Estimation and Inference for Causal Explainability

Weihan Zhang WEIHAN_ZHANG2001@BERKELEY.EDU
Division of Biostatistics, University of California, Berkeley

Zijun Gao ZIJUNGAO@MARSHALL.USC.EDU
Department of Data Science and Operations, University of Southern California

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Understanding how much each variable contributes to an outcome is a central question across disciplines. A causal view of explainability is favorable for its ability in uncovering underlying mechanisms and generalizing to new contexts. Based on a family of causal explainability quantities, we develop methods for their estimation and inference. In particular, we construct a one-step correction estimator using semi-parametric efficiency theory, which explicitly leverages the independence structure of variables to reduce the asymptotic variance. For a null hypothesis on the boundary, i.e., zero explainability, we show its equivalence to Fisher's sharp null, which motivates a randomization-based inference procedure. Finally, we illustrate the empirical efficacy of our approach through simulations as well as an immigration experiment dataset, where we investigate how features and their interactions shape public opinion toward admitting immigrants.

Keywords: Explainability, Causality, ANOVA, Semi-parametric efficiency, Randomization inference

1. INTRODUCTION

"The truth is rarely pure and never simple." — Oscar Wilde. Outcomes typically rely on multiple factors rather than a single cause: customer ratings are influenced by various attributes of an product; biological traits are influenced by multiple genes (Fisher, 1918; Visscher et al., 2017; Watanabe et al., 2019; Sivakumaran et al., 2011). In the presence of multiple factors, it is practically relevant to quantify their explainability, that is how much a set of factors and their interactions account for the variability in the outcome.

Compared with association-based measures, causality-based explainability can illuminate mechanisms and transfer across contexts reliably, and is sometimes regarded as the foundation of scientific explanation (Pearl, 2009; VanderWeele, 2015). This motivates us to adopt the Causal ANOVA (Gao and Zhao, 2025), a set of quantities generalizing the functional ANOVA (Hooker, 2007) and Sobol indices (Sobol', 2001) to the counterfactual perspective, to measure the explainability. For the statistical analysis of causal ANOVA quantities, prior work has primarily focused on settings in which an oracle[1], typically a black-box machine learning model, is available to evaluate the outcome at arbitrarily chosen input values, with the resulting explainabilities used to assess feature importance. However, in many domains including social science and genetics, researchers rely on collected data samples and have no access to such oracles.

1. For "oracle", we refer to an outcome-generating mechanism: given a user-chosen input vector (e.g., treatment values), one can evaluate the corresponding potential outcome as a function value. This is the standard setup in global sensitivity analysis for black-box simulators, such as Sobol/functional-ANOVA decompositions (Sobol', 2001).

Semi-parametric methods target the estimation and inference for finite-dimensional parameters while allowing nuisance components to be non-parametric, and operate on pre-collected datasets requiring no oracle. This aligns with our scope: a scalar target (explainability summarized by a number), minimal modeling assumptions, and no oracle access. Therefore, we develop semi-parametric methods for the estimation and inference of Causal ANOVA quantities. Particularly, we derive the efficient influence function and construct the associated one-step correction estimator, and explicitly show how the independence structure in Causal ANOVA can be leveraged to gain efficiency. Degenerate nulls (the parameter of interest lies on the boundary of its space) are a known challenge in semi-parametric statistics (Verdinelli and Wasserman, 2024), and we design a valid randomization-based procedure for such nulls in contrast to the common practice based on sample splitting or noise injection.

Contributions. We investigate the estimation and inference for the explainability attributable to individual factors and their interactions through the lens of semi-parametric statistics.

- Our approach can be used to identify influential factors and interactions by testing for non-zero explainability, and further quantify their importance by providing confidence intervals. Applied to an immigration-experiment dataset, multiple factors are shown to exhibit nonzero causal explainability for public opinions toward admitting immigrants, and the interaction between *Job Plan* and *Job Experience* is found to be influential as well.
- We derive the efficient influence function that explicitly leverages the structure, i.e., (conditional) independence, in the probability model and construct the associated one-step correction estimator, which attains smaller asymptotic variance compared to the counterpart that ignores the structure. The benefits of exploiting independence structure as well as our derivation extend to other applications with independent components, such as factorial A/B tests.
- We propose a randomization-based procedure to test whether the null is degenerate, which requires neither sample splitting nor noise injection. Further combined with the confidence interval using the asymptotic distribution of one-step correction estimators (valid under non-degenerate nulls), we introduce a sequential procedure that returns valid confidence sets regardless of the null degeneracy.

Organization. In Section 2, we review Causal ANOVA, semi-parametric estimation theory, and randomization test. In Section 3, we exploit the independence structure in the probability model to derive a more efficient influence function and provide the associated one-step estimator. In Section 4, we derive the one-step estimator's asymptotic distribution under non-degenerate nulls, introduce a randomization-based test for degenerate nulls, and combine them into a sequential procedure. In Section 5, Section 6, we demonstrate our procedure in simulations and a real world dataset. Proofs, additional simulations are provided in the appendix.

Notation. Let ξ denote the causal estimand of interest. Let Y be the outcome, $\boldsymbol{W} = (W_1, \ldots, W_k)$ be the vector of factors; for $\mathcal{S} \subseteq [K]$, write $\boldsymbol{W}_{\mathcal{S}}$ for the corresponding subvector. Let $\mathbb{P}$, μ and υ denote the true joint distribution of $(Y, \boldsymbol{W})$ and their expectation operators $\mathbb{E}[Y \mid \boldsymbol{W}]$ and $\mathbb{E}[Y^2 \mid \boldsymbol{W}]$. Write $\mathbb{P}_{\boldsymbol{W}}$ for the marginal distribution of $\boldsymbol{W}$, $\mathbb{P}_{W_k}$ for the marginal of W_k, and $\mathbb{P}_{Y|\boldsymbol{W}}$ for the conditional distribution of Y given $\boldsymbol{W}$. Their estimated counterparts are denoted by $\widehat{\mathbb{P}}$ and $\widehat{\mu}$, with $\widehat{\mathbb{P}}_{\boldsymbol{W}}$, $\widehat{\mathbb{P}}_{W_k}$, and $\widehat{\mathbb{P}}_{Y|\boldsymbol{W}}$ defined analogously. For a probability model $\mathcal{P}$ and $\mathbb{P} \in \mathcal{P}$, let $\dot{\mathcal{P}}_{\mathbb{P}}$ denote

the tangent space at $\mathbb{P}$ (subscript $\mathbb{P}$ dropped when unambiguous). We write $\varphi^{\xi}_{\text{IF}}$, $\varphi^{\xi}_{\text{EIF}}$, or simply φ^{ξ} for influence functions of quantity ξ (superscript ξ dropped when unambiguous). For a more comprehensive summary of the notations, see Table 3.

2. PROBLEM FORMULATION

2.1. Causal ANOVA

Causal ANOVA is set in the potential outcome framework (Rubin, 1974), with K randomized treatments $W_k \in \mathcal{W}_k$ and n units. Each unit i has potential outcomes $\boldsymbol{Y}_i(\cdot)$, but only $Y_i = Y_i(\boldsymbol{W}_i)$ is observed for the realized assignment $\boldsymbol{W}_i = (W_{1,i}, \ldots, W_{K,i})$. To ensure that Causal ANOVA quantities generalize to future observations, we assume the observed units are drawn i.i.d. from a super-population model. Formally,

$$\boldsymbol{W}_i \sim \mathbb{P}_{\boldsymbol{W}},\ \boldsymbol{Y}_i(\cdot) \sim \mathbb{P}_{Y(\cdot)},\ \boldsymbol{W}_i \perp\!\!\!\perp Y_i(\cdot), \tag{1}$$

independently across units, where $\mathbb{P}_{\boldsymbol{W}}$ is the (unknown) distribution of treatment assignments and $\mathbb{P}_{Y(\cdot)}$ is the unknown distribution of potential outcomes.

Causal ANOVA defines explainability via contrasts among potential outcomes, i.e., comparing factuals and counterfactuals. While Causal ANOVA can accommodate any set of factors $\boldsymbol{W}$ that respect a directed acyclic graph (DAG), identification of the explainabilities under dependent factors generally requires additional assumptions (e.g., comonotonicity of potential outcomes). As a result, in this paper we focus on independent treatments specified below.

Assumption 1 (Independent treatments) *The factors W_k, $k \in [K]$, are mutually independent.*

Let $\mathcal{P}_{\text{ind}}$ denote the model satisfying Assumption 1, a proper subset of the non-parametric model $\mathcal{P}_{\text{np}}$ in (1). Assumption 1 is satisfied in many experiment designs, for example, conjoint analysis randomizes the value of attributes independently. More generally, settings with block-wise independence of treatment groups can be analyzed in a similar manner, where each block of treatments is regarded as a single combined factor. In Appendix A, we show Assumption 1 can be relaxed to the conditional independence: W_k are mutually independent given some covariates $\boldsymbol{X}$.

Let $\boldsymbol{W}'$ be an independent copy of $\boldsymbol{W}$. Under Assumption 1, Causal ANOVA uses the variance of the difference obtained by replacing $\boldsymbol{W}_{\mathcal{S}}$ with $\boldsymbol{W}'_{\mathcal{S}}$ to define the total explainability of $\boldsymbol{W}_{\mathcal{S}}$, $\mathcal{S} \subseteq [K]$.

Definition 1 (Total explainability) *For $\mathcal{S} \subseteq [K]$,*

$$\xi(\vee_{k\in\mathcal{S}} W_k) := \frac{\mathsf{Var}\left(Y(\boldsymbol{W}) - Y(\boldsymbol{W}'_{\mathcal{S}}, \boldsymbol{W}_{-\mathcal{S}})\right)}{2\mathsf{Var}\left(Y(\boldsymbol{W})\right)}. \tag{2}$$

The explainability of interaction is defined as the variance of a difference-in-difference contrast.

Definition 2 (Interaction explainability) *For $k \neq k' \in [K]$,*

$$\xi(W_k \wedge W_{k'}) := \frac{\mathsf{Var}\left(I(W'_k, W'_{k'})\right)}{4\,\mathsf{Var}\left(Y(\boldsymbol{W})\right)}, \tag{3}$$

where $I(W'_k, W'_{k'}) = Y(W'_k, W'_{k'}) - Y(W'_k, W_{k'}) - Y(W_k, W'_{k'}) + Y(W_k, W_{k'})$.

Definition 2 defines the first-order interaction, and can be generalized to higher-order interactions. However, we omit them due to the reduced usefulness and increased complexity.

When Y is not fully determined by $\boldsymbol{W}$, the Causal ANOVA quantities are generally not identifiable. The non-parametric structural equation model with additive, independent errors, a.k.a. NPSEM-IE (Pearl, 2009), offers a useful balance between identifiability and generality.

Assumption 2 (NPSEM-IE) *The outcome satisfies $Y = f_Y(\boldsymbol{W}) + E_Y$ for some function f_Y and an error term E_Y independent of $\boldsymbol{W}$.*

NPSEM-IE is a commonly used model for causal discovery and inference (Hoyer et al., 2008; Peters et al., 2014). We discuss its extensions in Appendix A.2.

Proposition 1 *Under Assumption 1 and Assumption 2, for any $\mathcal{S} \subseteq [K]$, Definition 1 is identifiable and admits the form,*

$$\xi(\vee_{k\in\mathcal{S}} W_k) = \frac{\mathbb{E}[\mathbb{E}^2[Y \mid \boldsymbol{W}]]}{\mathsf{Var}(Y)} - \frac{\mathbb{E}\left[\mathbb{E}^2[Y \mid \boldsymbol{W}_{-\mathcal{S}}]\right]}{\mathsf{Var}(Y)}. \tag{4}$$

By the inclusion-exclusion principle for causal explainabilities, that is $\xi(W_k \wedge W_{k'}) = \xi(W_k) + \xi(W_{k'}) - \xi(W_k \vee W_{k'})$, we obtain an analogue of Proposition 1 for the interaction explainabilities.

Corollary 1 *Under the assumption of Proposition 1, for $k \neq k' \in [K]$, Definition 2 is identifiable and admits the form,*

$$\begin{aligned}\xi(W_k \wedge W_{k'}) =& \frac{\mathbb{E}[\mathbb{E}^2[Y \mid \boldsymbol{W}]]}{\mathsf{Var}(Y)} - \frac{\mathbb{E}\left[\mathbb{E}^2[Y \mid \boldsymbol{W}_{-k}]\right]}{\mathsf{Var}(Y)} \\ & - \frac{\mathbb{E}\left[\mathbb{E}^2[Y \mid \boldsymbol{W}_{-k'}]\right]}{\mathsf{Var}(Y)} + \frac{\mathbb{E}\left[\mathbb{E}^2[Y \mid \boldsymbol{W}_{-\{k,k'\}}]\right]}{\mathsf{Var}(Y)}.\end{aligned} \tag{5}$$

2.2. Influence function and one-step correction

Influence functions are critical in semi-parametric inference for constructing approximately unbiased and efficient estimators robust to nuisance estimation (Bickel et al., 1993). Let the quantity of interest ξ be some functional of $\mathbb{P}$, e.g., the causal explainability for a set of factors. Let $\mathcal{P}$ be a class of probabilities for the law $\mathbb{P}$ of $(\boldsymbol{W}_i, Y_i)$, and define its tangent space $\dot{\mathcal{P}}$ as the closure of the linear span of its scores. Suppose the functional ξ admits the von Mises expansion

$$\xi(\mathbb{P}_\varepsilon) - \xi(\mathbb{P}) = \int \varphi(\boldsymbol{w}, y; \mathbb{P}_\varepsilon)\, d(\mathbb{P}_\varepsilon - \mathbb{P})(\boldsymbol{w}, y) + R_2(\mathbb{P}_\varepsilon, \mathbb{P}), \tag{6}$$

for a perturbed distribution $\mathbb{P}_\varepsilon$ of $\mathbb{P}$, then $\varphi(\boldsymbol{w}, y; \mathbb{P})$ is a mean-zero, finite-variance function satisfying $\int \varphi(\boldsymbol{w}, y; \mathbb{P}) d\mathbb{P}(\boldsymbol{w}, y) = 0$, typically called an influence function. Influence functions are not unique; however, the projection of any influence function onto the tangent space $\dot{\mathcal{P}}$ is unique, which is called the *efficient influence function*, denoted by $\varphi_{\text{EIF}}(\boldsymbol{w}, y; \mathbb{P})$, and the semi-parametric efficiency bound of ξ is $\mathsf{Var}(\varphi_{\text{EIF}}(\boldsymbol{W}, Y; \mathbb{P}))$. We emphasize that for the same functional, the efficient influence function depends on the chosen model class $\mathcal{P}$. A smaller class $\mathcal{P}$ yields a smaller tangent space $\dot{\mathcal{P}}$ and hence likely a lower semi-parametric efficiency bound.

Influence functions provide a principled way to correct plug-in estimators and construct less biased, more efficient estimators. Particularly, let $\widehat{\mathbb{P}}$ be an estimator of the true distribution $\mathbb{P}$. A

natural plug-in estimator of $\xi(\mathbb{P})$ is $\xi(\widehat{\mathbb{P}})$. Given an influence function φ, the associated one-step corrected estimator is

$$\widehat{\xi} = \xi(\widehat{\mathbb{P}}) + \mathbb{P}_n\widehat{\varphi}, \quad \mathbb{P}_n\widehat{\varphi} := \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{W}_i, Y_i; \widehat{\mathbb{P}}). \tag{7}$$

If $\xi(\widehat{\mathbb{P}})$ is regular and efficient, then the one-step corrected estimator $\widehat{\xi}$ is also efficient, in the sense of attaining the semi-parametric efficiency bound.

We also review the literature on degenerate nulls (the parameter of interest lies on the boundary of the parameter space). For degenerate nulls, the influence functions of the estimator vanish. As a result, the estimator typically converges to a degenerate distribution, which can not be used to construct valid confidence intervals. Currently there are three major approaches to addressing the above problem (Dai et al., 2022; Williamson et al., 2023; Verdinelli and Wasserman, 2024). As noted by Verdinelli and Wasserman (2024), all these approaches rely on $O(n^{-1/2})$-level expansions to preserve validity under the null, at the cost of efficiency. In Hudson (2023), an estimator is proposed, which converges at rate n^{-1} under the null and at rate $n^{-1/2}$ away from the null. However, the method does not yield valid confidence intervals in the null hypothesis's neighborhood.

2.3. Fisher's randomization test

Fisher's randomization test originally evaluates a test statistic over the randomization distribution induced by the assignment mechanism under some sharp null (Fisher, 1918). Explicitly, given a test statistic T, we first compute $T(Y, \boldsymbol{W})$ under the observed assignment $\boldsymbol{W}$. Under the sharp null, we impute the potential outcomes under the partially sharp null, draw assignments $\boldsymbol{W}^{(b)}$ from the known assignment mechanism, evaluate $T(Y, \boldsymbol{W}^{(b)})$ to form the randomization distribution, and compute the p-value

$$P := \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{\, T(Y, \boldsymbol{W}^{(b)}) \geq T(Y, \boldsymbol{W})\,\}}{1 + B}. \tag{8}$$

With any $B < \infty$, P is always a finite-sample valid p-value in the sense that $\mathbb{P}(P \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ (Phipson and Smyth, 2010). In Causal ANOVA, the degenerate null (explainability being zero) implies Fisher's sharp null (Section 4), motivating the use of randomization-based inference.

3. ESTIMATION

We build on the semi-parametric efficiency theory to derive the influence function for the Causal ANOVA quantities (1) and then incorporate it into the standard cross-fitting procedure (Algorithm 1) (Schick, 1986; Klaassen, 1987) to obtain the one-step correction estimator (Pfanzagl, 1990; Bickel et al., 1993). Among the one-step correction estimators satisfying the asymptotic distribution established in Section 4 below, the one with a smaller asymptotic variance, i.e., higher efficiency, is preferred. In particular, the influence function that yields an estimator attaining the semi-parametric efficiency bound regarding a model $\mathcal{P}$ ($\mathcal{P}$ denoting a class of candidate joint distributions of $(Y, \boldsymbol{W})$) is optimal and referred to as the efficient influence function (for $\mathcal{P}$).

According to Proposition 1 and Corollary 1, for the total or interaction explainabilities, it suffices to estimate quantities of the form $\mathbb{E}\left[\mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}]^2\right]/\mathsf{Var}\,(Y)$, $\mathcal{S} \subseteq [K]$, which is pathwise differentiable (Williamson et al., 2021).

3.1. Baseline estimator

To begin with, consider the non-parametric model $\mathcal{P}_{\text{np}}$ for $(Y, \boldsymbol{W})$. Under this model, the efficient influence function for $\mathbb{E}[\mathbb{E}(Y \mid \boldsymbol{W}_{-\mathcal{S}})^2]/\mathsf{Var}(Y)$ takes the form (derivation in Appendix D.1),

$$\varphi_{\text{IF}}(Y, \boldsymbol{W}; \mathbb{P}) := \frac{2Y\mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}] - \mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}]^2}{\mathsf{Var}\left(Y\right)} - \mathbb{E}\left[\mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}]^2\right]\left(\frac{Y - \mathbb{E}\left[Y\right]}{\mathsf{Var}\left(Y\right)}\right)^2. \tag{9}$$

3.2. Estimator leveraging independence

We leverage the independence among W_k (1) to refine φ_{IF} in (9) to gain efficiency. The key step is to project φ_{IF} onto the tangent space of the restricted model $\dot{\mathcal{P}}_{\text{ind}}$, a proper subspace of $\dot{\mathcal{P}}_{\text{np}}$, to get a more efficient influence function φ_{EIF}. The difference $\varphi_{\text{IF}} - \varphi_{\text{EIF}}$, while may be influential under $\mathcal{P}_{\text{np}}$, has zero influence on the target estimand when W_k are independent, and variation along these directions merely inflates the variability. By removing them, we do not lose any information of the target estimand under the independence (1) and achieve a reduction in variance. Remark 1 provides an alternative explanation that demonstrates the projection procedure and the resulting efficiency gain. The next result shows the efficient influence function incorporating the independence (1).

