

Bias and Variance Inflation in Hierarchical Multi-Treatment DML: Diagnosis and Correction

Deep Nayak
Amazon
Bangalore, India
deepnyk@amazon.com

Prasanth Naidu Karaka
Amazon
Bangalore, India
kpnaidu@amazon.com

Mohit Kumar
Amazon
Bangalore, India
mohitkrg@amazon.com

Sivaramakrishnan Kaveri
Amazon
Bangalore, India
kavers@amazon.com

Abstract

Many business and clinical decisions depend on the conditional average treatment effect (CATE) of each of several continuous treatments that move together: advertising spend and discount levels driven by the same campaign, or drug doses adjusted in a common protocol. Several estimator families produce CATE at every query point with a confidence interval, including meta-learners, doubly robust learners, Causal Forests and Double Machine Learning (DML); DML in particular gives both in closed form and is widely deployed. We show that, when treatments remain correlated after conditioning on covariates, the bias of the standard DML estimator is amplified along the directions in which two treatments are nearly indistinguishable. The plug-in confidence interval is asymptotically valid as long as the residual correlation does not reach one, but the convergence is so slow under strong correlation that coverage is far from nominal at practical sample sizes. We propose a hierarchy-aware bias correction that combines two CATE estimators of different complexity through a projector built from a user-supplied grouping of treatments by shared cause: the more flexible estimator on the directions where the data is informative and the less expressive estimator on the contrast directions among sibling treatments, where it has lower bias. Across four synthetic data-generating processes, the IHDP public benchmark and a treatment graph with two disjoint sibling triples, the correction lifts coverage on the correlated treatments from 30% for vanilla DML to 94%, reduces mean squared error by a factor of 51 and produces 30% narrower intervals. A second contribution adds a bias-augmented variance term that lifts coverage further to 99% at the cost of wider intervals, for use in extreme-correlation regimes where residual bias persists.

CCS Concepts

• **Computing methodologies** → **Causal reasoning and diagnostics.**



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

CIKM '26, Rome, Italy

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2539-5/2026/11

<https://doi.org/10.1145/3799682.3841043>

Keywords

Causal Inference; Double Machine Learning

ACM Reference Format:

Deep Nayak, Mohit Kumar, Prasanth Naidu Karaka, and Sivaramakrishnan Kaveri. 2026. Bias and Variance Inflation in Hierarchical Multi-Treatment DML: Diagnosis and Correction. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*, November 07–11, 2026, Rome, Italy. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3799682.3841043>

1 Introduction

A widely encountered task in observational causal inference is estimating, for each unit, the causal effect of every member of a small collection of continuous treatments on an outcome, given a vector of pre-treatment covariates. Unit-level effect estimates support targeted decision-making: a marketing analyst wants to know which of several controllable inputs caused a drop in sales, a clinician wants to attribute a patient's outcome to one of several co-prescribed drug doses and a labour economist wants to attribute wage changes to hours, training and tenure. In each setting, the practitioner needs not only an estimate but also a defensible measure of uncertainty.

The methodological landscape for this problem is broad. Meta-learners [28], doubly robust learners [27], Bayesian Additive Regression Trees [21], Causal Forests [45] and the Double Machine Learning (DML) family [10] all target the same conditional average treatment effect (CATE). The methods differ in how they quantify uncertainty at a query point. Meta-learners and the R-learner [33] have no native pointwise confidence intervals and require the bootstrap. BART produces posterior credible intervals from MCMC samples, which are not frequentist coverage statements. Causal Forests [45] produce asymptotic-normal pointwise intervals via an empirical infinitesimal-jackknife variance. DML and its calibrated variant [43] provide influence-function-based pointwise intervals in closed form, which is a property that has driven its adoption in deployed industrial settings such as Microsoft/EconML's TripAdvisor study [41] and Amazon's causal-attribution work [36]. We focus on the DML family in this paper.

A practical difficulty arises because the treatments are typically not freely varying. They often share an external cause that is not itself a treatment of interest, such as a marketing campaign that simultaneously raises advertising spend and discount levels, or a clinical guideline that adjusts the doses of several co-prescribed drugs

together. Conditioning on observed covariates removes confounding bias but not the dependence among the treatments themselves and the residual correlation is rarely uniform across the population: a global remedy that collapses two treatments or applies a pooled orthogonal transform distorts the per-unit effects outside the high-correlation region and erases the per-treatment interpretation the practitioner asked for. In this regime, a classical result applies: the asymptotic variance of the DML estimator on each treatment is governed by a variance inflation factor [17, 30, 37] that diverges when the residualised treatment can be predicted from the others, motivating shrinkage-style remedies in deployed pipelines [12].

Variance inflation is not the whole story. We prove (§4) that under any imperfect nuisance fit, generic in finite samples, the bias of the DML estimator is amplified by the same factor that drives the variance inflation. The plug-in confidence interval is asymptotically valid as long as the residual correlation stays strictly below one, but the standardised bias (bias divided by standard deviation) scales inversely with the square root of the local treatment covariance’s smallest eigenvalue: as siblings become more nearly indistinguishable conditional on covariates, the sample size required to drive standardised bias below the nominal threshold grows without bound. At practical sample sizes the plug-in interval miscovers the truth along precisely the direction where the practitioner most cares about uncertainty: the contrast between two treatments sharing a cause.

