(Y_{1m'}|X_1 = \gamma)}.$$

Athey and Wager (2018) show that

$$\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big] \geq \frac{(p-1)!}{2^{p+1}\log^p(s)} \cdot \frac{1}{ks}, \tag{47}$$

[6] An alternative argument is to notice that the term in the integrand consists of multiples of the first and second moments of the outcome variables Y_{1m} and $Y_{1m'}$. Since these moments are continuous, they are bounded. Thus, their expectation is also bounded.

where k is the minimum number of observations in a given terminal node. Combining (42) and (47) yields the order of diagonal and off-diagonal terms:

$$\mathbb{V}\big(\mathring{T}\big)_{mm} = o\bigg(\frac{1}{\log^p(s)}\bigg), \text{ and } \mathbb{V}\big(\mathring{T}\big)_{mm'} = o\bigg(\frac{1}{\log^p(s)}\bigg). \tag{48}$$

Now we prove that $\frac{s}{N} tr\Bigg(\big(\mathbb{V}(\mathring{T})\big)^{-1}\mathbb{V}(T)\Bigg) \to 0$ in a more general framework. Consider, we have two square matrices C and B with diagonal (c_{ii}, b_{ii}) and non-diagonal terms (c_{ij}, b_{ij}), respectively. Moreover, they have the following properties:

$$1.\ b_{ii} \geq \eta \text{ for some } \eta \in \mathbb{R}^+ \text{ and for all } i = 1, \ldots d, \tag{49}$$

$$2.\ c_{ii} \geq \frac{b_{ii}}{\log(N)}, \tag{50}$$

$$3.\ c_{ij} = o\bigg(\frac{1}{\log(N)}\bigg). \tag{51}$$

Then we show that $\frac{s}{N} tr(C^{-1}B) \to 0$. Recall that the Leibniz formula for the determinant is given as follows:

$$det(C) = \sum_{\pi}\Bigg(\operatorname{sgn}(\pi)\prod_{i=1}^{d} c_{i,\pi_i}\Bigg), \tag{52}$$

where π is a permutation function that reorders the set $\{1, \ldots, d\}$. Diagonal and off-diagonal terms are on the same order, their product is also on the same order. Therefore, $det(C)$ is asymptotically equivalent to either $\prod_{i=1}^{d} c_{ii}$ or $\prod_{i=1}^{d} c_{ij}$ where $i \neq j$. For simplicity, we keep the notation that $det(C) \sim^a \prod_{i=1}^{d} c_{ii}$, where " $\sim^a$ " denotes asymptotic equivalence. Based on Cramer's rule, we can write i-th diagonal term of the inverse of C:

$$(C^{-1})_{ii} = \frac{det(C_{-i})}{det(C)}.$$

C_{-i} is the matrix where we remove the i-th row and the i-th column. By the same argument, $det(C_{-i}) \sim^a \prod_{j=1}^{d-1} c_{jj}$. Then we end up with:

$$(C^{-1})_{ii} \sim^a \frac{\prod_{j=1}^{M-1} c_{jj}}{\prod_{i=1}^{M} c_{ii}} = \frac{1}{c_{ii}}.$$

The i-th diagonal entry of the matrix

$$(C^{-1}B)_{ii} = (c^{-1})_{ii}b_{ii} + \sum_{j\neq i}(c^{-1})_{ij}b_{ji} \sim^{a} \frac{b_{ii}}{c_{ii}} \leq \log(N).$$

The last equality follows from Property 2 in (50). Therefore, the trace of $(C^{-1}B)$ is also on the order of $\log(N)$. We take the limit of $\frac{s}{N}tr(C^{-1}B)$, where $s = N^{\beta}$ and $\beta < 1$. L'Hôpital's rule yields:

$$\lim_{N\longrightarrow\infty} \frac{s}{N}\log(N) = \lim_{N\longrightarrow\infty} \frac{1}{(1-\beta)N^{1-\beta}} \rightarrow 0. \tag{53}$$

The proof is complete by letting $C = \big(\mathbb{V}(\mathring{T})\big)^{-1}$ and $B = \mathbb{V}(T)$.

The proof is equivalent for the generalized random forest for treatment effect estimation. Athey and Wager (2018) show that,

$$\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big] \geq \frac{(p-1)!}{2^{p+1}\log^{p}(s)} \cdot \frac{\epsilon}{ks}, \tag{54}$$

where, ϵ is a constant from Assumption 9. This does not change the results of the proofs, as the order of $\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]$ is still $o\bigg(\frac{1}{\log^{p}(s)}\bigg)$.

□