Proposition 2 *Under the assumption of Proposition 1, for any $\mathcal{S} \subseteq [K]$, the efficient influence function of $\mathbb{E}\left[\mathbb{E}\left[Y \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2\right]/\mathsf{Var}\left(Y\right)$ for distribution $\mathbb{P}$ of $(Y, \boldsymbol{W})$ with independent treatments, i.e., $\mathbb{P}_{\boldsymbol{W}} = \prod_{k\in[K]} \mathbb{P}_{W_k}$, takes the form*

$$\begin{aligned}\varphi_{EIF}(Y, \boldsymbol{W}; \mathbb{P}) =& \frac{2\left(Y - \mathbb{E}[Y \mid \boldsymbol{W}]\right)\cdot\mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}]}{\mathsf{Var}\left(Y\right)} + \frac{2\sum_{k\in\mathcal{S}}\mathbb{E}\left[\mathbb{E}[Y \mid \boldsymbol{W}]\cdot\mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}] \mid W_k\right]}{\mathsf{Var}\left(Y\right)} \\ &+ \frac{\sum_{k\in-\mathcal{S}}\mathbb{E}\left[\mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \mid W_k\right]}{\mathsf{Var}\left(Y\right)} - \frac{(2|\mathcal{S}| + |-\mathcal{S}|)\cdot\mathbb{E}\left[\mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}]^2\right]}{\mathsf{Var}(Y)} \\ &- \frac{\varphi_{EIF}^{\mathsf{Var}(Y)}(Y, \boldsymbol{W}; \mathbb{P})\cdot\mathbb{E}\left[\mathbb{E}\left[Y \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2\right]}{\mathsf{Var}\left(Y\right)^2},\end{aligned} \tag{10}$$

$$\begin{aligned}\varphi_{EIF}^{\mathsf{Var}(Y)}(Y, \boldsymbol{W}; \mathbb{P}) =& \left(Y - \mathbb{E}\left[Y\right]\right)^2 - \mathsf{Var}\left(Y \mid \boldsymbol{W}\right) - \left(\mathbb{E}[Y \mid \boldsymbol{W}] - \mathbb{E}[Y]\right)^2 \\ &+ \sum_{k\in[K]}\mathbb{E}\left[\left(Y - \mathbb{E}[Y]\right)^2 \mid W_k\right] - |K|\cdot\mathsf{Var}\left(Y\right).\end{aligned} \tag{11}$$

The proof of Proposition 2 relies on explicitly characterizing the tangent space $\dot{\mathcal{P}}_{\text{ind}}$ under (1) and then performing the projection. Details are provided in Appendix D. In Section 5, we demonstrate the efficiency gain of φ_{EIF} over φ_{IF} in simulated datasets.

Remark 1 *According to Proposition 1. in Chapter 3 of Bickel et al. (1993), the efficient influence function can be decomposed as $\varphi_{EIF} = \varphi - \Pi\{\varphi \mid \dot{\mathcal{P}}_{\text{ind}}^{\perp}\}$, where φ is any influence function and φ_{EIF} is orthogonal to $\Pi\{\varphi \mid \dot{\mathcal{P}}_{\text{ind}}^{\perp}\}$. The estimator based on φ_{IF}, φ_{EIF} has asymptotic variance $\mathsf{Var}(\varphi_{IF})$, $\mathsf{Var}(\varphi_{EIF})$, respectively, for any $\mathbb{P}$, $\mathsf{Var}(\varphi_{IF}) = \mathsf{Var}(\varphi_{EIF}) + \mathsf{Var}\left(\Pi\{\varphi \mid \dot{\mathcal{P}}_{\text{ind}}^{\perp}\}\right) \geq \mathsf{Var}(\varphi_{EIF})$.*

3.3. Estimators leveraging additional structure

We briefly discuss the possibility of further improving the efficiency of φ_{EIF} in (10) by leveraging additional structures of $\mathcal{P}$. If we further include the additive and independent noise assumption (2), known as the location-shift regression model (Tsiatis, 2006), φ_{EIF} can be refined accordingly and the improved influence function requires the distribution of the error term E_Y (in (2)). However, E_Y is unobserved, and its distribution is difficult to estimate with sufficient accuracy to ensure the asymptotic distribution in Section 4. Therefore, the efficiency gain from incorporating (2) is primarily conceptual rather than practical, and we therefore continue with φ_{EIF} in what follows (see Chapter 5 of Tsiatis (2006) for a more detailed discussion).

If $\mathbb{P}_{\boldsymbol{W}}$ is known, φ_{EIF} can be further improved, and we provide the derivation of this efficient influence function in Corollary 3. However, in many real-data scenarios, including the application in Section 6, the distribution $\mathbb{P}_{\boldsymbol{W}}$ is unavailable, and thus we regard φ_{EIF} as more relevant and useful.

4. INFERENCE

4.1. Asymptotic distribution of one-step correction estimator

We derive the asymptotic distribution of the one-step correction estimator in Algorithm 1 using φ_{IF} in (9) and φ_{EIF}) in Proposition 2, respectively. We define $\widehat{\varphi}_{\text{IF}}$ as the influence function with plug-in nuisance estimators, i.e. $\varphi_{\text{IF}}(Y, \boldsymbol{W}; \widehat{\mathbb{P}})$. Similarly for $\widehat{\varphi}_{\text{EIF}}$.

Proposition 3 (One-step correction estimator based on φ_{IF}) *We assume the following.*

(C1) (Boundedness) There exists $C < \infty$ such that $|Y|$, $|\widehat{\mu}|$, and $|\mu| \leq C$, a.s..

(C2) (Consistency of nuisance estimators) $\|\widehat{\mu} - \mu\| = o_{\mathbb{P}}(1)$, $\|\widehat{\mathbb{P}}_{\boldsymbol{W}} - \mathbb{P}_{\boldsymbol{W}}\| = o_{\mathbb{P}}(1)$.

(C3) (Convergence rate of nuisance estimators) $\|\widehat{\mu} - \mu\| = o_{\mathbb{P}}(n^{-1/4})$, $\|\widehat{\mathbb{P}}_{\boldsymbol{W}} - \mathbb{P}_{\boldsymbol{W}}\| = o_{\mathbb{P}}(n^{-1/4})$.

The following asymptotic approximation holds

$$\sqrt{n}\big(\widehat{\xi}_{IF} - \xi\big) = \sqrt{n}\,\mathbb{P}_n\widehat{\varphi}_{IF} + o_{\mathbb{P}}(1),$$

which implies $\widehat{\xi}_{IF}$ is a regular and asymptotic linear estimator. Moreover, if $\mathsf{Var}(\widehat{\varphi}_{IF}) < \infty$, then

$$\sqrt{n}\big(\widehat{\xi}_{IF} - \xi\big) \xrightarrow{d} \mathcal{N}\big(0, \mathsf{Var}(\varphi_{IF})\big).$$

Proposition 4 (One-step correction estimator based on φ_{EIF}) *Under Assumption 1 and the assumptions of Proposition 3 where the condition on the joint distribution estimator $\|\widehat{p}_{\boldsymbol{W}} - p_{\boldsymbol{W}}\|$ is replaced by conditions on the marginal distribution estimators $\|\widehat{p}_{W_k,-l} - p_{W_k}\|$, $k \in [K]$, and further assume $\|\widehat{\upsilon} - \upsilon\| = o_{\mathbb{P}}(n^{-1/4})$ as well as*

(C4.) (Positivity) There exists $\underline{c} > 0$ such that $\inf_k p_{\boldsymbol{w}_k} \geq \underline{c}$ and $\inf_k \widehat{p}_{\boldsymbol{w}_k} \geq \underline{c}$,

(C5.) (Donsker class) There exists a nonrandom function class $\mathcal{F}$ that is $\mathbb{P}$–Donsker and satisfies

$$\mathbb{P}\Big((p_{W_k} - \widehat{p}_{W_k,-l})^2 \in \mathcal{F}\Big) \longrightarrow 1,$$

the following holds

$$\sqrt{n}\big(\widehat{\xi}_{EIF} - \xi\big) = \sqrt{n}\,\mathbb{P}_n\widehat{\varphi}_{EIF} + o_{\mathbb{P}}(1), \quad \sqrt{n}\big(\widehat{\xi}_{EIF} - \xi\big) \xrightarrow{d} \mathcal{N}\big(0, \mathsf{Var}(\varphi_{EIF})\big).$$

Proofs of Propositions 3 and 4 are provided in Appendix D. Condition (C3) implies that, for $\widehat{\xi}_{\text{IF}}$, $\widehat{\xi}_{\text{EIF}}$, the nuisance functions only need to be estimated at rate $o_{\mathbb{P}}(n^{-1/4})$ to guarantee the $o_{\mathbb{P}}(n^{-1/2})$ convergence of the estimator of the causal estimand. Comparing Proposition 3 and Proposition 4, there are several differences: (1) the asymptotic variance satisfies $\mathsf{Var}(\widehat{\varphi}_{\text{EIF}}) \leq \mathsf{Var}(\widehat{\varphi}_{\text{IF}})$, implying that $\widehat{\xi}_{\text{EIF}}$ is more efficient; (2) Proposition 4 relaxes the requirements (C2, C3) on the estimator of the joint distribution of all factors $\widehat{p}_{\boldsymbol{W}}$ to the marginal distributions $\widehat{p}_{W_k,-l}$, $k \in [K]$, which are easier to satisfy; (3) Proposition 4 additionally requires estimating υ (i.e., $\mathbb{E}[Y^2 \mid \boldsymbol{W} = \boldsymbol{w}]$), which enters $\varphi_{\text{EIF}}^{\mathsf{Var}(Y)}$ for $\mathsf{Var}(Y)$.

4.2. Non-degenerate null

For $\mathsf{Var}(\varphi) > 0$, where φ may denote either φ_{IF} or φ_{EIF}, Propositions 3 and 4 imply that we can construct the confidence interval for any significance level $\alpha \in (0,1)$ as

$$\Big[\widehat{\xi} - \Phi^{-1}(1-\alpha/2)\frac{\widehat{\sigma}}{\sqrt{n}},\ \widehat{\xi} + \Phi^{-1}(1-\alpha/2)\frac{\widehat{\sigma}}{\sqrt{n}}\Big]. \tag{12}$$

Corollary 2 *Under the assumptions of Proposition 4, if* $\mathsf{Var}(\varphi_{EIF}) > 0$*, then the confidence interval in* (12) *achieves asymptotic coverage. Similarly for* $\widehat{\xi}_{IF}$*.*

4.3. Degenerate null

When $\xi(W_{\mathcal{S}}) = 0$ for some $\mathcal{S} \subseteq [K]$, the leading term in the expansion of the estimation error vanishes (Propositions 3 and 4), and the asymptotic distribution cannot be directly used to construct confidence interval[2]. Here, we exploit the structure of the null and the independence (1) to introduce a randomization-based inference procedure. Compared with existing approaches to null degeneracy (Section 2.2), our method is finite-sample valid and avoids sample splitting or added noise (and preventing the associated loss of power). As the first step, we establish an equivalent statement of the zero total explainability null.

Proposition 5 *Assume Assumption 1, the null* $\xi(W_{\mathcal{S}}) = 0$ *is equivalent to*

$$Y(w_{\mathcal{S}}, w_{-\mathcal{S}}) = Y(w'_{\mathcal{S}}, w_{-\mathcal{S}}),$$

for any pairs $(w_{\mathcal{S}}, w_{-\mathcal{S}})$ *and* $(w'_{\mathcal{S}}, w_{-\mathcal{S}})$ *that both occur with positive probability in the support of* W *for discrete* W*, and almost surely for continuous* W*.*

Proof of Proposition 5 is provided in Appendix D. Proposition 5 implies that, holding components $W_{-\mathcal{S}}$ fixed, changing $W_{\mathcal{S}}$ does not affect the outcome Y. In finite samples, this translates to the partially sharp null (Fisher, 1918; Zhang and Zhao, 2023),

$$Y_i(w_{\mathcal{S}}, w_{-\mathcal{S}}) = Y_i(w'_{\mathcal{S}}, w_{-\mathcal{S}}), \quad \forall i \in [n]. \tag{13}$$

2. For interaction nulls, zero interaction explainability does not necessarily imply a zero influence function, and the limiting distribution remains nondegenerate and valid for inference. Example: consider independent $W_1, W_2, E_Y \sim \mathcal{N}(0,1)$ and $Y = W_1 + W_2 + E_Y$, we have $\xi(W_1)$, $\xi(W_2)$, $\xi(W_1 \vee W_2) > 0$ while $\xi(W_1 \wedge W_2) = 0$. Then $\varphi_{\text{IF}} = -2W_1W_2/\mathsf{Var}(Y)$ for $\xi(W_1 \wedge W_2)$, which does not vanish.

Under the null (13), for any hypothetical reassignment $W^*_{\mathcal{S},i}$, we can impute the corresponding potential outcome as $Y_i\big(W^*_{\mathcal{S},i}, W_{-\mathcal{S},i}\big) = Y_i$.

Randomization tests (Section 2.3) are standard for the partially sharp null (13). By Proposition 5, we propose to adopt randomization tests for degenerate nulls. Explicitly, we choose a computable test statistic that tends to be large under the alternative $\xi(W_{\mathcal{S}}) > 0$, such as an estimator of $\xi(W_{\mathcal{S}})$. We then generate $W^*_{\mathcal{S},i}$ from the same distribution as $W_{\mathcal{S},i}$ conditional on $W_{-\mathcal{S},i}$, impute the potential outcomes using (13), and recompute the test statistic based on $W^*_{\mathcal{S},i}$, the imputed outcomes, and compare the observed test statistic to those using $W^*_{\mathcal{S},i}$ to compute the p-value.

Note that standard randomization tests require knowing the distribution of $\boldsymbol{W}$ in order to generate $W^*_{\mathcal{S},i}$. However, the joint distribution of $\boldsymbol{W}$ may not be available in our setting. Fortunately, under the independence assumption (1) and the i.i.d. nature of the observations, we have

$$\big(\boldsymbol{W}_{\mathcal{S},\rho(i)}, \boldsymbol{W}_{-\mathcal{S},i}\big) \stackrel{d}{=} \big(\boldsymbol{W}_{\mathcal{S},i}, \boldsymbol{W}_{-\mathcal{S},i}\big), \tag{14}$$

where ρ denotes a permutation of the indices $i \in [n]$. Eq. (14) implies that, to perform the randomization test, we can generate $(\boldsymbol{W}^*_{\mathcal{S},i}, \boldsymbol{W}_{-\mathcal{S},i})$ by simply permuting $\boldsymbol{W}_{\mathcal{S},i}$ while keeping $\boldsymbol{W}_{-\mathcal{S},i}$ fixed. We summarize the procedure in Algorithm 2. The result below establishes the validity of the randomization test, and the proof is provided in Appendix D. We discuss extensions to other treatment assignment mechanisms, such as blocked design, in Appendix A.4.

Proposition 6 *Assume Assumption 1, under the null $\xi(\boldsymbol{W}_{\mathcal{S}}) = 0$, the randomization test in Algorithm 2 controls type I error.*

4.4. Sequential procedure adapting to null degeneracy

Since it is unknown a priori whether $\xi(W_{\mathcal{S}}) = 0$, a procedure is needed to distinguish between the degenerate and non-degenerate cases in order to apply the suitable inference procedure above. Explicitly, we propose to use the following sequential approach.

1. Run the randomization test Algorithm 2 for the degenerate null as in Section 4.3 at level α. If the null is not rejected, output $\{0\}$; otherwise, proceed to Step 2.

2. Compute Eq. (12) in Section 4.2 to construct a $(1-\alpha)$ confidence interval and output.

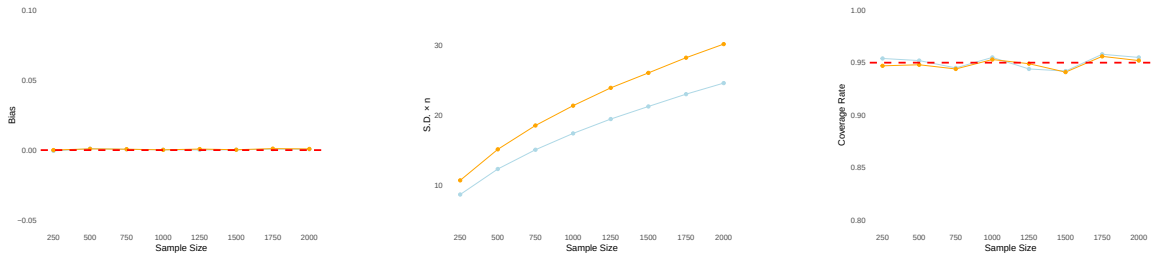
A detailed algorithm of the sequential procedure is provided in Appendix B. We emphasize that although two tests are conducted in the sequential procedure, there is no need to split the significance level α. This is because both tests concern the same quantity $\xi(W_{\mathcal{S}})$. The following result justifies the validity of the sequential procedure. The proof is provided in Appendix D.

Proposition 7 *The sequential procedure achieves asymptotically valid coverage under the assumptions of Corollary 2 and Proposition 6, and that the power of the randomization test converges to one as the sample size tends to infinity. If the power of the randomization test does not converge to one, replacing the first-step confidence interval $\{0\}$ with $[0, \infty)$ still guarantees the valid coverage.*

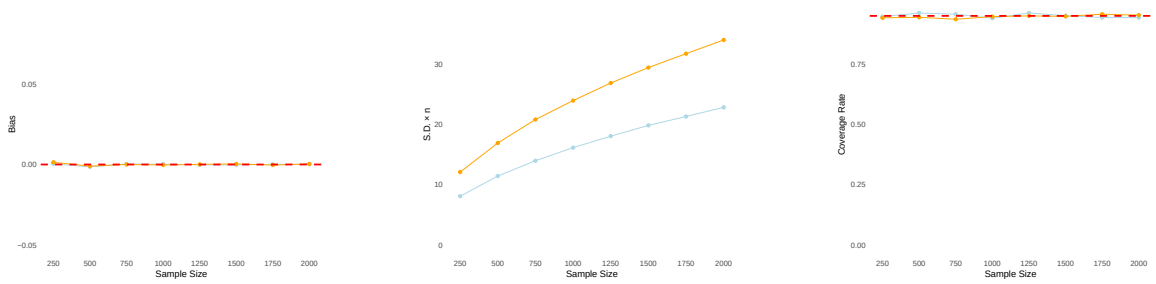
5. SIMULATION

We consider the following data generating mechanism and estimand: independent treatments $W_k \sim \mathcal{N}(0,1)$, $K = 3$, $E_Y \sim \mathcal{N}(0, 0.5)$. The response follows $Y = W_1^3 + W_2^3 + W_3^3 + W_1 W_2^2 + \sigma \cdot$

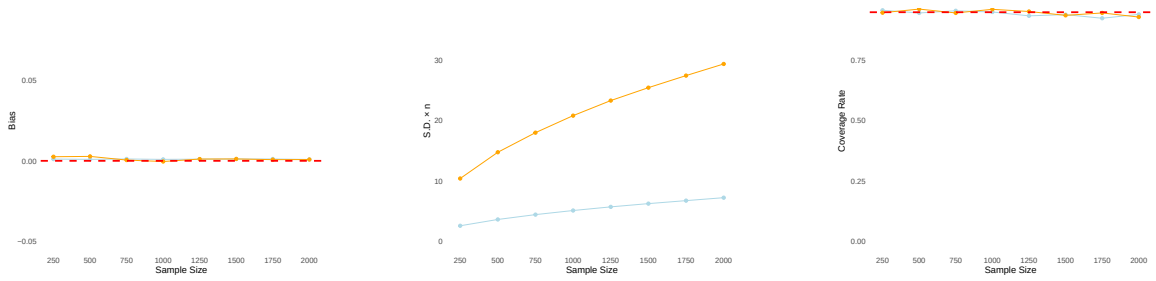
$W_1^2 W_3 + E_Y$. The estimands of interest are $\xi(W_3)$, $\xi(W_1 \vee W_3)$ and $\xi(W_1 \wedge W_3)$. We evaluate two different estimators based on influence functions: (1) one-step correction estimator accounting for the independence among $\boldsymbol{W}$ (φ_{EIF}-based estimator); (2) one-step correction estimator ignoring the independence among $\boldsymbol{W}$ (φ_{IF}-based estimator). To start, we implement the two methods using the true nuisance functions.



(a) $\xi(W_3)$; True nuisance functions; Varying sample size; Noise magnitude = 1



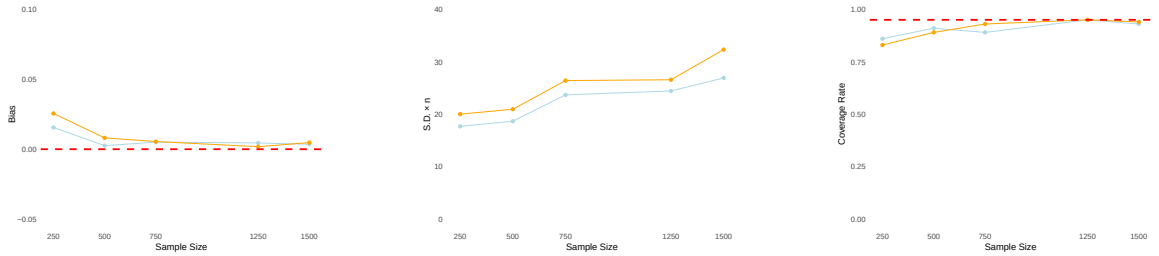
(b) $\xi(W_1 \vee W_3)$; True nuisance functions; Varying sample size; Noise magnitude = 1



(c) $\xi(W_1 \wedge W_3)$; True nuisance functions; Varying sample size; Noise magnitude = 1

Figure 1: Comparison of bias (left panel), estimated standard deviation times sample size (mid panel), and coverage rate (right panel; significance level $\alpha = 0.05$). φ_{EIF}-based method in blue, φ_{IF}-based method in gold. True nuisance functions are used. Results aggregated over 1000 trials.

In Figure 1, we vary the sample size. Both methods achieve exact coverage. The φ_{EIF}-based method exhibits lower variance, validating that exploiting independence can improve efficiency. We provide the results with varying noise magnitude in Appendix F.1. In Figure 2, we implement the two methods with nuisance functions estimated by the highly adaptive LASSO (HAL) and super learner (package HAL9001 and SUPER LEARNER) (van der Laan et al., 2007; Hejazi et al., 2020). Note that φ_{EIF} involves more nuisance functions than φ_{IF}, for example, φ_{EIF} additionally requires $\mathbb{E}\left[\mathbb{E}[Y \mid \boldsymbol{W}] \cdot \mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}] \mid W_k\right]$ for all $k \in [K]$. With estimated nuisance functions, the φ_{EIF}-



$\xi(W_3)$; Estimated nuisance functions; Varying sample size; Noise magnitude = 1

Figure 2: Comparison using estimated nuisance functions (details in the caption of Figure 1). Results aggregated over 100 trials.

based method still exhibits variance reduction, although these gains are less substantial than those obtained with true nuisance functions. In Appendix F.2, we provide the comparison for interaction explainabilities, where the nuisance estimation is even more complex and the φ_{IF}-based estimator appears competitive. We noticed that the efficient influence function requires more nuisance components, which would result in unstable finite-sample performance (especially under coverage) of the one-step bias-correcting estimator that could offset theoretical efficiency gains. We recommend using the nonparametric influence function (NIF, or IF mentioned in our paper) based one-step bias-correcting estimator in finite samples, especially when the practitioners are not so confident about their estimated nuisances, since it only contains two nuisances. As shown in Appendix F.2, the performance of the IF-based estimator is more robust compared to the φ_{EIF}-based estimator in finite samples using the estimated nuisances. The φ_{EIF}-based estimator is preferred when the nuisances are well-estimated to gain a narrower confidence interval.

6. REAL DATA ANALYSIS

Conjoint analysis is a factorial survey-based experiment in which participants choose between pairs of candidate profiles with several randomly selected characteristics. One of the most cited applications of conjoint analysis is Hainmueller et al. (2015), who examined the role of immigrants' attributes in shaping support for their admission to the United States. In the experiment, 1396 respondents were asked to act as immigration officials and to decide which of a pair of immigrants they would choose for admission. Respondents evaluated a total of five pairings. The attributes were then randomly chosen to form immigrant profiles. We apply the explainability metric and the associated estimation and inference tools proposed to this real dataset. We use the φ_{IF}-based estimator as it requires a simpler set of nuisance functions to estimate[3]. For the nulls of zero total explainability, we do not rule out the degenerate null issue, and thus apply the sequential procedure in Section 4.4. Since the data have a cluster structure, we treat the respondent as the independent unit: let $O_i = \{O_{ir}\}_{r=1}^{m_i}$ denote all observations from respondent i, assume $\{O_i\}_{i=1}^{n}$ are i.i.d. across i, and allow arbitrary dependence within i. We compute row-level influence-function contributions ϕ_{ir} (with cross-fitting done by respondent folds), and aggregate to respondent-level scores $U_i = \sum_{r=1}^{m_i} \phi_{ir}$. The cluster-robust variance for $\widehat{\xi} = \xi(\hat{\eta}) + N^{-1} \sum_{i,r} \phi_{ir}$

3. And is not susceptible to the degeneracy problem for the interaction nulls.

Table 1: Total explainability

Variable	Total explainability	95% CI
Gender	0.0012	$[-0.0003, 0.0026]$
Language	0.0319	$[0.0240, 0.0398]$
Job experience	0.0164	$[0.0109, 0.0219]$
Job plans	0.0884	$[0.0751, 0.1016]$
Prior trips to U.S.	0.0366	$[0.0275, 0.0457]$
Education & Profession	0.0579	$[0.0471, 0.0687]$
Origin & Application reason	0.0155	$[0.0101, 0.0210]$

Note: The permutation test for Gender returns a p-value of 1, failing to reject the null hypothesis.

Table 2: Interaction explainability

Factor 1	Factor 2	Interaction explainability	95% CI
Job plan	Prior trips to U.S.	-0.0002	$[-0.0032, 0.0029]$
Job plan	Job experience	0.0014	$[-0.0006, 0.0034]$
Job plan	Language	0.0010	$[-0.0017, 0.0038]$

is then $\widehat{\mathrm{Var}}(\widehat{\xi}) = \frac{1}{N^2} \cdot \frac{n}{n-1} \sum_{i=1}^{n} (U_i - \overline{U})^2$, $\overline{U} = \frac{1}{n} \sum_{i=1}^{n} U_i$, and we use $t_{n-1,\,1-\alpha/2}$ critical values for confidence intervals. Any randomization test must preserve the actual assignment mechanism. Under clustering, we resample/permutate $\boldsymbol{W}_{\mathcal{S}}$ within the respondent (and within blocks/strata if present), preserving task structure and any design constraints (e.g., paired-profile choice sets, restricted attribute combinations). We then recompute the test statistic on each resampled dataset using the same respondent-level cross-fitting and cluster-robust studentization. This yields finite-sample validity when the resampling scheme matches the true constrained randomization distribution for $\boldsymbol{W}_{\mathcal{S}}$. Table 1 shows that the total explainabilities of *Language*, *Job Experience*, *Job Plan*, *Prior Trips to the U.S.*, *Education & Profession*, and *Origin & Application* reason are all significantly positive, except *Gender*. Among these factors, *Job Plan*, *Education & Profession* exhibit the two highest explainabilities, suggesting an encouraging pattern that immigrants' skills and qualifications are seriously evaluated. The interaction explainability analysis results (Table 2) indicates no significant interactions for several variables. Appendix G provides a comparison with other existing methods.

7. DISCUSSION

In this work, we study the estimation and inference of causal ANOVA quantities, which quantify the explainability of individual factors and their interactions with respect to some outcome. Building on semi-parametric efficiency theory, we derive one-step corrected estimators both with and without incorporating independence structures across factors. The estimator that incorporates this structure achieves smaller asymptotic variance theoretically and when nuisance components can be accurately estimated, whereas the estimator that does not explicitly leverage independence has a simpler form and may exhibit greater robustness in small samples. To address the issue of null degeneracy, we develop a randomization-based inference procedure and integrate it with the semi-parametric analysis. Applied to a real immigration dataset, our methods identify multiple factors with nonzero explainability, as well as a significant interaction between *Job Plan* and *Job Experience*.

We outline a few future directions. When testing a set of factors and their interactions for nonzero explainability simultaneously, multiplicity arises, and it is of interest to explore proper adjustment. Causal ANOVA quantities exhibit a hierarchical structure (for example, if the total explainability of a factor is zero, then any interaction involving this factor must be zero), which can be exploited via closed testing or gatekeeping to improve power while maintaining error control. In addition, in sequential decision problems (Markov decision processes), the Markov property yields a natural conditional independence. A direction for future research is to investigate the influence functions incorporating this conditional independence to reduce estimator's asymptotic variance. Moreover, when factors are dependent, Causal ANOVA quantities are generally not point-identified. A natural direction is to characterize the identified set and extend the semi-parametric analysis to the estimation and inference for this set.

References

Vladimir I. Averbukh and Oleg Georgievich Smolyanov. The theory of differentiation in linear topological spaces. *Russian Mathematical Surveys*, 22(6):201, 1967.

Peter J. Bickel, Ya'acov Ritov, Chris A. J. Klaassen, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Springer, 1993.

Ben Dai, Xiaotong Shen, and Wei Pan. Significance tests of feature relevance for a black-box learner. *IEEE transactions on neural networks and learning systems*, 35(2):1898–1911, 2022.

R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

Zijun Gao and Qingyuan Zhao. Counterfactual explanability of black-box prediction models. In Biwei Huang and Mathias Drton, editors, *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pages 1174–1174. PMLR, 07–09 May 2025.

Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22(1):1–30, 2015.

Dae Woong Ham, Kosuke Imai, and Lucas Janson. Using machine learning to test causal hypotheses in conjoint analysis. *Political Analysis*, 32(3):329–344, 2024.

Nima S Hejazi, Jeremy R Coyle, and Mark J van der Laan. hal9001: Scalable highly adaptive lasso regression in R. *Journal of Open Source Software*, 2020.

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

Aaron Hudson. Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space. *arXiv preprint arXiv:2306.07492*, 2023.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pages 207–236, 2024.

Edward H. Kennedy, Sivaraman Balakrishnan, and Max G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48(4):2008 – 2030, 2020.

Chris AJ Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, 15(4):1548–1562, 1987.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2 edition, 2009.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Johann Pfanzagl. Estimation in semiparametric models. In *Estimation in Semiparametric Models: Some Recent Developments*, pages 17–22. Springer, 1990.

Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

Shonali Sivakumaran, Felix Agakov, Evropi Theodoratou, James G. Prendergast, Lina Zgaga, Teri Manolio, Igor Rudan, Paul McKeigue, James F. Wilson, and Harry Campbell. Abundant pleiotropy in human complex diseases and traits. *Proceedings of the National Academy of Sciences*, 108(21):13924–13929, 2011.

Ilya M. Sobol'. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3):271–280, 2001.

Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*, volume 4. Springer, 2006.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007.

Aad van der Vaart. On differentiable functionals. *The Annals of Statistics*, pages 178–204, 1991.

Aad W. van der Vaart and Jon A. Wellner. *Empirical Processes*, pages 127–384. Springer International Publishing, Cham, second edition, 2023.

Tyler J. VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015.

Tyler J. VanderWeele and Eric J. Tchetgen Tchetgen. Attributing effects to interactions. *Epidemiology*, 25(5):711–722, 2014.

Isabella Verdinelli and Larry Wasserman. Feature importance: A closer look at shapley values and loco. *Statistical Science*, 39(4):623–636, 2024.

Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

Kyoko Watanabe, Sven Stringer, Oleksandr Frei, Marija Umićević Mirkov, Christiaan A. de Leeuw, Tinca J. C. Polderman, Sophie van der Sluis, Ole A. Andreassen, Benjamin M. Neale, and Danielle Posthuma. A global view of pleiotropy and genetic architecture across complex traits. *Nature Genetics*, 51:1339–1348, 2019.

Brian D. Williamson, Peter B. Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021.

Brian D. Williamson, Peter B. Gilbert, Noah R. Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.

Yao Zhang and Qingyuan Zhao. What is a Randomization Test? *Journal of the American Statistical Association*, 0(0):1–15, 2023.

Appendix for "Estimation and Inference for Causal Explainability"

Organization. In Appendix A.1, we provide additional definitions of causal ANOVA quantities and discuss their hierarchical structure. In Appendix A.2, we discuss extensions of the NPSEM-IE model (2). In Appendix A.3, we extend the independence condition (1) to conditional independence. In Appendix A.4, we discuss extensions of standard randomization tests to various treatment assignment mechanisms. In Appendix B, we present additional algorithms. In Appendix C, we collect supplementary lemmas and corollaries. In Appendix D, we provide detailed proofs. In Appendix E, we give additional tables. In Appendix F, we report additional empirical studies.

Appendix A. Method extension

A.1. Interaction explainability

We define the interaction effect of a set of treatments iteratively.

Definition 3 (Interaction effect) *For $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^K$, let $I_{\emptyset,\boldsymbol{w}'}(\boldsymbol{w}) := Y(\boldsymbol{w}')$, $I_{k,\boldsymbol{w}'}(\boldsymbol{w}) := Y(\boldsymbol{W}_k, \boldsymbol{w}'_{-k}) - Y(\boldsymbol{w}')$ for $k \in [K]$. For $\mathcal{S} \subseteq [K]$,*

$$I_{\mathcal{S},\boldsymbol{w}'}(\boldsymbol{w}) := Y(\boldsymbol{w}_{\mathcal{S}}, \boldsymbol{w}'_{-\mathcal{S}}) - \sum_{\mathcal{S}' \subsetneq \mathcal{S}} I_{\mathcal{S}',\boldsymbol{w}'}(\boldsymbol{w}).$$

Equivalently, the interaction term can be represented as a linear combination of the potential outcomes evaluated at a combination of $\boldsymbol{w}_{\mathcal{S}'}, \boldsymbol{w}'_{-\mathcal{S}'}$ for $\mathcal{S}' \subseteq \mathcal{S}$ with coefficients in $\{1, -1\}$,

$$I_{\mathcal{S},\boldsymbol{w}'}(\boldsymbol{w}) = \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}-\mathcal{S}'|} Y(\boldsymbol{w}_{\mathcal{S}'}, \boldsymbol{w}'_{-\mathcal{S}'}).$$

The interaction effect of W_k and W'_k takes the form $I_{\{k,k'\},\boldsymbol{w}'}(\boldsymbol{w}) = Y(W_k, W_{k'}, \boldsymbol{w}'_{-\{k,k'\}}) - Y(W_k, \boldsymbol{w}'_{-k}) - Y(W_{k'}, \boldsymbol{w}'_{-k'}) + Y(\boldsymbol{w}')$. For $K = 2$ and binary W_1, W_2, VanderWeele and Tchetgen (2014) uses $I_{\{1,2\},[0,0]}([1,1])$ as the interactive effect of W_1 and W_2.

The interaction effect $I_{\mathcal{S},\boldsymbol{W}'}(\boldsymbol{W})$ of treatments in $\mathcal{S}$ only requires access to the independent copy of treatments in $\mathcal{S}$. The treatments outside $\mathcal{S}$ remain unchanged. We refer to this property as "self-sufficiency". The self-sufficiency property is desirable in the sense that the definition of $I_{\mathcal{S},\boldsymbol{W}'}$ is invariant to the specification of the complete set of treatments. Explicitly, let $\tilde{Y}(\boldsymbol{W}_{\mathcal{S}}) = Y(\boldsymbol{W}_{\mathcal{S}}, \boldsymbol{W}_{-\mathcal{S}})$, then $\tilde{I}_{\mathcal{S},\boldsymbol{W}'}$ defined based on $\tilde{Y}(\boldsymbol{W}_{\mathcal{S}})$ is equivalent to that using $Y(\boldsymbol{W}_{\mathcal{S}}, \boldsymbol{W}_{-\mathcal{S}})$.

We define the explainability of the interaction of a set of treatments $W_{\mathcal{S}}$, $\mathcal{S} \subseteq [K]$.

Definition 4 *Let $\boldsymbol{W}'$ be an independent copy of $\boldsymbol{W}$. For $\mathcal{S} \subseteq [K]$,*

$$\xi(\wedge_{k \in \mathcal{S}} W_k) = \frac{\mathsf{Var}\left(I_{\mathcal{S},\boldsymbol{W}'}(\boldsymbol{W})\right)}{2^{|\mathcal{S}|}\mathsf{Var}\left(Y\right)}. \tag{15}$$

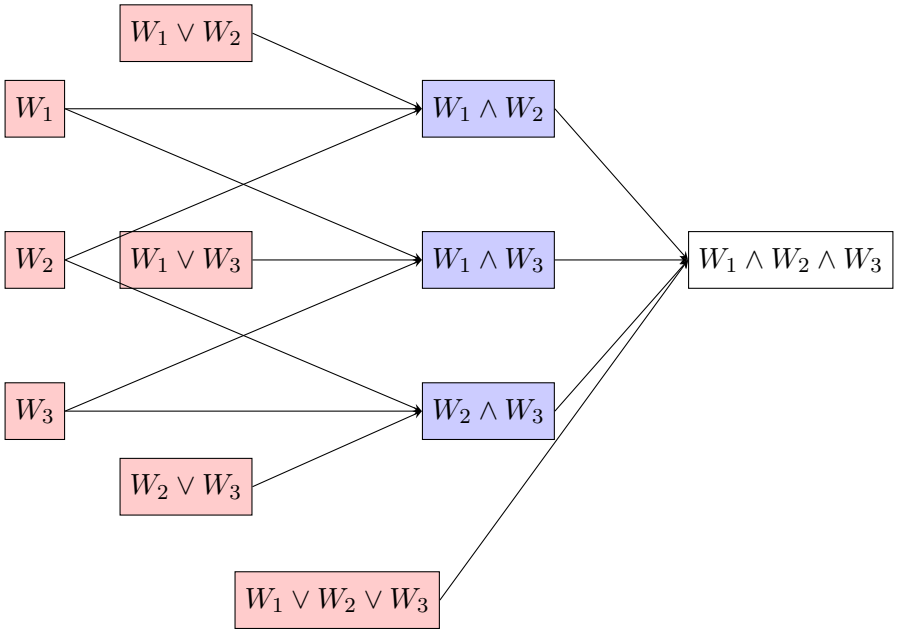


Figure 3: Influence functions of total and interaction explainabilities: zeroth-order, first-order, second-order interaction terms are highlighted in red, blue, and white, respectively. The dependence is derived from the inclusion-exclusion principle and the linearity of influence functions.

The difference $Y - Y(\boldsymbol{W}')$ can be decomposed into the sum of interaction effects with $\mathcal{S} \neq \emptyset$,

$$Y - Y(\boldsymbol{W}') = \sum_{\emptyset \neq \mathcal{S} \subseteq [K]} I_{\mathcal{S},\boldsymbol{W}'}(\boldsymbol{W}). \tag{16}$$

Note that $I_{\mathcal{S},\boldsymbol{W}'}(\boldsymbol{W})$, $I_{\mathcal{S}',\boldsymbol{W}'}(\boldsymbol{W})$ are correlated, therefore $\mathsf{Var}(Y - Y(\boldsymbol{W}'))$ typically does not equal the sum of the variance of $I_{\mathcal{S},\boldsymbol{W}'}(\boldsymbol{W})$. However, we show for independent treatments, $\mathsf{Var}(Y - Y(\boldsymbol{W}'))$ can be represented as a linear combination of the variance of $I_{\mathcal{S},\boldsymbol{W}'}(\boldsymbol{W})$.

Proposition 8 (Anchored variance decomposition) *Under Assumption 1,*

$$\xi(\vee_{k\in\mathcal{S}} W_k) = \sum_{\emptyset \neq \mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}'|-1}\, \xi(\wedge_{k\in\mathcal{S}'} W_k), \tag{17}$$

$$\xi(\wedge_{k\in\mathcal{S}} W_k) = \sum_{\emptyset \neq \mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}'|-1}\, \xi(\vee_{k\in\mathcal{S}'} W_k). \tag{18}$$

Proposition 8 indicates the total explainability and the explainability of interactions admits a relationship resembling the inclusion-exclusion principle in its structural form. Based on it, the influence functions for the explainability of interactions among treatments can be derived from the inclusion-exclusion principle and the linearity property of influence functions (Figure 3).

A.2. Extension of NPSEMs

We next discuss three NPSEM models of increasing generality, with NPSEM-IE (2) as the starting point.

NPSEM-IE. NPSEM-IE is a commonly used framework for causal discovery and inference (Hoyer et al., 2008; Peters et al., 2014). Under NPSEM-IE, when the nonparametric model is correctly specified, the multivariate distribution can not only identify the underlying causal structure and effects, but also typically allows for higher statistical accuracy compared to models with independent but heteroskedastic errors (below). When correctly specified, it identifies causal structure and effects and can yield higher statistical efficiency than the extensions below.

NPSEM with additive noise. Consider the NPSEM model where the noise remains additive, but its distribution may depend on the treatment level , i.e., heteroskedastic errors. In this setting, the counterfactual model satisfies the co-monotone property (Gao and Zhao, 2025) under which the joint distribution of all potential outcomes is identifiable. In principle, one could estimate this joint distribution, construct an oracle from it, and sample from this oracle to evaluate the Causal ANOVA quantities. However, learning this joint distribution is statistically challenging, and the inferential properties of such a procedure remain largely unexplored.

General NPSEM. Finally, when the noise is not necessarily additive, the co-monotonicity structure may be violated, and the Causal ANOVA quantities become only partially identifiable. Characterizing and conducting inference for the resulting partial identification set is important but beyond the scope of this paper.

Given the above discussion on extensions beyond NPSEM-IE, we next describe a practical way to assess whether NPSEM-IE or the additive-error NPSEM is more appropriate. Specifically, one can estimate the residuals for each possible and examine whether the residual distribution varies across treatment levels . If the residual distributions are approximately invariant across , this provides empirical support for the NPSEM-IE assumption; otherwise, the additive-error NPSEM framework may be more suitable.

A.3. Covariates adjustment

If there exists a set of covariates $\boldsymbol{X}$ such that the treatments are unconfounded given $\boldsymbol{X}$, that is $Y(\cdot)_i \perp\!\!\!\perp \boldsymbol{W}_i \mid \boldsymbol{X}_i$, the above analysis can be extended by conditioning on $\boldsymbol{X}$.

Assumption 3 (Unconfounded treatments) *The treatments are independent of the potential outcomes conditional on the covariates,*

$$Y_i(\cdot) \perp\!\!\!\perp \boldsymbol{W}_i \mid \boldsymbol{X}_i.$$

Assumption 4 (Conditionally independent treatments) *Treatments W_k, $k \in [K]$ are mutually independent conditional on the covariates.*

Definition 5 (Total explainability with covariates) *Let $\boldsymbol{W}'$ be an independent copy of $\boldsymbol{W}$ conditional on $\boldsymbol{X}$. For $\mathcal{S} \subseteq [K]$,*

$$\xi(\vee_{k\in\mathcal{S}} W_k) := \frac{\mathbb{E}\left[\mathsf{Var}\left(Y(\boldsymbol{W}) - Y(\boldsymbol{W}'_{\mathcal{S}}, \boldsymbol{W}_{-\mathcal{S}}) \mid \boldsymbol{X}\right)\right]}{2\mathsf{Var}\left(Y(\boldsymbol{W})\right)}. \tag{19}$$

Lemma 1 (Law of total conditional variance) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $Y \in \mathcal{L}_2(\mathbb{P})$, and let X and W be random elements. Define $\mathcal{G} := \sigma(X)$ and $\mathcal{H} := \mathcal{G} \vee \sigma(W)$. Then, $\mathbb{P}$ a.s.,*

$$\mathsf{Var}(Y \mid \mathcal{G}) = \mathbb{E}[\,\mathsf{Var}(Y \mid \mathcal{H}) \,|\, \mathcal{G}] \;+\; \mathsf{Var}(\mathbb{E}[Y \mid \mathcal{H}] \,|\, \mathcal{G})\,.$$

Equivalently,

$$\mathsf{Var}(Y \mid X) = \mathbb{E}[\,\mathsf{Var}(Y \mid X, W) \,|\, X] \;+\; \mathsf{Var}(\mathbb{E}[Y \mid X, W] \,|\, X) \quad \textit{a.s.}$$

The proof comes from the law of total variance.

Lemma 2 *Under Assumption 3 and Assumption 4, the definition (19) admits an equivalent form*

$$\xi(\vee_{k\in\mathcal{S}} W_k) = \frac{\mathbb{E}[\mathbb{E}^2\left[Y \mid \boldsymbol{W}, \boldsymbol{X}\right]]}{\mathsf{Var}\left(Y\right)} - \frac{\mathbb{E}\left[\mathbb{E}^2[Y \mid \boldsymbol{W}_{-\mathcal{S}}, \boldsymbol{X}]\right]}{\mathsf{Var}\left(Y\right)}. \tag{20}$$

The proof is similar to that of Proposition 1.

Definition 6 (Interaction explainability with covariates) *Let $\boldsymbol{W}'$ be an independent copy of $\boldsymbol{W}$ conditional on $\boldsymbol{X}$. For $\mathcal{S} \subseteq [K]$,*

$$\xi(\wedge_{k\in\mathcal{S}} W_k) = \frac{\mathbb{E}\left[\mathsf{Var}\left(I_{\mathcal{S},\boldsymbol{W}'}(\boldsymbol{W}) \mid \boldsymbol{X}\right)\right]}{2^{|\mathcal{S}|}\mathsf{Var}\left(Y(\boldsymbol{W})\right)}. \tag{21}$$

A.4. Extensions of randomization tests

When the assignment mechanism admits additional constraints, we can use a conditional randomization/permutation test that respects the design. We list a few examples.

1. Under blocking, permutations shall be carried out within each block.
2. Under fixed treatment margins, treatment assignments shall be permuted while keeping the margins fixed.
3. Under repeated measurements, treatment assignments shall be permuted within the measurements of each individual.

The key requirement for a conditional randomization/permutation test to be valid is that the treatment assignment W and the resampled W^* are exchangeable under the conditioning implied by the design.

Appendix B. Additional algorithms

Appendix C. Additional lemmas and corollaries

Lemma 3 (*Theorem 5.6.* in Tsiatis (2006)) *If no restrictions are put on the conditional density $p_{W|Z}(w|z)$, where the marginal density of Z is assumed to be from the semi-parametric model $p_Z(z, \beta, \eta)$, then the orthogonal complement of the tangent space $\mathscr{T}^{WZ}$ for the semi-parametric model for the joint distribution of (W, Z) (i.e., $\mathscr{T}^{WZ^\perp}$) is equal to the orthogonal complement of the tangent space $\mathscr{T}^Z$ for the semi-parametric model of the marginal distribution for Z alone (i.e., $\mathscr{T}^{Z^\perp}$).*

4. There might exist a degenerate null problem for zero interaction explainability, see simulation results in Section 5.

Algorithm 1: Cross-fitting of one-step correction estimator

Input: Data $\{\boldsymbol{W}_i, Y_i\}_{i=1}^n$; number of folds $L \geq 2$; an influence function φ; nuisance learners for $\mathbb{P}$, $\mu(\boldsymbol{w}) = \mathbb{E}[Y \mid \boldsymbol{W} = \boldsymbol{w}]$, and $\nu(\boldsymbol{w}) = \mathbb{E}[Y^2 \mid \boldsymbol{W} = \boldsymbol{w}]$.

Partition: Partition the sample into L folds $\{\mathcal{I}_\ell\}_{\ell=1}^L$ with (approximately) equal sizes $n_\ell = |\mathcal{I}_\ell|$.

for $\ell = 1, \ldots, L$ **do**

Nuisance estimation: On $\mathcal{I}_{-\ell}$, obtain nuisance estimators $\widehat{\mu}_{-\ell}(\boldsymbol{w})$ for $\mu(\boldsymbol{w})$, $\widehat{\nu}_{-\ell}(\boldsymbol{w})$ for $\nu(\boldsymbol{w})$, $\widehat{\mathbb{P}}_{W_k,-\ell}$ for $\mathbb{P}_{W_k}$, $k = 1, \ldots, K$, and compute

$$\widehat{\mathbb{P}}_{\boldsymbol{W},-\ell} := \prod_{k=1}^{K} \widehat{\mathbb{P}}_{W_k,-\ell}.$$

One-step correction: On the held-out fold $\mathcal{I}_\ell$, compute

$$\widehat{\xi}_\ell := \frac{1}{n_\ell} \sum_{i \in \mathcal{I}_\ell} \widehat{\xi}_i, \qquad \widehat{\xi}_i := \xi\big(\widehat{\mathbb{P}}_{-\ell}\big) + \varphi\big(Y_i, \boldsymbol{W}_i; \widehat{\mathbb{P}}_{-\ell}\big).$$

The details of $\xi(\widehat{\mathbb{P}}_{-\ell})$ and $\varphi(Y_i, \boldsymbol{W}_i; \widehat{\mathbb{P}}_{-\ell})$ are given in Table 4.

end

Aggregation: Average the estimates over folds:

$$\widehat{\xi} := \frac{1}{n} \sum_{\ell=1}^{L} n_\ell \widehat{\xi}_\ell.$$

Compute the variance estimator

$$\widehat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{\xi}_i - \widehat{\xi}\big)^2.$$

Output: Estimator $\widehat{\xi}$ and its estimated standard error $\widehat{\sigma}/\sqrt{n}$.

Algorithm 2: Randomization test for degenerate null

Input: Data $\{\boldsymbol{W}_i, Y_i\}_{i=1}^n$; test statistic $T(\boldsymbol{W}, Y)$ (e.g. an arbitrary estimator of $\xi(\vee_{k \in \mathcal{S}} W_k)$); number of permutations B.

for $b = 1, \ldots, B$ **do**

Generate $\boldsymbol{W}^{(b)} := (\boldsymbol{W}_{\mathcal{S},i}^{(b)}, \boldsymbol{W}_{-\mathcal{S},i})$ by permuting $\boldsymbol{W}_{\mathcal{S},i}$ while keeping $\boldsymbol{W}_{-\mathcal{S},i}$ fixed

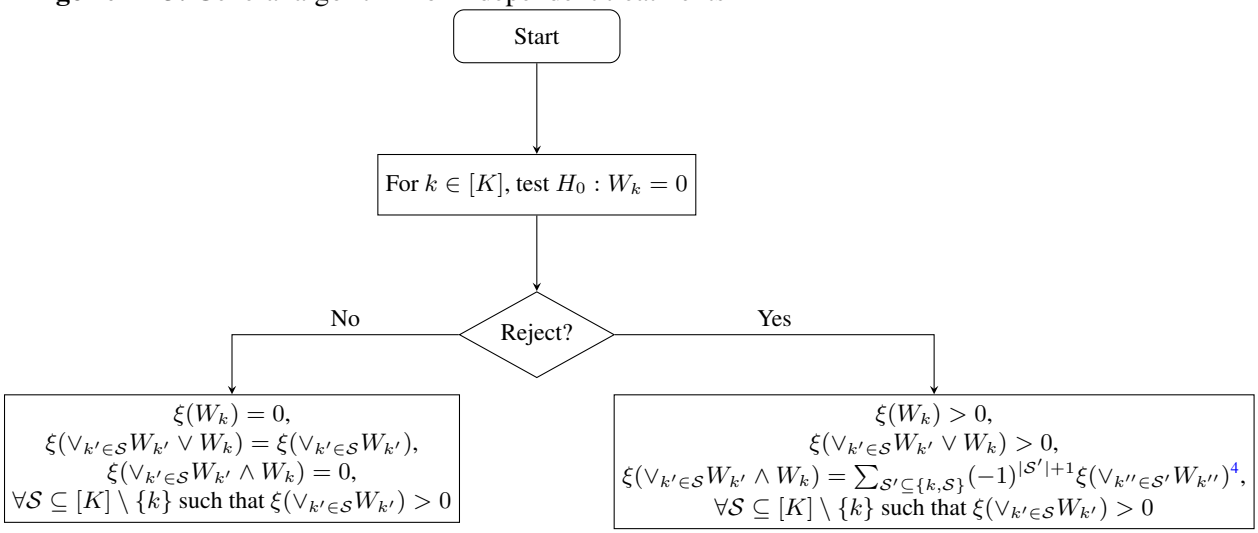
Impute $Y_i(\boldsymbol{W}^{(b)}) = Y_i$ and compute $T(\boldsymbol{W}^{(b)}, Y)$

end

Compute $P := \dfrac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbb{1}\Big\{T(\boldsymbol{W}^{(b)}, Y) \geq T(\boldsymbol{W}, Y)\Big\}\right]$

Output: p-value P

Algorithm 3: General algorithm for independent treatments



Lemma 4 *Fix $\mathcal{S} \subseteq [K]$ such that $0 < |-\mathcal{S}| < K$. Consider the semiparametric model $\mathcal{P}_{blk}$ for $(Y, \boldsymbol{W}) = (Y, \boldsymbol{W}_{-\mathcal{S}}, \boldsymbol{W}_{\mathcal{S}})$ defined by*

$$p(y, \boldsymbol{w}_{-\mathcal{S}}, \boldsymbol{w}_{\mathcal{S}}) = p_{Y|\boldsymbol{W}} \cdot p_{\boldsymbol{W}_{-\mathcal{S}}} \cdot p_{\boldsymbol{W}_{\mathcal{S}}},$$

where we put no restrictions on $p_{Y|\boldsymbol{W}}$, $p_{\boldsymbol{W}_{-\mathcal{S}}}$, or $p_{\boldsymbol{W}_{\mathcal{S}}}$ beyond domination and square integrability of scores.

Let $\dot{\mathcal{P}}_{\text{blk}}$ be the tangent space of $\mathcal{P}_{blk}$ at $\mathbb{P}$. Let $\mathcal{P}_{np}(Y, \boldsymbol{W}_{-\mathcal{S}})$ be the unrestricted model for the marginal law of $(Y, \boldsymbol{W}_{-\mathcal{S}})$, with tangent space $\dot{\mathcal{P}}_{Y,\boldsymbol{W}_{-\mathcal{S}}} = L_2^0(\mathbb{P}_{Y,\boldsymbol{W}_{-\mathcal{S}}})$.

Then

$$\dot{\mathcal{P}}^{\perp}_{Y,\boldsymbol{W}_{-\mathcal{S}}} = \{0\}, \text{ but } \dot{\mathcal{P}}^{\perp}_{\text{blk}} \neq \{0\}.$$

Proof Since $\mathcal{P}_{\text{np}}(Y, \boldsymbol{W}_{-\mathcal{S}})$ is unrestricted, its tangent space is $L_2^0(\mathbb{P}_{Y,\boldsymbol{W}_{-\mathcal{S}}})$ and hence $\dot{\mathcal{P}}^{\perp}_{Y,\boldsymbol{W}_{-\mathcal{S}}} = \{0\}$.

Under $\mathcal{P}_{\text{blk}}$, the score contributions from $p_{Y|\boldsymbol{W}}$, $p_{\boldsymbol{W}_{-\mathcal{S}}}$, and $p_{\boldsymbol{W}_{\mathcal{S}}}$ vary independently, so the tangent space decomposes orthogonally as

$$\dot{\mathcal{P}}_{\text{blk}} = \dot{\mathcal{P}}_{Y|\boldsymbol{W}} \ \oplus^{\perp} \ \dot{\mathcal{P}}_{\boldsymbol{W}_{-\mathcal{S}}} \ \oplus^{\perp} \ \dot{\mathcal{P}}_{\boldsymbol{W}_{\mathcal{S}}},$$

where

$$\dot{\mathcal{P}}_{Y|\boldsymbol{W}} = \{s(Y, \boldsymbol{W}) : \mathbb{E}[s \mid \boldsymbol{W}] = 0\}, \quad \dot{\mathcal{P}}_{\boldsymbol{W}_{-\mathcal{S}}} = \{s(\boldsymbol{W}_{-\mathcal{S}}) : \mathbb{E}[s] = 0\}, \quad \dot{\mathcal{P}}_{\boldsymbol{W}_{\mathcal{S}}} = \{s(\boldsymbol{W}_{\mathcal{S}}) : \mathbb{E}[s] = 0\}.$$

Hence $h(Y, \boldsymbol{W}) \in \dot{\mathcal{P}}^{\perp}_{\text{blk}} = \dot{\mathcal{P}}^{\perp}_{Y|\boldsymbol{W}} \ \cap \ \dot{\mathcal{P}}^{\perp}_{\boldsymbol{W}_{-\mathcal{S}}} \ \cap \ \dot{\mathcal{P}}^{\perp}_{\boldsymbol{W}_{\mathcal{S}}}$ iff it is orthogonal to each component. First, any $h(\boldsymbol{W})$ is orthogonal to $\dot{\mathcal{P}}_{Y|\boldsymbol{W}}$ because for $s \in \dot{\mathcal{P}}_{Y|\boldsymbol{W}}$,

$$\mathbb{E}[h(\boldsymbol{W})s(Y, \boldsymbol{W})] = \mathbb{E}\{h(\boldsymbol{W})\mathbb{E}[s(Y, \boldsymbol{W}) \mid \boldsymbol{W}]\} = 0.$$

So it suffices to find a nonzero $h(\boldsymbol{W})$ orthogonal to both $\dot{\mathcal{P}}_{\boldsymbol{W}_{-\mathcal{S}}}$ and $\dot{\mathcal{P}}_{\boldsymbol{W}_{\mathcal{S}}}$.

Pick square-integrable functions $u(\boldsymbol{W}_{\mathcal{S}})$ and $v(\boldsymbol{W}_{-\mathcal{S}})$ such that $\mathbb{E}[u(\boldsymbol{W}_{\mathcal{S}})] = \mathbb{E}[v(\boldsymbol{W}_{-\mathcal{S}})] = 0$ and neither is a.s. zero. Define

$$h(\boldsymbol{W}) := u(\boldsymbol{W}_{\mathcal{S}})\, v(\boldsymbol{W}_{-\mathcal{S}}).$$

Then $h(\boldsymbol{W}) \not\equiv 0$. Moreover, for any mean-zero $s(\boldsymbol{W}_{\mathcal{S}})$, using block independence and $\mathbb{E}[v] = 0$,

$$\mathbb{E}[h(\boldsymbol{W})\, s(\boldsymbol{W}_{\mathcal{S}})] = \mathbb{E}[v(\boldsymbol{W}_{-\mathcal{S}})]\ \mathbb{E}[u(\boldsymbol{W}_{\mathcal{S}})s(\boldsymbol{W}_{\mathcal{S}})] = 0,$$

and similarly for any mean-zero $s(\boldsymbol{W}_{-\mathcal{S}})$,

$$\mathbb{E}[h(\boldsymbol{W})\, s(\boldsymbol{W}_{-\mathcal{S}})] = \mathbb{E}[u(\boldsymbol{W}_{\mathcal{S}})]\ \mathbb{E}[v(\boldsymbol{W}_{-\mathcal{S}})s(\boldsymbol{W}_{-\mathcal{S}})] = 0.$$

Thus $h \in \dot{\mathcal{P}}_{\text{blk}}^{\perp}$ and the orthocomplement is nontrivial. ■

Lemma 5 *Assume $K \geq 2$ and the full-independence model $\mathcal{P}_{ind}$ for $(Y, \boldsymbol{W})$:*

$$p(y, \boldsymbol{w}) = p_{Y|\boldsymbol{W}} \prod_{k=1}^{K} p_{w_k},$$

with no restrictions on $p_{Y|\boldsymbol{W}}$ and each p_{w_k} beyond domination and square-integrable scores. Let $\dot{\mathcal{P}}_{ind}$ be its tangent space at $\mathbb{P}$. Let $\mathcal{P}_{np}(Y)$ be the unrestricted model for the marginal law of Y alone, with tangent space $\dot{\mathcal{P}}_Y = L_2^0(\mathbb{P}_Y)$.

Then

$$\dot{\mathcal{P}}_Y^{\perp} = \{0\},\ \textit{but}\ \dot{\mathcal{P}}_{ind}^{\perp} \neq \{0\}.$$

Proof Since $\mathcal{P}_{\text{np}}(Y)$ is unrestricted, $\dot{\mathcal{P}}_Y = L_2^0(\mathbb{P}_Y)$ and thus $\dot{\mathcal{P}}_Y^{\perp} = \{0\}$.

Under $\mathcal{P}_{\text{ind}}$, the tangent space decomposes orthogonally as

$$\dot{\mathcal{P}}_{\text{ind}} = \bigoplus_{k\in[K]}^{\perp} \dot{\mathcal{P}}_{W_k} \oplus^{\perp} \dot{\mathcal{P}}_{Y|\boldsymbol{W}},$$

where $\dot{\mathcal{P}}_{Y|\boldsymbol{W}} = \{s(Y, \boldsymbol{W}) : \mathbb{E}[s \mid \boldsymbol{W}] = 0\}$ and $\dot{\mathcal{P}}_{W_k} = \{s(W_k) : \mathbb{E}[s] = 0\}$.

Pick distinct indices $i \neq j$ and choose square-integrable $u(W_i), v(W_j)$ with $\mathbb{E}[u(W_i)] = \mathbb{E}[v(W_j)] = 0$ and not a.s. zero. Let

$$h(\boldsymbol{W}) := u(W_i)v(W_j).$$

Then $h(\boldsymbol{W}) \not\equiv 0$. As in Lemma 4, $h(\boldsymbol{W}) \perp \dot{\mathcal{P}}_{Y|\boldsymbol{W}}$. For any $s(W_k) \in \dot{\mathcal{P}}_{W_k}$, mutual independence and centering imply

$$\mathbb{E}[h(\boldsymbol{W})s(W_i)] = \mathbb{E}[v(W_j)]\,\mathbb{E}[u(W_i)s(W_i)] = 0,\ \mathbb{E}[h(\boldsymbol{W})s(W_j)] = \mathbb{E}[u(W_i)]\,\mathbb{E}[v(W_j)s(W_j)] = 0,$$

and for $k \notin \{i, j\}$,

$$\mathbb{E}[h(\boldsymbol{W})s(W_k)] = \mathbb{E}[h(\boldsymbol{W})]\,\mathbb{E}[s(W_k)] = 0.$$

Thus $h \in \dot{\mathcal{P}}_{\text{ind}}^{\perp} = \bigcap_{k\in[K]} \dot{\mathcal{P}}_{W_k}^{\perp} \cap \dot{\mathcal{P}}_{Y|\boldsymbol{W}}^{\perp}$, proving the orthocomplement is nontrivial. ■

Corollary 3 *If $\mathbb{P}_{\boldsymbol{W}}$ is known, the efficient influence function takes the form*

$$\frac{2Y\mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}] - 2\mathbb{E}[Y \mid \boldsymbol{W}_{-\mathcal{S}}]^2}{\mathsf{Var}\,(Y)} - \frac{\left\{ (Y - \mathbb{E}\,(Y))^2 - \mathsf{Var}\,(Y) \right\} \cdot \mathbb{E}\left[\mathbb{E}\left[Y \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2\right]}{\mathsf{Var}\,(Y)^2}.$$

The proof is simple as we only need to remove the irrelevant parts from the original EIF according to the new tangent space.

Lemma 6 (*Lemma 2.* in Kennedy et al. (2020)) *Let $\widehat{f}(\mathbf{o})$ be a function estimated from a sample $\mathbf{O}^N = (\mathbf{O}_{n+1}, \ldots, \mathbf{O}_N)$, and let $\mathbb{P}_n$ denote the empirical measure over $(\mathbf{O}_1, \ldots, \mathbf{O}_n)$, which is independent of $\mathbf{O}^N$. Then*

$$(\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - f\|}{\sqrt{n}}\right).$$

Lemma 7 *Let $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ be i.i.d. from $\mathbb{P}$. Fix a fold $l \subset \{1, \ldots, n\}$ with $m := |l|$ and let $n_{-l} := n - m$. Define the empirical measure on the training sample $-l$ by*

$$\widehat{\mathbb{P}}_{-l} f := \frac{1}{n_{-l}} \sum_{i \notin l} f(\boldsymbol{W}_i).$$

Let $h_{-l} : \mathcal{W} \to \mathbb{R}$ be $\sigma(\boldsymbol{W}_i : i \notin l)$-measurable. Assume there exists a nonrandom function class $\mathcal{H}$ such that $h_{-l} \in \mathcal{H}$ a.s. for all l. Let $\mathcal{F} := \{h^2 : h \in \mathcal{H}\}$. If $\mathcal{F}$ is $\mathbb{P}$–Donsker, then

$$\left| \|h_{-l}\|^2_{L_2(\widehat{\mathbb{P}}_{-l})} - \|h_{-l}\|^2_{L_2(\mathbb{P})} \right| = \left| (\widehat{\mathbb{P}}_{-l} - \mathbb{P})\, h^2_{-l} \right| = O_{\mathbb{P}}(n_{-l}^{-1/2}).$$

In particular, under fixed K-fold cross-fitting, $n_{-l} \asymp n$, hence the difference is $o_{\mathbb{P}}(1/\sqrt{n})$.

Proof Note that $\|h_{-l}\|^2_{L_2(\widehat{\mathbb{P}}_{-l})} = \widehat{\mathbb{P}}_{-l}[h^2_{-l}]$ and $\|h_{-l}\|^2_{L_2(\mathbb{P})} = \mathbb{P}[h^2_{-l}]$, so the left-hand side equals $|(\widehat{\mathbb{P}}_{-l} - \mathbb{P})h^2_{-l}|$.

Because $h_{-l} \in \mathcal{H}$ a.s., we have $h^2_{-l} \in \mathcal{F}$ a.s., hence the domination

$$\left|(\widehat{\mathbb{P}}_{-l} - \mathbb{P})h^2_{-l}\right| \le \sup_{f \in \mathcal{F}} \left|(\widehat{\mathbb{P}}_{-l} - \mathbb{P})f\right|.$$

If $\mathcal{F}$ is $\mathbb{P}$–Donsker, then the empirical process $\mathbb{G}_{-l} := \sqrt{n_{-l}}(\widehat{\mathbb{P}}_{-l} - \mathbb{P})$ is asymptotically tight in $\ell^\infty(\mathcal{F})$, which implies

$$\sup_{f \in \mathcal{F}} |\mathbb{G}_{-l} f| = O_{\mathbb{P}}(1).$$

Therefore,

$$\sqrt{n_{-l}}\left|(\widehat{\mathbb{P}}_{-l} - \mathbb{P})h^2_{-l}\right| \le \sup_{f \in \mathcal{F}} |\mathbb{G}_{-l} f| = O_{\mathbb{P}}(1),$$

i.e. $|(\widehat{\mathbb{P}}_{-l} - \mathbb{P})h^2_{-l}| = O_{\mathbb{P}}(n_{-l}^{-1/2})$. ■

Appendix D. Omitted proofs

D.1. Proof of Proposition 1

Proof

$$
\begin{aligned}
\xi(\vee_{k\in\mathcal{S}}W_k) :=& \frac{\mathsf{Var}\left(Y - Y(\boldsymbol{W}'_{\mathcal{S}}, \boldsymbol{W}_{-\mathcal{S}})\right)}{2\mathsf{Var}\left(Y\right)} \\
=& \frac{\mathbb{E}\left[\mathsf{Var}\left(Y - Y(\boldsymbol{W}'_{\mathcal{S}}, \boldsymbol{W}_{-\mathcal{S}}) \mid Y(\cdot), \boldsymbol{W}_{-\mathcal{S}}\right)\right]}{2\mathsf{Var}\left(Y\right)} \\
&+ \frac{\mathsf{Var}\left[\mathbb{E}\left(Y - Y(\boldsymbol{W}'_{\mathcal{S}}, \boldsymbol{W}_{-\mathcal{S}}) \mid Y(\cdot), \boldsymbol{W}_{-\mathcal{S}}\right)\right]}{2\mathsf{Var}\left(Y\right)} \text{ (Equation (1))} \\
=& \frac{\mathbb{E}\left[\mathsf{Var}\left(Y \mid Y(\cdot), \boldsymbol{W}_{-\mathcal{S}}\right)\right] + \mathbb{E}\left[\mathsf{Var}\left(Y(\boldsymbol{W}'_{\mathcal{S}}, \boldsymbol{W}_{-\mathcal{S}}) \mid Y(\cdot), \boldsymbol{W}_{-\mathcal{S}}\right)\right]}{2\mathsf{Var}\left(Y\right)} \text{ (Assumption 1)} \\
=& \frac{\mathbb{E}\left[\mathsf{Var}\left(Y \mid Y(\cdot), \boldsymbol{W}_{-\mathcal{S}}\right)\right]}{\mathsf{Var}\left(Y\right)} \text{ (independent copy)} \\
=& \frac{\mathbb{E}[\mathsf{Var}(Y \mid \boldsymbol{W}_{-\mathcal{S}}, E_Y)]}{\mathsf{Var}\left(Y\right)} \quad \text{(Definition 1)} \\
=& \frac{\mathbb{E}[\mathsf{Var}(\mathbb{E}[Y \mid \boldsymbol{W}] + E_Y \mid \boldsymbol{W}_{-\mathcal{S}}, E_Y)]}{\mathsf{Var}\left(Y\right)} \quad \text{(Assumption 2)} \\
=& \frac{\mathbb{E}\left[\mathsf{Var}(\mathbb{E}[Y \mid \boldsymbol{W}] \mid \boldsymbol{W}_{-\mathcal{S}})\right]}{\mathsf{Var}\left(Y\right)} \quad (\boldsymbol{W} \perp\!\!\!\perp E_Y) \\
=& \frac{\mathbb{E}[\mathbb{E}^2\left[Y \mid \boldsymbol{W}\right]]}{\mathsf{Var}\left(Y\right)} - \frac{\mathbb{E}\left[\mathbb{E}^2[Y \mid \boldsymbol{W}_{-\mathcal{S}}]\right]}{\mathsf{Var}\left(Y\right)}. \quad \text{(Definition of variance)}
\end{aligned}
$$

■

D.2. Proof of Proposition 2

Proof

We can rewrite the parameter of interest as $\Psi := \psi \circ T$, where $\Psi : \mathcal{P} \to \mathbb{R}$, $\psi : \mathbb{D} \to \mathbb{R}$ and $T : \mathcal{P} \to \mathbb{D}$. Let $T_1 := \mathbb{E}\Big[\mathbb{E}\{Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}\}^2\Big]$, $T_2 := \mathsf{Var}(Y)$, and define $T := (T_1, T_2)$, $\psi(x, y) := x/y$. Assume that $T_2 > 0$, so that $T \in \mathbb{D} := \{(x, y) \in \mathbb{R}^2 : y > 0\}$. Since ψ is $\mathbb{C}^1$ on the open set $\mathbb{D}$, it is Fréchet differentiable there, and hence Hadamard differentiable. By *Lemma 1.* in Williamson et al. (2021), T is pathwise differentiable, i.e. Hadamard differentiable along the tangent space $\dot{\mathcal{P}}$ (van der Vaart, 1991; Bickel et al., 1993). Since Hadamard differentiability allows chain rule (Averbukh and Smolyanov, 1967), $\Psi = \psi \circ T$ is Hadamard differentiable along $\dot{\mathcal{P}}$, i.e. pathwise differentiable, with derivative

$$
\dot{\Psi}(S) = \dot{\psi}_T \circ \dot{T} = \frac{\dot{T}_1(S)}{T_2} - \frac{T_1}{T_2^2}\dot{T}_2(S) \in \mathbb{R}, \quad S \in \dot{\mathcal{P}}.
$$

Then by Riesz-Fréchet representation theorem, there exists a unique $\varphi_{\text{EIF}} \in \dot{\mathcal{P}}$ such that

$$
\dot{\Psi}(S) = \langle \varphi_{\text{EIF}}, S\rangle_{L_2} = \langle \frac{\varphi_{T_1,\text{EIF}}}{T_2} - \frac{T_1}{T_2^2}\varphi_{T_2,\text{EIF}}, S\rangle_{L_2}.
$$

If we ignore the independence among treatments, then the model we consider is a non-parametric model $\mathcal{P}_{\text{np}}$ with tangent space $\dot{\mathcal{P}}_{\text{np}} = L_2^0(\mathbb{P})$. Under the assumption of discrete treatments, for $-\mathcal{S} \subseteq [K]$, we use the operator $\mathbb{IF}$ for pathwise differentiable functionals (Kennedy, 2024) to get the non-parametric influence function of the numerator:

$$\begin{aligned}
\varphi_{T_1,\text{IF}} &= \mathbb{IF}\Big\{\mathbb{E}\left[\mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2\right]\Big\} \\
&= \mathbb{IF}\Big\{\int \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}} = \boldsymbol{w}_{-\mathcal{S}}\right]^2 p_{\boldsymbol{w}_{-\mathcal{S}}} \lambda(d\boldsymbol{w}_{-\mathcal{S}})\Big\} \\
&= \int \mathbb{IF}\{\mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}} = \boldsymbol{w}_{-\mathcal{S}}\right]^2\} p_{\boldsymbol{w}_{-\mathcal{S}}} \lambda(d\boldsymbol{w}_{-\mathcal{S}}) \\
&\quad + \int \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}} = \boldsymbol{w}_{-\mathcal{S}}\right]^2 \mathbb{IF}(p_{\boldsymbol{w}_{-\mathcal{S}}}) \lambda(d\boldsymbol{w}_{-\mathcal{S}}) \\
&= \int 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}} = \boldsymbol{w}_{-\mathcal{S}}] \frac{\mathbb{1}(\boldsymbol{W}_{-\mathcal{S}} = \boldsymbol{w}_{-\mathcal{S}})}{p_{\boldsymbol{w}_{-\mathcal{S}}}} \left(Y(\boldsymbol{W}) - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}} = \boldsymbol{w}_{-\mathcal{S}}]\right) p_{\boldsymbol{w}_{-\mathcal{S}}} \lambda(d\boldsymbol{w}_{-\mathcal{S}}) \\
&\quad + \int \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}} = \boldsymbol{w}_{-\mathcal{S}}]^2 \left(\mathbb{1}(\boldsymbol{W}_{-\mathcal{S}} = \boldsymbol{w}_{-\mathcal{S}}) - p_{\boldsymbol{w}_{-\mathcal{S}}}\right) \lambda(d\boldsymbol{w}_{-\mathcal{S}}) \\
&= 2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 - \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2\right].
\end{aligned}$$

For the denominator, we have its non-parametric influence function as

$$\varphi_{T_2,\text{IF}} = \mathbb{IF}\Big\{\mathsf{Var}\left(Y(\boldsymbol{W})\right)\Big\} = \left(Y(\boldsymbol{W}) - \mathbb{E}\left(Y(\boldsymbol{W})\right)\right)^2 - \mathsf{Var}\left(Y(\boldsymbol{W})\right).$$

So the non-parametric influence function of the parameter of interest is

$$\begin{aligned}
\varphi_{\text{IF}} &= \frac{2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 - \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2\right]}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)} \\
&\quad - \frac{\left(Y(\boldsymbol{W}) - \mathbb{E}\left(Y(\boldsymbol{W})\right)\right)^2 \cdot \mathbb{E}\left[\mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2\right] - \mathsf{Var}\left(Y(\boldsymbol{W})\right) \cdot \mathbb{E}\left[\mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2\right]}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)^2} \\
&= \frac{2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)} \\
&\quad - \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2\right] \cdot \left\{\frac{Y(\boldsymbol{W}) - \mathbb{E}\left(Y(\boldsymbol{W})\right)}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)}\right\}^2.
\end{aligned}$$

Given $-\mathcal{S} \subsetneq [K]$ and fully independent treatments, the joint probability measure of observed data can be factorized as $\mathbb{P}_{Y,\boldsymbol{W}} = \mathbb{P}_{Y|\boldsymbol{W}} \cdot \prod_{j \in -\mathcal{S}} \mathbb{P}_{W_j} \cdot \prod_{i \in \mathcal{S}} \mathbb{P}_{W_i}$. Let $\mathcal{P}_{\text{ind}}$ be the set of all regular densities $p_{Y,\boldsymbol{W}}$ such that

$$p_{Y,\boldsymbol{W}} = p_{Y|\boldsymbol{W}} \cdot \prod_{j \in -\mathcal{S}} p_{W_j} \cdot \prod_{i \in \mathcal{S}} p_{W_i}.$$

The reason why we have $\{W_i : i \in \mathcal{S}\}$ here is that, even though they are not needed to define our parameter of interest, fully independent treatments implies that $p_{\boldsymbol{W}_{-\mathcal{S}}|\boldsymbol{W}_{\mathcal{S}}} = p_{\boldsymbol{W}_{-\mathcal{S}}}$, then by Lemma 4, the orthogonal complement of the tangent space $\dot{\mathcal{P}}_{Y,\boldsymbol{W}_{-\mathcal{S}},\boldsymbol{W}_{\mathcal{S}}}$ for the joint distribution of

$(Y, \boldsymbol{W}_{-\mathcal{S}}, \boldsymbol{W}_{\mathcal{S}})$, $\dot{\mathcal{P}}^{\perp}_{Y,\boldsymbol{W}_{-\mathcal{S}},\boldsymbol{W}_{\mathcal{S}}}$, is not the same as the orthogonal complement of the tangent space $\dot{\mathcal{P}}_{Y,\boldsymbol{W}_{-\mathcal{S}}}$ for the joint distribution of $(Y, \boldsymbol{W}_{-\mathcal{S}})$, $\dot{\mathcal{P}}^{\perp}_{Y,\boldsymbol{W}_{-\mathcal{S}}}$. So we can still gain efficiency if these auxiliary variables are included. Then for this semi-parametric model, we can project the non-parametric influence function of the numerator onto the corresponding tangent space,

$$\begin{aligned}
\dot{\mathcal{P}}_{\text{ind}} &= \bigoplus_{k\in[K]}^{\perp} \dot{\mathcal{P}}_{W_k} \oplus^{\perp} \dot{\mathcal{P}}_{Y|\boldsymbol{W}} \\
&= \bigoplus_{k\in[K]}^{\perp} \{S(W_k) : \mathbb{E}(S) = 0\} \oplus^{\perp} \{S(Y, \boldsymbol{W}) : \mathbb{E}(S \mid \boldsymbol{W}) = 0\},
\end{aligned}$$

to get the efficient influence function

$$\begin{aligned}
\varphi_{T_1,\text{EIF}} =& \sum_{k\in[K]} \prod \left(\mathbb{IF}\Big\{ \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \right] \Big\} \mid \dot{\mathcal{P}}_{W_k} \right) \\
&+ \prod \left(\mathbb{IF}\Big\{ \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \right] \Big\} \mid \dot{\mathcal{P}}_{Y|\boldsymbol{W}} \right) \\
=& \sum_{k\in[K]} \mathbb{E}\left(\mathbb{IF}\Big\{ \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \right] \Big\} \mid W_k \right) + \mathbb{IF}\Big\{ \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \right] \Big\} \\
&- \mathbb{E}\left(\mathbb{IF}\Big\{ \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \right] \Big\} \mid \boldsymbol{W} \right) \\
=& 2\left(Y(\boldsymbol{W}) - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\right) \cdot \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] \\
&+ \sum_{j\in-\mathcal{S}} \Big\{ \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \mid W_j \right] - \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \right] \Big\} \\
&+ 2\sum_{i\in\mathcal{S}} \Big\{ \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}] \cdot \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] \mid W_i \right] - \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \right] \Big\}.
\end{aligned}$$

For the denominator, we can also project its non-parametric influence function onto $\dot{\mathcal{P}}_{\text{ind}}$. Because constraining the marginal distribution of $\boldsymbol{W}$ to be factorizable (mutually independent) implicitly restricts the conditional density $p_{\boldsymbol{W}|Y}$ to only those specific forms that preserve this independence after integrating out Y, which violates the condition of Lemma 3. So Lemma 5 kicks in and we get

$$\begin{aligned}
\varphi_{T_2,\text{EIF}} =& \left(Y - \mathbb{E}\left[Y\right]\right)^2 - \mathsf{Var}\left(Y \mid \boldsymbol{W}\right) - \left(\mathbb{E}[Y \mid \boldsymbol{W}] - \mathbb{E}[Y]\right)^2 \\
&+ \sum_{k\in[K]} \Big\{ \mathbb{E}\left[(Y - \mathbb{E}[Y])^2 \mid W_k \right] - \mathsf{Var}\left(Y\right) \Big\}.
\end{aligned}$$

So the efficient influence function of the parameter of interest for $\mathcal{P}_{\text{ind}}$ is

$$\begin{aligned}\varphi_{\text{EIF}} = & \frac{2\left(Y(\boldsymbol{W}) - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\right) \cdot \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)} \\ & + \frac{2\sum_{i\in\mathcal{S}}\left\{\mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}] \cdot \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] \mid W_i\right] - \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2\right]\right\}}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)} \\ & + \frac{\sum_{j\in-\mathcal{S}}\left\{\mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \mid W_j\right] - \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2\right]\right\}}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)} \\ & - \frac{\left\{\left(Y - \mathbb{E}\left[Y\right]\right)^2 - \mathsf{Var}\left(Y \mid \boldsymbol{W}\right) - \left(\mathbb{E}[Y \mid \boldsymbol{W}] - \mathbb{E}[Y]\right)^2\right\} \cdot \mathbb{E}\left[\mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2\right]}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)^2} \\ & - \frac{\left\{\sum_{k\in[K]}\left\{\mathbb{E}\left[(Y - \mathbb{E}[Y])^2 \mid W_k\right] - \mathsf{Var}\left(Y\right)\right\}\right\} \cdot \mathbb{E}\left[\mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2\right]}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)^2}.\end{aligned}$$

■

D.3. Proof of Proposition 3

See *Theorem 1.* in Williamson et al. (2021).

D.4. Proof of Proposition 4

Proof Define the remainder term based on probability measures $\overline{\mathbb{P}}$ and $\mathbb{P}$ as $R(\overline{\mathbb{P}}, \mathbb{P}) = \psi_{\overline{\mathbb{P}}} - \psi + \mathbb{P}\{\varphi_{\overline{\mathbb{P}}}\}$, where ψ is any estimand. We begin with the numerator, denote it by Θ and its efficient influence function by θ. For $l = 1, \ldots, L$,

$$\begin{aligned}\widehat{\Theta}_l - \Theta &= \Theta_{\widehat{\mathbb{P}}_{-l}} + \mathbb{P}_n^l\{\theta_{\widehat{\mathbb{P}}_{-l}}\} - \Theta \\ &= (\mathbb{P}_n^l - \mathbb{P})\{\theta_{\widehat{\mathbb{P}}_{-l}}\} + R(\widehat{\mathbb{P}}_{-l}, \mathbb{P}) \\ &= (\mathbb{P}_n^l - \mathbb{P})\{\theta\} + (\mathbb{P}_n^l - \mathbb{P})\{\theta_{\widehat{\mathbb{P}}_{-l}} - \theta\} + R(\widehat{\mathbb{P}}_{-l}, \mathbb{P}).\end{aligned}$$

By the Central Limit Theorem, the first term, a de-meaned sample average of the true efficient influence function, will behave as a normal random variable with variance $\mathsf{Var}\{\theta\}/n$, up to error $o_{\mathbb{P}}(1/\sqrt{n})$, as long as $\mathsf{Var}\{\theta\} < \infty$. For the empirical process term, according to Lemma 6, cross-fitting enables us to conclude that,

$$(\mathbb{P}_n^l - \mathbb{P})\{\theta_{\widehat{\mathbb{P}}_{-l}} - \theta\} = O_{\mathbb{P}}\left(||\theta_{\widehat{\mathbb{P}}_{-l}} - \theta||/\sqrt{n}\right),$$

which means we only need to show that $\theta_{\widehat{\mathbb{P}}_{-l}}$ converge to θ in $L_2(\mathbb{P})$ norm to get the desired convergence rate. The exact form of the first half of $\theta_{\widehat{\mathbb{P}}_{-l}} - \theta$ is

$$
\begin{aligned}
(\theta_{\widehat{\mathbb{P}}_{-l}} - \theta)_{\text{first half}} = \; & 2Y(\boldsymbol{W})\widehat{\mu}_{-l} - 2\widehat{\mu}^2_{-l} + \sum_{k\in[K]} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-l} \mid W_k] - [K]\cdot\widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-l}] \\
& - \Big\{2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}] - 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]^2 + \sum_{k\in[K]} \mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]^2 \mid W_k\big] - [K]\cdot\mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]^2\big]\Big\} \\
= \; & \underbrace{2Y(\boldsymbol{W})\widehat{\mu}_{-l} - 2\widehat{\mu}^2_{-l} - 2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}] + 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]^2}_{:=a} \\
& + \underbrace{\sum_{k\in[K]} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-l} \mid W_k] - [K]\cdot\widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-l}] - \sum_{k\in[K]} \mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]^2 \mid W_k\big] + [K]\cdot\mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]^2\big]}_{:=b}.
\end{aligned}
$$

Note that, under C1. and C2., we have

$$
\begin{aligned}
& ||a|| = ||2Y(\boldsymbol{W})\widehat{\mu}_{-l} - 2\widehat{\mu}^2_{-l} - 2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}] + 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]^2|| \\
= \; & ||2Y(\boldsymbol{W})\cdot(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]) - 2\,(\widehat{\mu}_{-l} + \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}])\cdot(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}])\,|| \\
= \; & ||2\,(Y(\boldsymbol{W}) - \widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}])\cdot(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}])\,|| \\
\lesssim \; & ||\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]|| = o_{\mathbb{P}}(1)
\end{aligned}
$$

and

$$
\begin{aligned}
\|b\| &= \Big\| \sum_{k\in[K]} \widehat{\mathbb{E}}_{-l}\big[\widehat{\mu}^2_{-l} \mid W_k\big] - [K]\cdot\widehat{\mathbb{E}}_{-l}\big[\widehat{\mu}^2_{-l}\big] - \sum_{k\in[K]} \mathbb{E}\Big[\mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \mid W_k\Big] + [K]\cdot\mathbb{E}\Big[\mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2\Big]\Big\| \\
&\le \sum_{k\in[K]} \Big\|\widehat{\mathbb{E}}_{-l}\big[\widehat{\mu}^2_{-l} \mid W_k\big] - \mathbb{E}\Big[\mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \mid W_k\Big]\Big\| + [K]\cdot\Big|\widehat{\mathbb{E}}_{-l}\big[\widehat{\mu}^2_{-l}\big] - \mathbb{E}\Big[\mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2\Big]\Big|.
\end{aligned}
$$

For the first term, for each fixed $k \in [K]$, we have

$$
\begin{aligned}
& \widehat{\mathbb{E}}_{-l}\big[\widehat{\mu}^2_{-l} \mid W_k\big] - \mathbb{E}\Big[\mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \mid W_k\Big] \\
& \quad = \int \Big(\widehat{\mu}^2_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2\Big) \prod_{k'\neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} \;+\; \int \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \Big(\prod_{k'\neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} - \prod_{k'\neq k} d\mathbb{P}_{W_{k'}}\Big).
\end{aligned}
$$

Hence, by the triangle inequality and the fact that this quantity is W_k-measurable,

$$
\begin{aligned}
\Big\|\widehat{\mathbb{E}}_{-l}\big[\widehat{\mu}^2_{-l} \mid W_k\big] - \mathbb{E}\Big[\mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \mid W_k\Big]\Big\| &= \Big\|\widehat{\mathbb{E}}_{-l}\big[\widehat{\mu}^2_{-l} \mid W_k\big] - \mathbb{E}\Big[\mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \mid W_k\Big]\Big\|_{L_2(\mathbb{P}_{W_k})} \\
&\le \Big\| \int \Big(\widehat{\mu}^2_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2\Big) \prod_{k'\neq k} d\widehat{\mathbb{P}}_{W_{k'},-l}\Big\|_{L_2(\mathbb{P}_{W_k})} + \Big\| \int \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \Big(\prod_{k'\neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} - \prod_{k'\neq k} d\mathbb{P}_{W_{k'}}\Big)\Big\|_{L_2}
\end{aligned}
$$

Write $\widehat{\mu}_{-l}^2 - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 = \Big(\widehat{\mu}_{-l} + \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]\Big)\Big(\widehat{\mu}_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]\Big)$. For the left-hand side term, under C1., by Minkowski's integral inequality,

$$
\begin{aligned}
&\bigg\| \int \Big(\widehat{\mu}_{-l}^2 - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2\Big) \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} \bigg\|_{L_2(\mathbb{P}_{W_k})} \\
&\quad \leq \int \Big\| \Big(\widehat{\mu}_{-l} + \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]\Big)\Big(\widehat{\mu}_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]\Big) \Big\|_{L_2(\mathbb{P}_{W_k})} \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} \\
&\quad \leq 2C \int \Big\| \widehat{\mu}_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big] \Big\|_{L_2(\mathbb{P}_{W_k})} \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} \\
&\quad \leq 2C \bigg(\int \Big\| \widehat{\mu}_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big] \Big\|_{L_2(\mathbb{P}_{W_k})}^2 \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} \bigg)^{1/2} \\
&\quad = 2C \bigg(\int \Big(\widehat{\mu}_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]\Big)^2 d\mathbb{P}_{W_k} \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} \bigg)^{1/2}.
\end{aligned}
$$

Using $0 \leq (\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}])^2 \leq (2C)^2$ and the telescoping identity for products, we get

$$
\begin{aligned}
&\bigg| \int \Big(\widehat{\mu}_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]\Big)^2 d\mathbb{P}_{W_k} \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} - \int \Big(\widehat{\mu}_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]\Big)^2 d\mathbb{P}_{\boldsymbol{W}} \bigg| \\
&\quad = \bigg| \int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{W_k} \Big(\prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} - \prod_{k' \neq k} d\mathbb{P}_{W_{k'}} \Big) \bigg| \\
&\quad \leq \sum_{j \neq k} \int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{W_k} \Big| d\widehat{\mathbb{P}}_{W_j,-l} - d\mathbb{P}_{W_j} \Big| \prod_{\substack{r \neq k \\ r<j}} d\widehat{\mathbb{P}}_{W_r,-l} \prod_{\substack{r \neq k \\ r>j}} d\mathbb{P}_{W_r} \\
&\quad \leq (2C)^2 \sum_{j \neq k} \int \Big| d\widehat{\mathbb{P}}_{W_j,-l} - d\mathbb{P}_{W_j} \Big| \underbrace{\int d\mathbb{P}_{W_k} \prod_{\substack{r \neq k \\ r<j}} d\widehat{\mathbb{P}}_{W_r,-l} \prod_{\substack{r \neq k \\ r>j}} d\mathbb{P}_{W_r}}_{=1} \\
&\quad = (2C)^2 \sum_{j \neq k} \int \Big| d\widehat{\mathbb{P}}_{W_j,-l} - d\mathbb{P}_{W_j} \Big| \\
&\quad = (2C)^2 \sum_{j \neq k} \int \big| \widehat{p}_{W_j,-l} - p_{W_j} \big| \, d\lambda_j.
\end{aligned}
$$

So under C2., we have

$$
\Big| \int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{W_k} \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} - \int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{\boldsymbol{W}} \Big| = o_{\mathbb{P}}(1).
$$

Now use the inequality for $x, y \geq 0$,

$$
\big|\sqrt{x} - \sqrt{y}\big| = \frac{|x-y|}{\sqrt{x}+\sqrt{y}} \leq \sqrt{|x-y|}.
$$

Applying this with

$$x = \int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{W_k} \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l},\ y = \int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{\boldsymbol{W}},$$

gives

$$\begin{aligned}
&\left(\int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{W_k} \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l}\right)^{1/2} \\
&\quad \leq \left(\int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{\boldsymbol{W}}\right)^{1/2} \\
&\quad + \left|\int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{W_k} \prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} - \int \Big(\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big)^2 d\mathbb{P}_{\boldsymbol{W}}\right|^{1/2} \\
&\quad = \Big\|\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\Big\|_{L_2(\mathbb{P}_{\boldsymbol{W}})} + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).
\end{aligned}$$

For the right-hand side term, we have

$$\begin{aligned}
&\left\| \int \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \Big(\prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} - \prod_{k' \neq k} d\mathbb{P}_{W_{k'}} \Big) \right\|_{L_2(\mathbb{P}_{W_k})} \\
&\quad \leq \sup_{w_k} \left| \int \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \Big(\prod_{k' \neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} - \prod_{k' \neq k} d\mathbb{P}_{W_{k'}} \Big) \right| \\
&\quad \leq C^2 \sum_{k' \neq k} \int \big|\widehat{p}_{W_{k'},-l} - p_{W_{k'}}\big| \, d\lambda_{k'} \ = \ o_{\mathbb{P}}(1).
\end{aligned}$$

Combining the two parts, for each k,

$$\Big\|\widehat{\mathbb{E}}_{-l}\big[\widehat{\mu}^2_{-l} \mid W_k\big] - \mathbb{E}\Big[\mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \mid W_k\Big]\Big\| = o_{\mathbb{P}}(1).$$

Similarly, for the second term,

$$\begin{aligned}
&\Big|\widehat{\mathbb{E}}_{-l}\big[\widehat{\mu}^2_{-l}\big] - \mathbb{E}\Big[\mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2\Big]\Big| \\
&\quad = \left| \int \Big(\widehat{\mu}^2_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2\Big) \prod_k d\widehat{\mathbb{P}}_{W_k,-l} + \int \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]^2 \Big(\prod_k d\widehat{\mathbb{P}}_{W_k,-l} - \prod_k d\mathbb{P}_{W_k} \Big) \right| \\
&\quad \leq 2C \left(\int \Big(\widehat{\mu}_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]\Big)^2 \prod_k d\widehat{\mathbb{P}}_{W_k,-l} \right)^{1/2} + C^2 \sum_k \int \big|\widehat{p}_{W_k,-l} - p_{W_k}\big| \, d\lambda_k \\
&\quad \leq 2C \Big\|\widehat{\mu}_{-l} - \mathbb{E}\big[Y(\boldsymbol{W}) \mid \boldsymbol{W}\big]\Big\|_{L_2(\mathbb{P}_{\boldsymbol{W}})} + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).
\end{aligned}$$

Since K is fixed, $\|b\| \leq \sum_{k \in [K]} o_{\mathbb{P}}(1) + [K] \cdot o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$. So we conclude that $\|b\| = o_{\mathbb{P}}(1)$ and $||(\theta_{\widehat{\mathbb{P}}_{-l}} - \theta)_{\text{first half}}|| \leq ||a|| + ||b|| = o_{\mathbb{P}}(1)$.

For the second half,

$$
\begin{aligned}
(\theta_{\widehat{\mathbb{P}}_{-l}} - \theta)_{\text{second half}} = \; & 2Y(\boldsymbol{W})\widehat{\mu}_{-\mathcal{S},-l} - 2\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} + \sum_{j\in-\mathcal{S}} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-\mathcal{S},-l} \mid W_j] - |-\mathcal{S}| \cdot \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-\mathcal{S},-l}] \\
& + 2\sum_{i\in\mathcal{S}} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \mid W_i] - 2|\mathcal{S}| \cdot \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-\mathcal{S},-l}] \\
& - \Big\{ 2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] - 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] + \sum_{j\in-\mathcal{S}} \mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \mid W_j\big] \\
& - |-\mathcal{S}| \cdot \mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2\big] + 2\sum_{i\in\mathcal{S}} \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] \mid W_i\right] \\
& - 2|\mathcal{S}| \cdot \mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2\big] \Big\} \\
= \; & \underbrace{2Y(\boldsymbol{W})\widehat{\mu}_{-\mathcal{S},-l} - 2\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} - 2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] + 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]}_{:=a} \\
& + \underbrace{\sum_{j\in-\mathcal{S}} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-\mathcal{S},-l} \mid W_j] - |-\mathcal{S}| \cdot \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-\mathcal{S},-l}] - \sum_{j\in-\mathcal{S}} \mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2 \mid W_j\big]}_{:=b} \\
& \underbrace{+|-\mathcal{S}| \cdot \mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2\big]}_{:=b \text{ continued}} \\
& + \underbrace{2\sum_{i\in\mathcal{S}} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \mid W_i] - 2|\mathcal{S}| \cdot \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-\mathcal{S},-l}] - 2\sum_{i\in\mathcal{S}} \mathbb{E}\left[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] \mid W_i\right]}_{:=c} \\
& + \underbrace{2|\mathcal{S}| \cdot \mathbb{E}\big[\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]^2\big]}_{:=c \text{ continued}}.
\end{aligned}
$$

Also, under C1. and C2., we have

$$
\begin{aligned}
& ||a|| = ||2Y(\boldsymbol{W})\widehat{\mu}_{-\mathcal{S},-l} - 2\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} - 2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] + 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]|| \\
= \; & ||2Y(\boldsymbol{W}) \cdot (\widehat{\mu}_{-\mathcal{S},-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]) - 2\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} + 2\widehat{\mu}_{-l}\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] \\
& - 2\widehat{\mu}_{-l}\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] + 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]|| \\
= \; & ||2(Y(\boldsymbol{W}) - \widehat{\mu}_{-l}) \cdot (\widehat{\mu}_{-\mathcal{S},-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]) - 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}] \cdot (\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}])|| \\
= \; & \max\Big\{ O_{\mathbb{P}}(||\widehat{\mu}_{-\mathcal{S},-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}]||), O_{\mathbb{P}}(||\widehat{\mu}_{-l} - \mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]||) \Big\} \\
= \; & o_{\mathbb{P}}(1).
\end{aligned}
$$

We can also get $||b|| = o_{\mathbb{P}}(1)$ and $||c|| = o_{\mathbb{P}}(1)$, where the proofs are similar to that in the first half. So we can conclude that $||c|| = o_{\mathbb{P}}(1)$ and $||(\theta_{\widehat{\mathbb{P}}_{-l}} - \theta)_{\text{second half}}|| \leq ||a|| + ||b|| + ||c|| = o_{\mathbb{P}}(1)$, and the empirical process term is of the order $o_{\mathbb{P}}(1/\sqrt{n})$.

The exact form of the first half of the remainder term $R(\widehat{\mathbb{P}}_{-l}, \mathbb{P}) = \Theta_{\widehat{\mathbb{P}}_{-l}} - \Theta + \mathbb{P}\{\theta_{\widehat{\mathbb{P}}_{-l}}\}$ is

$$\begin{aligned}
R(\widehat{\mathbb{P}}_{-l}, \mathbb{P})_{\text{first half}} &= \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2] - \int \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}\right]^2 d\mathbb{P}_{\boldsymbol{W}} \\
&\quad + \int \Big\{2Y(\boldsymbol{W})\widehat{\mu}_{-l} - 2\widehat{\mu}_{-l}^2 + \sum_{k\in[K]} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2 \mid W_k] - [K]\cdot\widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2]\Big\} d\mathbb{P}_{\boldsymbol{W}} \\
&= \underbrace{\int \Big\{2Y(\boldsymbol{W})\widehat{\mu}_{-l} - \widehat{\mu}_{-l}^2 - \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}\right]^2\Big\} d\mathbb{P}_{\boldsymbol{W}}}_{:=a} \\
&\quad + \underbrace{\int \Big\{ \sum_{k\in[K]} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2 \mid W_k] - [K]\cdot\widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2] + \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2] - \widehat{\mu}_{-l}^2\Big\} d\mathbb{P}_{\boldsymbol{W}}}_{:=b}.
\end{aligned}$$

Note that, under C1. and C3.,

$$\begin{aligned}
|a| &= \Big|\int \Big\{2Y(\boldsymbol{W})\widehat{\mu}_{-l} - \widehat{\mu}_{-l}^2 - \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}\right]^2\Big\} d\mathbb{P}_{\boldsymbol{W}}\Big| \\
&= \int (\widehat{\mu}_{-l} - \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}\right])^2 d\mathbb{P}_{\boldsymbol{W}} \\
&= ||\widehat{\mu}_{-l} - \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}\right]||^2 \\
&= o_{\mathbb{P}}(1/\sqrt{n}).
\end{aligned}$$

We observe that, for b,

$$\int \left(\widehat{\mu}_{-l}^2 - \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2]\right) d\mathbb{P}_{\boldsymbol{W}} = \int \widehat{\mu}_{-l}^2 \prod_{k\in[K]} d\mathbb{P}_{W_k} - \int \widehat{\mu}_{-l}^2 \prod_{k\in[K]} d\widehat{\mathbb{P}}_{W_k,-l}.$$

Expanding the difference of product measures yields the exact identity

$$\begin{aligned}
\int \left(\widehat{\mu}_{-l}^2 - \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2]\right) d\mathbb{P}_{\boldsymbol{W}} &= \int \sum_{k\in[K]} \widehat{\mu}_{-l}^2 \Big(d\mathbb{P}_{W_k} - d\widehat{\mathbb{P}}_{W_k,-l}\Big) \prod_{k'\neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} \\
&\quad + \int \sum_{k\neq k'} \widehat{\mu}_{-l}^2 \Big(d\mathbb{P}_{W_k} - d\widehat{\mathbb{P}}_{W_k,-l}\Big)\Big(d\mathbb{P}_{W_{k'}} - d\widehat{\mathbb{P}}_{W_{k'},-l}\Big) \prod_{k''\neq k,k'} d\widehat{\mathbb{P}}_{W_{k''},-l} \\
&\quad + \text{higher-order terms.}
\end{aligned}$$

Moreover, for each fixed $k \in [K]$,

$$\int \widehat{\mu}_{-l}^2 \Big(d\mathbb{P}_{W_k} - d\widehat{\mathbb{P}}_{W_k,-l}\Big) \prod_{k'\neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} = \int \left(\widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2 \mid W_k] - \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2]\right) d\mathbb{P}_{\boldsymbol{W}},$$

so summing over k gives

$$\begin{aligned}
\int \left(\widehat{\mu}_{-l}^2 - \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2]\right) d\mathbb{P}_{\boldsymbol{W}} &= \int \Big(\sum_{k\in[K]} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2 \mid W_k] - [K]\cdot\widehat{\mathbb{E}}_{-l}[\widehat{\mu}_{-l}^2]\Big) d\mathbb{P}_{\boldsymbol{W}} \\
&\quad + \int \sum_{k\neq k'} \widehat{\mu}_{-l}^2 \Big(d\mathbb{P}_{W_k} - d\widehat{\mathbb{P}}_{W_k,-l}\Big)\Big(d\mathbb{P}_{W_{k'}} - d\widehat{\mathbb{P}}_{W_{k'},-l}\Big) \prod_{k''\neq k,k'} d\widehat{\mathbb{P}}_{W_{k''},-l} \\
&\quad + \text{higher-order terms.}
\end{aligned}$$

By the definition of b,

$$b = \int \Big(\sum_{k\in[K]} \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-l} \mid W_k] - [K]\cdot \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-l}] \Big)\, d\mathbb{P}_{\boldsymbol{W}} - \int \Big(\widehat{\mu}^2_{-l} - \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-l}] \Big)\, d\mathbb{P}_{\boldsymbol{W}}$$
$$= -\int \sum_{k\neq k'} \widehat{\mu}^2_{-l} \Big(d\mathbb{P}_{W_k} - d\widehat{\mathbb{P}}_{W_k,-l} \Big) \Big(d\mathbb{P}_{W_{k'}} - d\widehat{\mathbb{P}}_{W_{k'},-l} \Big) \prod_{k''\neq k,k'} d\widehat{\mathbb{P}}_{W_{k''},-l} - \text{higher-order terms.}$$

Consequently, under C1.,

$$|b| \le O_{\mathbb{P}}\left(\int \max_k \big(p_{W_k} - \widehat{p}_{W_k,-l}\big)^2 \prod_{k''\neq k,k'} d\widehat{\mathbb{P}}_{W_{k''},-l}\, d\lambda_k\, d\lambda_{k'} \right)$$
$$= O_{\mathbb{P}}\left(|| \frac{\max_k (p_{W_k} - \widehat{p}_{W_k,-l})}{\widehat{p}_{W_k}\widehat{p}_{W_{k'}}} ||^2_{L_2(\widehat{\mathbb{P}}_{-l})} \right).$$

Fix a fold l and write, for a fixed $k' \in [K]$,

$$\Delta_{k,-l}(w_k) := p_{W_k} - \widehat{p}_{W_k,-l},\ r_{k,-l}(w) := \frac{\Delta_{k,-l}(w_k)}{\widehat{p}_{W_k}\,\widehat{p}_{W_{k'}}},$$

so that

$$\Big\| \max_k r_{k,-l} \Big\|^2_{L_2(\widehat{\mathbb{P}}_{-l})} = \widehat{\mathbb{P}}_{-l}\Big[\Big(\max_k r_{k,-l} \Big)^2 \Big].$$

Using the inequality $(\max_k |a_k|)^2 \le \sum_{k=1}^K a_k^2$, we obtain

$$\Big\| \max_{k\in[K]} r_{k,-l} \Big\|^2_{L_2(\widehat{\mathbb{P}}_{-l})} = \widehat{\mathbb{P}}_{-l}\Big[\Big(\max_{k\in[K]} r_{k,-l} \Big)^2 \Big]$$
$$\le \widehat{\mathbb{P}}_{-l}\Big[\sum_{k=1}^K r^2_{k,-l} \Big] = \sum_{k=1}^K \widehat{\mathbb{P}}_{-l}\big[r^2_{k,-l} \big].$$

Under C4., for every $k \in [K]$,

$$\frac{1}{\widehat{p}^2_{W_k}\,\widehat{p}^2_{W_{k'}}} \ \le\ \underline{c}^{-4},$$

and hence

$$\widehat{\mathbb{P}}_{-l}\big[r^2_{k,-l} \big] = \widehat{\mathbb{P}}_{-l}\Big[\frac{\Delta_{k,-l}(W_k)^2}{\widehat{p}^2_{W_k}\,\widehat{p}^2_{W_{k'}}} \Big]$$
$$\le \underline{c}^{-4}\, \widehat{\mathbb{P}}_{-l}\big[\Delta_{k,-l}(W_k)^2 \big] = \underline{c}^{-4}\, \|\Delta_{k,-l}\|^2_{L_2(\widehat{\mathbb{P}}_{-l})}.$$

Combining the displays yields

$$\Big\| \max_{k\in[K]} r_{k,-l} \Big\|^2_{L_2(\widehat{\mathbb{P}}_{-l})} \le \underline{c}^{-4} \sum_{k=1}^K \|\Delta_{k,-l}\|^2_{L_2(\widehat{\mathbb{P}}_{-l})}.$$

Decompose

$$\begin{aligned}\|\Delta_{k,-l}\|^2_{L_2(\widehat{\mathbb{P}}_{-l})} &= \widehat{\mathbb{P}}_{-l}\big[\Delta^2_{k,-l}\big] \\ &= \mathbb{P}\big[\Delta^2_{k,-l}\big] + (\widehat{\mathbb{P}}_{-l} - \mathbb{P})\Delta^2_{k,-l} \\ &= \|\Delta_{k,-l}\|^2_{L_2(\mathbb{P})} + (\widehat{\mathbb{P}}_{-l} - \mathbb{P})\Delta^2_{k,-l}. \end{aligned} \tag{22}$$

By Lemma 7, and moreover assume

$$\|\Delta^2_{k,-l}\|_{L_2(\mathbb{P})} = \|\Delta_{k,-l}\|^2_{L_4(\mathbb{P})} \xrightarrow{\mathbb{P}} 0, \tag{23}$$

which holds under C1. that $\|\Delta_{k,-l}\|_\infty = O_{\mathbb{P}}(1)$, since then $\mathbb{P}[\Delta^4_{k,-l}] \le \|\Delta_{k,-l}\|^2_\infty \mathbb{P}[\Delta^2_{k,-l}] \to 0$.

Let $\mathbb{G}_{-l} := \sqrt{n_{-l}}(\widehat{\mathbb{P}}_{-l} - \mathbb{P})$ be the empirical process. Because $\mathcal{F}$ is Donsker by C5., $\mathbb{G}_{-l}$ is asymptotically tight in $\ell^\infty(\mathcal{F})$ and is asymptotically uniformly equicontinuous w.r.t. the $L_2(\mathbb{P})$ semimetric (van der Vaart and Wellner, 2023).

Let $d(\cdot,\cdot)$ be a semimetric on $\mathcal{F}$ and define the event $A := \{d(f_n, 0) < \delta\}$ for any $\delta > 0$. On A, the pair $(f_n, 0)$ is admissible for the supremum over $\{(f,g) : d(f,g) < \delta\}$, hence

$$|\mathbb{G}_{-l}(f_n) - \mathbb{G}_{-l}(0)| \le \sup_{\substack{f,g\in\mathcal{F}:\\ d(f,g)<\delta}} |\mathbb{G}_{-l}(f) - \mathbb{G}_{-l}(g)|.$$

Therefore,

$$\Big\{|\mathbb{G}_{-l}(f_n) - \mathbb{G}_{-l}(0)| > \varepsilon\Big\} \cap A \subseteq \left\{ \sup_{\substack{f,g\in\mathcal{F}:\\ d(f,g)<\delta}} |\mathbb{G}_{-l}(f) - \mathbb{G}_{-l}(g)| > \varepsilon \right\}.$$

Splitting the event by A vs. A^c yields

$$\begin{aligned}\mathbb{P}\Big(|\mathbb{G}_{-l}(f_n) - \mathbb{G}_{-l}(0)| > \varepsilon\Big) &\le \mathbb{P}\big(A^c\big) + \mathbb{P}\Big(\big\{|\mathbb{G}_{-l}(f_n) - \mathbb{G}_{-l}(0)| > \varepsilon\big\} \cap A\Big) \\ &\le \mathbb{P}\big(d(f_n,0) \ge \delta\big) + \mathbb{P}\left(\sup_{\substack{f,g\in\mathcal{F}:\\ d(f,g)<\delta}} |\mathbb{G}_{-l}(f) - \mathbb{G}_{-l}(g)| > \varepsilon \right).\end{aligned}$$

Hence, with $f_n := \Delta^2_{k,-l} \in \mathcal{F}$ and $\|f_n - 0\|_{L_2(\mathbb{P})} \xrightarrow{\mathbb{P}} 0$ by (23), we have

$$\mathbb{G}_{-l}(f_n) - \mathbb{G}_{-l}(0) \xrightarrow{\mathbb{P}} 0,$$

and since $\mathbb{G}_{-l}(0) = 0$, it follows that

$$\sqrt{n_{-l}}\,(\widehat{\mathbb{P}}_{-l} - \mathbb{P})\Delta^2_{k,-l} = \mathbb{G}_{-l}(\Delta^2_{k,-l}) = o_{\mathbb{P}}(1), \text{ i.e. } (\widehat{\mathbb{P}}_{-l} - \mathbb{P})\Delta^2_{k,-l} = o_{\mathbb{P}}(1/\sqrt{n}).$$

Combining with (22) gives

$$\|\Delta_{k,-l}\|^2_{L_2(\widehat{\mathbb{P}}_{-l})} = \|\Delta_{k,-l}\|^2_{L_2(\mathbb{P})} + o_{\mathbb{P}}(1/\sqrt{n}) = o_{\mathbb{P}}(1/\sqrt{n}).$$

Under positivity $\inf_k p_{W_k} \geq \underline{c} > 0$ and fixed K,

$$\Big\| \max_{k\in[K]} r_{k,-l} \Big\|^2_{L_2(\widehat{\mathbb{P}}_{-l})} \leq \underline{c}^{-4} \sum_{k=1}^{K} \|\Delta_{k,-l}\|^2_{L_2(\widehat{\mathbb{P}}_{-l})} = o_{\mathbb{P}}(1/\sqrt{n}).$$

So we get $|R(\widehat{\mathbb{P}}_{-l}, \mathbb{P})_{\text{first half}}| \leq |a| + |b| = o_{\mathbb{P}}(1/\sqrt{n})$.

For the second half,

$$\begin{aligned}
&R(\widehat{\mathbb{P}}_{-l}, \mathbb{P})_{\text{second half}} \\
&= \widehat{\mathbb{E}}_{-l}[\widehat{\mu}^2_{-\mathcal{S},-l}] - \int \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2 d\mathbb{P}_{\boldsymbol{w}_{-\mathcal{S}}} \\
&\quad + \int \Big\{ 2Y(\boldsymbol{W})\widehat{\mu}_{-\mathcal{S},-l} - 2\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} + \sum_{j\in-\mathcal{S}} \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}^2_{-\mathcal{S},-l} \mid W_j\right] - |-\mathcal{S}| \cdot \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}^2_{-\mathcal{S},-l}\right] \\
&\quad + 2\sum_{i\in\mathcal{S}} \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \mid W_i\right] - 2|\mathcal{S}| \cdot \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}^2_{-\mathcal{S},-l}\right] \Big\} d\mathbb{P}_{\boldsymbol{W}} \\
&= \int \Big\{ 2Y(\boldsymbol{W})\widehat{\mu}_{-\mathcal{S},-l} - \widehat{\mu}^2_{-\mathcal{S},-l} - \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2 \Big\} d\mathbb{P}_{\boldsymbol{W}} \\
&\quad + \int \Big\{ \sum_{j\in-\mathcal{S}} \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}^2_{-\mathcal{S},-l} \mid W_j\right] - |-\mathcal{S}| \cdot \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}^2_{-\mathcal{S},-l}\right] + \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}^2_{-\mathcal{S},-l}\right] - \widehat{\mu}^2_{-\mathcal{S},-l} \Big\} d\mathbb{P}_{\boldsymbol{W}} \\
&\quad + 2\int \Big\{ \sum_{k\in[K]} \widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \mid W_k] - |K| \cdot \widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l}] + \widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l}] - \widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \Big\} d\mathbb{P}_{\boldsymbol{W}} \\
&\quad - 2\int \Big\{ \sum_{j\in-\mathcal{S}} \widehat{\mathbb{E}}[\widehat{\mu}^2_{-\mathcal{S},-l} \mid W_j] - |-\mathcal{S}| \cdot \widehat{\mathbb{E}}[\widehat{\mu}^2_{-\mathcal{S},-l}] + \widehat{\mathbb{E}}[\widehat{\mu}^2_{-\mathcal{S},-l}] - \widehat{\mu}^2_{-\mathcal{S},-l} \Big\} d\mathbb{P}_{\boldsymbol{W}} \\
&= \underbrace{\int \Big\{ 2Y(\boldsymbol{W})\widehat{\mu}_{-\mathcal{S},-l} - \widehat{\mu}^2_{-\mathcal{S},-l} - \mathbb{E}\left[Y(\boldsymbol{W}) \mid \boldsymbol{W}_{-\mathcal{S}}\right]^2 \Big\} d\mathbb{P}_{\boldsymbol{W}}}_{:=a} \\
&\quad \underbrace{- \int \Big\{ \sum_{j\in-\mathcal{S}} \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}^2_{-\mathcal{S},-l} \mid W_j\right] - |-\mathcal{S}| \cdot \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}^2_{-\mathcal{S},-l}\right] + \widehat{\mathbb{E}}_{-l}\left[\widehat{\mu}^2_{-\mathcal{S},-l}\right] - \widehat{\mu}^2_{-\mathcal{S},-l} \Big\} d\mathbb{P}_{\boldsymbol{W}}}_{:=b} \\
&\quad \underbrace{+ 2\int \Big\{ \sum_{k\in[K]} \widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \mid W_k] - |K| \cdot \widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l}] + \widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l}] - \widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \Big\} d\mathbb{P}_{\boldsymbol{W}}}_{:=c}
\end{aligned}$$

The second equality holds because $\widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l}] = \widehat{\mathbb{E}}\left[\widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \mid \boldsymbol{W}_{-\mathcal{S}}]\right] = \widehat{\mathbb{E}}[\widehat{\mu}^2_{-\mathcal{S},-l}]$ and similarly $\widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \mid W_j] = \widehat{\mathbb{E}}\left[\widehat{\mathbb{E}}[\widehat{\mu}_{-l}\widehat{\mu}_{-\mathcal{S},-l} \mid \boldsymbol{W}_{-\mathcal{S}}]|W_j\right] = \widehat{\mathbb{E}}[\widehat{\mu}^2_{-\mathcal{S},-l} \mid W_j]$, for $j \in -\mathcal{S}$. It is obvious that $|a|$, $|b|$ and $|c|$ are of the order $o_{\mathbb{P}}(1/\sqrt{n})$ from the previous results. So if C1. and C3. hold, then the above results enable us to conclude that the remainder term is also of the order $o_{\mathbb{P}}(1/\sqrt{n})$. To sum up, we have $\widehat{\Theta} - \Theta = \frac{1}{L}\sum_{l=1}^{L}(\widehat{\Theta}_l - \Theta) = \frac{1}{L}\sum_{l=1}^{L}(\mathbb{P}_n^l - \mathbb{P})\{\theta\} + o_{\mathbb{P}}(1/\sqrt{n})$.

For the denominator, we denote its efficient influence function by η. The exact form of the empirical process term is

$$
\begin{aligned}
&\eta_{\widehat{\mathbb{P}}_{-l}} - \eta \\
=&Y(\boldsymbol{W})^2 - 2Y(\boldsymbol{W})\widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})] - \widehat{v}_{-l} + 2\widehat{\mu}_{-l}\widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})] + \sum_{k\in[K]} \widehat{v}_{k,-l} \\
&- 2\sum_{k\in[K]} \widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})]\widehat{\mu}_{k,-l} - |K|\cdot\widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})^2] + 2|K|\cdot\widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})]^2 \\
&- \Big\{Y(\boldsymbol{W})^2 - 2Y(\boldsymbol{W})\mathbb{E}[Y(\boldsymbol{W})] - \mathbb{E}[Y(\boldsymbol{W})^2 \mid \boldsymbol{W}] + 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\mathbb{E}[Y(\boldsymbol{W})]\Big\} \\
&- \Big\{\sum_{k\in[K]} \mathbb{E}[Y(\boldsymbol{W})^2 \mid W_k] - 2\sum_{k\in[K]} \mathbb{E}[Y(\boldsymbol{W})]\mathbb{E}[Y(\boldsymbol{W}) \mid W_k] - |K|\cdot\mathbb{E}[Y(\boldsymbol{W})^2] + 2|K|\cdot\mathbb{E}[Y(\boldsymbol{W})]^2\Big\} \\
=&2Y(\boldsymbol{W})\cdot\Big(\mathbb{E}[Y(\boldsymbol{W})] - \widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})]\Big) + \big(\mathbb{E}[Y(\boldsymbol{W})^2 \mid \boldsymbol{W}] - \widehat{v}_{-l}\big) \\
&+ \Big(2\widehat{\mu}_{-l}\widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})] - 2\mathbb{E}[Y(\boldsymbol{W}) \mid \boldsymbol{W}]\mathbb{E}[Y(\boldsymbol{W})]\Big) \\
&+ \sum_{k\in[K]} \big(\widehat{v}_{k,-l} - \mathbb{E}[Y(\boldsymbol{W})^2 \mid W_k]\big) + 2\sum_{k\in[K]} \Big(\mathbb{E}[Y(\boldsymbol{W})]\mathbb{E}[Y(\boldsymbol{W}) \mid W_k] - \widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})]\widehat{\mu}_{k,-l}\Big) \\
&+ |K|\cdot\Big(\mathbb{E}[Y(\boldsymbol{W})^2] - \widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})^2]\Big) + 2|K|\cdot\Big(\mathbb{E}[Y(\boldsymbol{W})]^2 - \widehat{\mathbb{E}}_{-l}[Y(\boldsymbol{W})]^2\Big),
\end{aligned}
$$

which can be easily proved to convergence at the rate of $o_{\mathbb{P}}(1)$ using the conditions and techniques we applied in the previous steps.

We then move on to the reminder term:

$$
\begin{aligned}
&\mathsf{Var}(Y)_{\widehat{\mathbb{P}}_{-l}} - \mathsf{Var}(Y) + \mathbb{P}\{\eta_{\widehat{\mathbb{P}}_{-l}}\} \\
=&\widehat{\mathbb{E}}_{-l}[Y^2] - \widehat{\mathbb{E}}_{-l}[Y]^2 - \mathbb{E}[Y^2] + \mathbb{E}[Y]^2 \\
&+ \int \Big\{Y^2 - 2Y\cdot\widehat{\mathbb{E}}_{-l}[Y] - \widehat{v}_{-l} + 2\widehat{\mu}_{-l}\cdot\widehat{\mathbb{E}}_{-l}[Y] + \sum_{k\in[K]} \widehat{v}_{k,-l} \\
&- 2\sum_{k\in[K]} \widehat{\mu}_{k,-l}\cdot\widehat{\mathbb{E}}_{-l}[Y] - |K|\cdot\widehat{\mathbb{E}}_{-l}[Y^2] + 2|K|\cdot\widehat{\mathbb{E}}_{-l}[Y]^2\Big\} d\mathbb{P} \\
=&\underbrace{\int \Big\{\mathbb{E}[Y]^2 - 2Y\cdot\widehat{\mathbb{E}}_{-l}[Y] + \widehat{\mathbb{E}}_{-l}[Y]^2\Big\} d\mathbb{P}}_{:=a} + \underbrace{\int \Big\{\widehat{\mathbb{E}}_{-l}[Y^2] - \widehat{v}_{-l} + \sum_{k\in[K]} \widehat{v}_{k,-l} - |K|\cdot\widehat{\mathbb{E}}_{-l}[Y^2]\Big\} d\mathbb{P}}_{:=b} \\
&+ \underbrace{\int \Big\{2\widehat{\mu}_{-l}\cdot\widehat{\mathbb{E}}_{-l}[Y] - 2\widehat{\mathbb{E}}_{-l}[Y]^2 + 2|K|\cdot\widehat{\mathbb{E}}_{-l}[Y]^2 - 2\sum_{k\in[K]} \widehat{\mu}_{k,-l}\cdot\widehat{\mathbb{E}}_{-l}[Y]\Big\} d\mathbb{P}}_{:=c}.
\end{aligned}
$$

For a,

$$
\begin{aligned}
&\int \Big\{\mathbb{E}[Y]^2 - 2Y\cdot\widehat{\mathbb{E}}_{-l}[Y] + \widehat{\mathbb{E}}_{-l}[Y]^2\Big\} d\mathbb{P} \\
=&\int \Big(\widehat{\mathbb{E}}_{-l}[Y] - \mathbb{E}[Y]\Big)^2 d\mathbb{P},
\end{aligned}
$$

where $\widehat{\mathbb{E}}_{-l}[Y]$ is the sample mean defined on the data excluding the l-th fold. Here we have $\widehat{\mathbb{E}}_{-l}[Y] - \mathbb{E}[Y] = o_{\mathbb{P}}(n^{-1/4})$ since by Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\widehat{\mathbb{E}}_{-l}[Y] - \mathbb{E}[Y]\right| > \varepsilon n_{-l}^{-1/4}\right) \leq \frac{\mathsf{Var}\left(\widehat{\mathbb{E}}_{-l}[Y]\right)}{\varepsilon^2 n_{-l}^{-1/2}} = \frac{\mathsf{Var}(Y)}{\varepsilon^2 n_{-l}^{1/2}} \xrightarrow{n\to\infty} 0.$$

For b,

$$\begin{aligned}
&\int \left\{\widehat{v}_{-l} - \widehat{\mathbb{E}}_{-l}[Y^2]\right\} d\mathbb{P} \\
=&\int Y^2 d\widehat{\mathbb{P}}_{Y|\boldsymbol{W}} d\mathbb{P}_{\boldsymbol{W}} - \int Y^2 d\widehat{\mathbb{P}}_Y \\
=&\int Y^2 d\widehat{\mathbb{P}}_{Y|\boldsymbol{W}} (\prod_k d\mathbb{P}_{W_k} - \prod_k d\widehat{\mathbb{P}}_{W_k,-l}) \\
=&\int \sum_{k\in[K]} Y^2 d\widehat{\mathbb{P}}_{Y|\boldsymbol{W}} (d\mathbb{P}_{W_k} - d\widehat{\mathbb{P}}_{W_k,-l}) \prod_{k'\neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} \\
&+ \int \sum_{k\neq k'} Y^2 (d\mathbb{P}_{W_k} - d\widehat{\mathbb{P}}_{W_k,-l})(d\mathbb{P}_{W_{k'}} - d\widehat{\mathbb{P}}_{W_{k'},-l}) \prod_{k''\neq k,k'} d\widehat{\mathbb{P}}_{W_{k''},-l} + \text{higher-order terms} \\
=&\int \Big\{ \sum_{k\in[K]} \widehat{v}_{k,-l} - |K| \cdot \widehat{\mathbb{E}}_{-l}[Y^2] \Big\} d\mathbb{P} + O_{\mathbb{P}}(\int \max_k (p_{W_k} - \widehat{p}_{W_k,-l})^2 \prod_{k''\neq k,k'} d\widehat{\mathbb{P}}_{W_{k''},-l} d\lambda_k d\lambda_{k'}).
\end{aligned}$$

For c,

$$\begin{aligned}
&\int \left\{2\widehat{\mu}_{-l} \cdot \widehat{\mathbb{E}}_{-l}[Y] - 2\widehat{\mathbb{E}}_{-l}[Y]^2\right\} d\mathbb{P} \\
=&2\widehat{\mathbb{E}}_{-l}[Y] \cdot \Big\{ \int Y d\widehat{\mathbb{P}}_{Y|\boldsymbol{W}} d\mathbb{P}_{\boldsymbol{W}} - \int Y d\widehat{\mathbb{P}}_Y \Big\} \\
=&2\widehat{\mathbb{E}}_{-l}[Y] \cdot \Big\{ \int Y d\widehat{\mathbb{P}}_{Y|\boldsymbol{W}} (\prod_k d\mathbb{P}_{W_k} - \prod_k d\widehat{\mathbb{P}}_{W_k,-l}) \Big\} \\
=&2\widehat{\mathbb{E}}_{-l}[Y] \cdot \Big\{ \int \sum_{k\in[K]} Y d\widehat{\mathbb{P}}_{Y|\boldsymbol{W}} (d\mathbb{P}_{W_k} - d\widehat{\mathbb{P}}_{W_k,-l}) \prod_{k'\neq k} d\widehat{\mathbb{P}}_{W_{k'},-l} \\
&+ \int \sum_{k\neq k'} Y (d\mathbb{P}_{W_k} - d\widehat{\mathbb{P}}_{W_k,-l})(d\mathbb{P}_{W_{k'}} - d\widehat{\mathbb{P}}_{W_{k'},-l}) \prod_{k''\neq k,k'} d\widehat{\mathbb{P}}_{W_{k''},-l} \Big\} + \text{higher-order terms} \\
=&2\widehat{\mathbb{E}}_{-l}[Y] \cdot \Big\{ \int \Big\{ \sum_{k\in[K]} \widehat{\mu}_{k,-l} - |K| \cdot \widehat{\mathbb{E}}_{-l}[Y] \Big\} d\mathbb{P} \Big\} + O_{\mathbb{P}}(\int \max_k (p_{W_k} - \widehat{p}_{W_k,-l})^2 \prod_{k''\neq k,k'} d\widehat{\mathbb{P}}_{W_{k''},-l} d\lambda_k d\lambda_{k'}).
\end{aligned}$$

The above terms can also be easily proved to convergence at the rate of $o_{\mathbb{P}}(1/\sqrt{n})$ using the conditions and techniques we applied in the previous steps.

Then by the delta method,

$$\begin{aligned}\widehat{\xi} - \xi &= \frac{\widehat{\Theta}}{\widehat{\mathsf{Var}}(Y)} - \frac{\Theta}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)} \\ &= (\mathbb{P}_n - \mathbb{P})\Big\{\frac{\mathsf{Var}\left(Y(\boldsymbol{W})\right)\theta - \Theta\eta}{\mathsf{Var}\left(Y(\boldsymbol{W})\right)^2}\Big\} + o_{\mathbb{P}}(1/\sqrt{n}) \\ &= (\mathbb{P}_n - \mathbb{P})\{\varphi\} + o_{\mathbb{P}}(1/\sqrt{n}).\end{aligned}$$

If $\mathsf{Var}\{\varphi\} < \infty$ holds, by the Central Limit Theorem,

$$\sqrt{n}\left(\widehat{\xi} - \xi\right) \rightsquigarrow \mathcal{N}\big(0, \mathsf{Var}\{\varphi\}\big).$$

That is, $\widehat{\xi}$ is a root-n consistent and asymptotically normal estimator.

■

D.5. Proof of Proposition 5

Proof

$$\begin{aligned}\xi(W_{\mathcal{S}}) = 0 \;&\Leftrightarrow \mathsf{Var}\big(Y(W) - Y(W'_{\mathcal{S}}, W_{-\mathcal{S}})\big) = 0 \\ &\Leftrightarrow Y(w_{\mathcal{S}}, w_{-\mathcal{S}}) = Y(w'_{\mathcal{S}}, w_{-\mathcal{S}}), \quad \forall w_{\mathcal{S}}, w'_{\mathcal{S}}, w_{-\mathcal{S}}.\end{aligned} \tag{24}$$

■

D.6. Proof of Proposition 6

Proof By Eq. (14), under Eq. (13) implied by the null $\xi(W_{\mathcal{S}}) = 0$, we have

$$(W^*_{\mathcal{S}}, W_{-\mathcal{S}}, Y) \overset{d}{=} (W_{\mathcal{S}}, W_{-\mathcal{S}}, Y).$$

The distributional equivalence guarantees that the reference distribution generated by the randomization procedure matches the null distribution of the test statistic, thereby ensuring the validity of the randomization test. ■

D.7. Proof of Proposition 7

Proof For $\xi_Y(W_{\mathcal{S}}) = 0$, by Proposition 6, the first step of the sequential test will not reject with probability at least $1 - \alpha$, and the resulting interval $\{0\}$ covers the truth.

For $\xi_Y(W_{\mathcal{S}}) > 0$, As the power of the randomization test converges to one, the first step will be passed with probability approaching one. Then, by Corollary 2, the subsequent confidence interval based on the central limit theorem achieves the correct asymptotic coverage.

If the power of the randomization test does not converge to one and the first-step interval is replaced by $[0, +\infty)$, the coverage error is upper bounded by that of Eq. (12) when $\xi_Y(W_{\mathcal{S}}) > 0$, and zero when $\xi_Y(W_{\mathcal{S}}) = 0$. ■

Appendix E. Additional tables

Expression	**Meaning**
Y	Outcome
$\boldsymbol{W} = (W_1, \ldots, W_k)$	Vector of factors
$\boldsymbol{W}_{\mathcal{S}}$	Subvector of $\boldsymbol{W}$ indexed by $\mathcal{S} \subseteq [K]$
W_k	k-th component of $\boldsymbol{W}$
$\mathbb{P}/\widehat{\mathbb{P}}$	True/estimated joint distribution of $(Y, \boldsymbol{W})$
$\mathbb{P}_{\boldsymbol{W}}/\widehat{\mathbb{P}}_{\boldsymbol{W}}$	True/estimated marginal distribution of $\boldsymbol{W}$
$\mathbb{P}_{W_k}/\widehat{\mathbb{P}}_{W_k}$	True/estimated marginal distribution of W_k
$\mathbb{P}_{Y\mid\boldsymbol{W}}/\widehat{\mathbb{P}}_{Y\mid\boldsymbol{W}}$	True/estimated conditional distribution of Y given $\boldsymbol{W}$
$\mathbb{E}[\cdot]/\widehat{\mathbb{E}}[\cdot]$	Expectation under $\mathbb{P}/\widehat{\mathbb{P}}$
$\mathcal{P}$	Probability model (set of data-generating distributions)
$\mathcal{P}_{\text{np}}$	Non-parametric probability model (no structural restrictions)
$\mathcal{P}_{\text{ind}}$	Probability model with independent components of $\boldsymbol{W}$
$\dot{\mathcal{P}}_{\mathbb{P}}$	Tangent space at $\mathbb{P} \in \mathcal{P}$
ξ	Causal estimand of interest
$\xi(\vee_{k\in\mathcal{S}} W_k), \mathcal{S} \subseteq [K]$	Causal explainability attributable to the union of factors in $\mathcal{S}$
$\xi(W_k \wedge W_{k'})$	Causal explainability attributable to the interaction of W_k and $W_{k'}$
$\varphi^{\xi}_{\text{IF}}$, $\varphi^{\xi}_{\text{EIF}}$, φ^{ξ}	(Efficient) influence functions for ξ

Table 3: Table of notations. When unambiguous, superscripts or subscripts may be omitted.

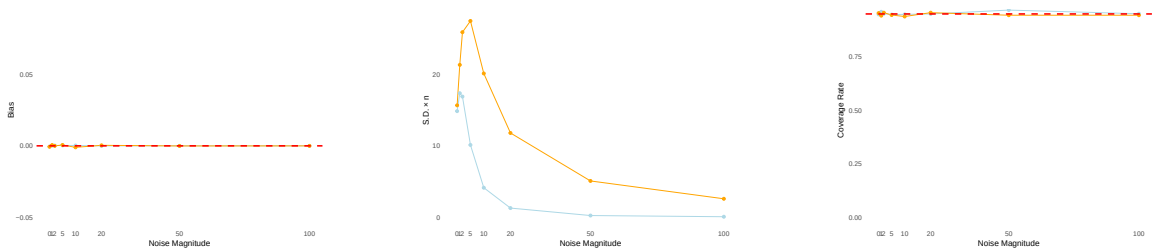
Expression	Formula
$\widehat{\mathbb{E}}[Y \mid \boldsymbol{W} = \boldsymbol{w}]$	$\widehat{\mu}(\boldsymbol{W})$
$\widehat{\mathbb{E}}[Y \mid \boldsymbol{W}_{-\mathcal{S}} = \boldsymbol{w}_{-\mathcal{S}}]$	$\int \widehat{\mu}(\boldsymbol{w}) d\widehat{\mathbb{P}}(\boldsymbol{w}_{\mathcal{S}})$
$\widehat{\mathbb{E}}[\widehat{\mathbb{E}}^2[Y \mid \boldsymbol{W}_{-\mathcal{S}}]]$	$\int \left(\int \widehat{\mu}(\boldsymbol{w}) d\widehat{\mathbb{P}}(\boldsymbol{w}_{\mathcal{S}})\right)^2 d\widehat{\mathbb{P}}(\boldsymbol{w}_{-\mathcal{S}})$
$\widehat{\mathbb{E}}\left[\widehat{\mathbb{E}}[Y \mid \boldsymbol{W}] \cdot \widehat{\mathbb{E}}[Y \mid \boldsymbol{W}_{-\mathcal{S}}] \mid W_i = W_i\right], i \in \mathcal{S}$	$\int \widehat{\mu}(\boldsymbol{w}) \cdot \left(\int \widehat{\mu}(\boldsymbol{w}) d\widehat{\mathbb{P}}(\boldsymbol{w}_{\mathcal{S}})\right) d\widehat{\mathbb{P}}(\boldsymbol{w}_{-i})$
$\widehat{\mathbb{E}}\left[\widehat{\mathbb{E}}^2[Y \mid \boldsymbol{W}_{-\mathcal{S}}] \mid W_j = W_j\right], j \notin \mathcal{S}$	$\int \left(\int \widehat{\mu}(\boldsymbol{w}) d\widehat{\mathbb{P}}(\boldsymbol{w}_{\mathcal{S}})\right)^2 d\widehat{\mathbb{P}}(\boldsymbol{w}_{-j})$
$\mathsf{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}^2[Y]$	$\int \widehat{\nu}(\boldsymbol{w}) d\widehat{\mathbb{P}}(\boldsymbol{w}) - \left(\int \widehat{\mu}(\boldsymbol{w}) d\widehat{\mathbb{P}}(\boldsymbol{w})\right)^2$

Table 4: Formulas for the one-step correction estimator.

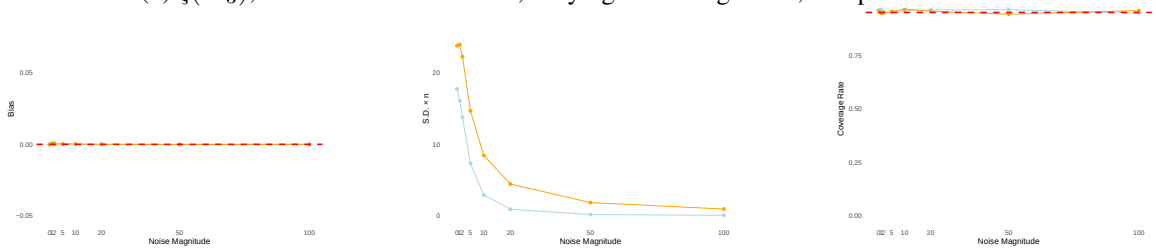
Appendix F. Additional empirical studies

F.1. Additional simulations with true nuisances

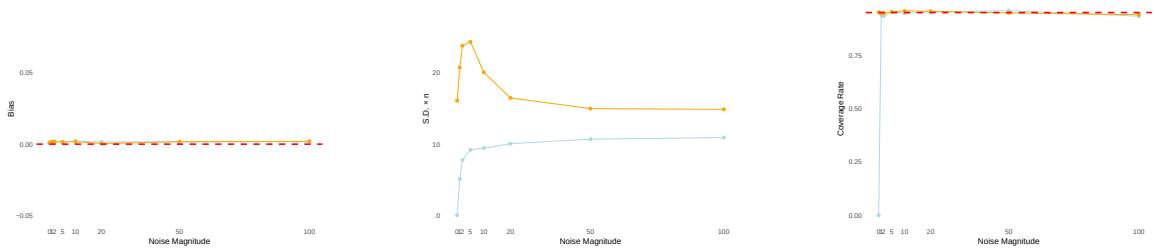
We implement the two methods using the true nuisance functions. In Figure 4, we vary the noise magnitude.



(d) $\xi(W_3)$; True nuisance functions; Varying noise magnitude; Sample size = 1000



(e) $\xi(W_1 \vee W_3)$; True nuisance functions; Varying noise magnitude; Sample size = 1000



(f) $\xi(W_1 \wedge W_3)$; True nuisance functions; Varying noise magnitude; Sample size = 1000

Figure 4: Comparison of bias (left panel), estimated standard deviation times sample size (mid panel), and coverage rate (right panel; significance level $\alpha = 0.05$) with varying noise magnitude. φ_{EIF}-based method in blue, φ_{IF}-based method in gold. True nuisance functions are used. Results aggregated over 1000 trials.

F.2. Additional simulations with estimated nuisances

We compared φ_{EIF}-based estimator and the φ_{IF}-based estimator for total explainability in Section 5, and here we focus on the more challenging problem of interaction explainability. Owing to the increased complexity of nuisance function estimation inherent in interaction effects, both methods exhibit coverage below the target level. Recall that φ_{EIF}-based estimator requires estimating more complicated nuisance components to exploit independence; in this setting, the theoretical gain in

asymptotic variance is offset by the increased variability from nuisance estimation. As a result, the φ_{EIF}-based estimator appears less favorable in terms of both MSE and coverage.

Table 5: Performance comparison of φ_{EIF} and φ_{IF}-based estimators with estimated nuisance function regarding interaction expliainability. Aggregated over 100 trials.

Method	Bias	SD $\times n$	Coverage ($\alpha = 0.05$)
φ_{EIF}	-0.11	2.87	0.79
φ_{IF}	0.07	2.58	0.83

Appendix G. Details of the immigration experiment

The attributes include each immigrant's gender, education level, employment plans, job experience, profession, language skills, country of origin, reasons for applying, and prior trips to the United States. Respondents' characteristics such as age, education, ethnicity, gender, and ethnocentrism are also available in the survey.

One thing should be noted is that there are two restrictions imposed on the possible combinations of immigrant attributes. First, those immigrant profiles who were fleeing persecution were restricted to come from countries where such an application was plausible (e.g., Iraq, Sudan). Second, those immigrants occupying high-skill occupations (e.g., research scientists, doctors) were constrained to have at least two years of college education. In this case, the distribution of immigrants' occupations is dependent on their education, but conditionally independent of the other attributes and the other profiles in the same choice task given their education (Hainmueller et al., 2015). However, as this is an artificial experiment, it is very unclear whether the attributes are *causally* dependent or not. The reason is that the restrictions are put in the design stage, so attributes are simultaneously randomized at the moment the profile is generated. To avoid the problem of unknown directions in a directed acyclic graph, we choose to collapse the two pairs of dependent variables into two variables —— *Origin & Application reason* and *Education & Profession*. Another thing is that although it is a randomized experiment, the authors did not provide any information on the randomization mechanism. So we do the analysis assuming that the propensity scores are unknown.

Table 6: Conditional Randomization Test (Ham et al., 2024)

Interaction	p **value**
Gender & Job Experience	0.9108911
Job Plan & Gender	0.6831683
Job Plan & Language	0.0990099
Job Plan & Job Experience	0.00990099**
Job Plan & Prior trips to U.S.	0.5049505

Table 7: Average Component Interaction Effects (ACIE) (Hainmueller et al., 2015)

Attribute	**Estimate**	**Std. Err**	**z value**	**Pr($>|z|$)**
Gender:Job Experience				
male:1-2 years	-0.0017453	0.023181	-0.07529	0.93998
male:3-5 years	-0.0215082	0.024037	-0.89481	0.37089
male:5+ years	-0.0088778	0.024508	-0.36224	0.71717
Gender:Job Plan				
male:contract with employer	0.0034950	0.023374	0.14953	0.88114
male:interviews with employer	-0.0185813	0.024573	-0.75616	0.44955
male:no plans to look for work	0.0042381	0.022811	0.18579	0.85261
Job Plan:Language				
contract with employer:broken English	0.0477691	0.031921	1.49649	0.134527
interviews with employer:broken English	0.0652677	0.033518	1.94725	0.051505
no plans to look for work:broken English	0.0117892	0.032696	0.36057	0.718423
contract with employer:tried English but unable	0.0100675	0.032462	0.31013	0.756463
interviews with employer:tried English but unable	0.0274033	0.032797	0.83555	0.403405
no plans to look for work:tried English but unable	0.0069235	0.031912	0.21696	0.828243
contract with employer:used interpreter	0.0150713	0.032272	0.46700	0.640497
interviews with employer:used interpreter	0.0499209	0.032747	1.52444	0.127399
no plans to look for work:used interpreter	0.0402530	0.031193	1.29046	0.196892

Number of Obs. = 13960, Number of Respondents = 1396.

Significance codes: 0 *** 0.001 ** 0.01 * 0.05

Table 8: Average Component Interaction Effects (ACIE)-continued

Attribute	**Estimate**	**Std. Err**	**z value**	**Pr($>\|z\|$)**
Job Experience:Job Plan				
1-2 years:contract with employer	0.03528079	0.032295	1.092449	0.274636
3-5 years:contract with employer	0.00069369	0.033193	0.020899	0.983327
5+ years:contract with employer	-0.02503323	0.033785	-0.740960	0.458718
1-2 years:interviews with employer	0.01152705	0.033719	0.341853	0.732462
3-5 years:interviews with employer	-0.02887657	0.034100	-0.846821	0.397095
5+ years:interviews with employer	-0.03515407	0.034809	-1.009902	0.312542
1-2 years:no plans to look for work	-0.00947529	0.031805	-0.297922	0.765762
3-5 years:no plans to look for work	-0.07701162	0.032256	-2.387519	0.016963*
5+ years:no plans to look for work	-0.06210952	0.032943	-1.885391	0.059377
Job Plan:Prior Entry				
contract with employer:once as tourist	-0.0306128	0.036545	-0.83767	0.402218
interviews with employer:once as tourist	0.0639714	0.037340	1.71320	0.086676
no plans to look for work:once as tourist	0.0469911	0.037612	1.24936	0.211533
contract with employer:many times as tourist	-0.0034264	0.036436	-0.09404	0.925077
interviews with employer:many times as tourist	0.0075229	0.038365	0.19609	0.844544
no plans to look for work:many times as tourist	0.0355192	0.037469	0.94797	0.343144
contract with employer:six months with family	0.0191688	0.036835	0.52039	0.602792
interviews with employer:six months with family	0.0508093	0.037914	1.34012	0.180206
no plans to look for work:six months with family	0.0392413	0.036617	1.07166	0.283871
contract with employer:once w/o authorization	0.0091775	0.036384	0.25224	0.800859
interviews with employer:once w/o authorization	0.0302793	0.037539	0.80661	0.419892
no plans to look for work:once w/o authorization	0.0620386	0.035214	1.76175	0.078112

Number of Obs. = 13960, Number of Respondents = 1396.

Significance codes: 0 *** 0.001 ** 0.01 * 0.05

Table 9: Average Component Interaction Effects (ACIE)-continued

Attribute	Estimate	Std. Err	z value	Pr($>\|z\|$)
Job:Job Plan				
waiter:contract with employer	-0.0327390	0.047473	-0.689639	0.490421
child care provider:contract with employer	-0.0900152	0.050735	-1.774211	0.076028
gardener:contract with employer	0.0311797	0.048505	0.642814	0.520345
financial analyst:contract with employer	-0.0484468	0.085065	-0.569525	0.569000
construction worker:contract with employer	-0.0493627	0.049477	-0.997686	0.318432
teacher:contract with employer	-0.0363747	0.050976	-0.713561	0.475499
computer programmer:contract with employer	-0.1469473	0.085449	-1.719704	0.085486
nurse:contract with employer	-0.0603685	0.047341	-1.275183	0.202245
research scientist:contract with employer	-0.1356852	0.080691	-1.681531	0.092660
doctor:contract with employer	-0.1538243	0.077780	-1.977672	0.047966*
waiter:interviews with employer	-0.0450223	0.049338	-0.912537	0.361486
child care provider:interviews with employer	-0.0383598	0.049890	-0.768895	0.441956
gardener:interviews with employer	-0.0413542	0.050545	-0.818161	0.413265
financial analyst:interviews with employer	-0.0863427	0.092088	-0.937614	0.348443
construction worker:interviews with employer	-0.0211739	0.051522	-0.410968	0.681096
teacher:interviews with employer	0.0301964	0.050484	0.598142	0.549746
computer programmer:interviews with employer	-0.1196731	0.088597	-1.350757	0.176773
nurse:interviews with employer	-0.0116562	0.049813	-0.234000	0.814985
research scientist:interviews with employer	-0.0679747	0.084104	-0.808220	0.418964
doctor:interviews with employer	-0.1211354	0.082133	-1.474867	0.140248
waiter:no plans to look for work	0.0028488	0.047582	0.059871	0.952258
child care provider:no plans to look for work	0.0035945	0.048850	0.073583	0.941343
gardener:no plans to look for work	0.0303911	0.048098	0.631859	0.527479
financial analyst:no plans to look for work	0.0443680	0.086945	0.510300	0.609842
construction worker:no plans to look for work	-0.0190767	0.048589	-0.392611	0.694607
teacher:no plans to look for work	0.0125936	0.049917	0.252292	0.800815
computer programmer:no plans to look for work	0.0012924	0.083344	0.015506	0.987628
nurse:no plans to look for work	-0.0131314	0.047923	-0.274010	0.784077
research scientist:no plans to look for work	0.0331237	0.085202	0.388766	0.697449
doctor:no plans to look for work	-0.0040783	0.079754	-0.051136	0.959217

Number of Obs. = 13960, Number of Respondents = 1396.

Significance codes: 0 *** 0.001 ** 0.01 * 0.05