The remedy uses the treatment hierarchy the practitioner already knows from the domain (e.g. “these three treatments share a marketing campaign”). Within a sibling group the small-eigenvalue directions of the local covariance are the contrasts among siblings; outside the group they are absent. We fit two CATE estimators that share the residualisation scaffolding but differ only in the expressivity of their nuisance learners. On the contrast subspace, the two disagree by an amount that to leading order equals the bias of the more flexible one; replacing the flexible point estimate with the less-expressive one on that subspace removes most of the bias and recovers coverage with the standard plug-in interval. Treatments outside the hierarchy are left untouched.

Contributions. This paper makes four contributions.

- (1) **Spectral diagnosis (Proposition 1).** The bias of every DML-style two-stage estimator concentrates along the directions where two treatments are nearly indistinguishable conditional on covariates and is amplified by the same inverse-eigenvalue factor that inflates the variance. The standardised bias diverges as treatments become more nearly indistinguishable. This adapts the classical multicollinearity diagnostic from variance to bias and explains why plug-in confidence intervals fail under correlated treatments.
- (2) **Hierarchy-aware bias correction (§5.3).** A point-estimate correction that uses the practitioner-supplied hierarchy to combine two CATE estimators of different complexity: the more flexible estimator on the directions in which the data is informative, the less expressive estimator on the contrast directions among siblings, where it has lower bias. The correction is recomputed at every query point and leaves treatments outside the hierarchy unchanged.

- (3) **Bias-augmented intervals (§5.4).** A variance augmentation that adds the empirical disagreement between the two estimators as an extra variance term, for use when residual bias persists at very high correlation; it widens the interval automatically wherever the spectral diagnosis predicts the plug-in is unreliable and matches it elsewhere.
- (4) **Empirical study (§8).** On heterogeneous data-generating processes spanning light, moderate and high conditional correlation, bias correction alone lifts coverage on the correlated treatments from 30.3% to 94.2%, reduces mean squared error by a factor of 51 relative to Causal Forest and produces intervals 30% narrower than the vanilla DML plug-in. We also evaluate on the IHDP benchmark, a multi-component hierarchical treatment graph and extreme correlation regimes.

2 Related Work

Heterogeneous treatment effect estimation. The partially linear model dates to Robinson [37], with the underlying partialling-out logic from the Frisch-Waugh-Lovell theorem [17, 30]. Double Machine Learning [10] scales this with cross-fit ML nuisance learners; EconML [6] provides LinearDML, KernelDML and Causal Forest implementations. Heterogeneous extensions include the R-learner [33], meta-learners [28], Causal Forest [45] and doubly robust learning [27]. The asymptotics of these estimators have been refined in two recent directions: [16] establishes oracle-rate guarantees for any orthogonal loss and [26] proves that doubly-robust estimation is minimax-optimal in the structure-agnostic regime. For continuous treatments specifically, [22] defines weighted average derivative effects and connects them to projection parameters in partially linear models, the same population estimand we work with. Pointwise confidence intervals for CATE come in several flavours: influence-function-based intervals from DML and its calibrated variant [43]; infinitesimal-jackknife intervals from Causal Forests [45]; conformal intervals on pseudo-outcomes [2, 29, 42]. Practical deployments in industry and medicine document the influence-function flavour at scale [5, 15, 36, 41].

Calibration, shrinkage and structured treatments. Recent DML refinements address failure modes distinct from ours. Calibrated DML [43] isotonicly calibrates the nuisance learners and shows that doubly robust inference still holds when one nuisance converges arbitrarily slowly; shrinkage-based DML [12] adds a global penalty on the second stage; multiply robust estimators [36] are valid under any one of several possible nuisance failures. Other lines impose structure on treatments for a different goal: causal abstraction [8, 38], mediation analysis [24, 44] and hierarchical/causal Shapley attribution [18, 20]. Our user-supplied hierarchy is similar in shape to these but is used to identify the directions along which the local treatment covariance is poorly conditioned, rather than to attribute or abstract.

Bias-aware confidence intervals. The closest precedent for our bias-augmented interval (§5.4) is the bias-aware honest-CI literature in econometrics. [4] construct finite-sample optimal CIs for average treatment effects under unconfoundedness by accounting for the worst-case bias of a non-parametric estimator. [13] extend the same machinery to sensitivity analysis under unmeasured confounding

via partial identification. Both motivate the addition of an explicit bias term to the variance-only Wald interval; ours is the conditional, hierarchy-aware analogue applied to per-unit DML.

Spectral diagnostics under collinearity. Eigen-based diagnostics for ill-conditioned regression are classical. Hoerl-Kennard [23] and Marquardt [31] link OLS instability and ridge bias to the small eigenvalues of the design matrix; Belsley-Kuh-Welsch [7] formalise the condition-index decomposition we adapt in §4.1. The minimum eigenvalue of the concentration matrix is the standard weak-IV diagnostic [11, 40]; the geometry there (small first-stage eigenvalue amplifies the second-stage point estimate) is structurally the same as in our setting. Wedin’s $\sin \theta$ bound [47] controls how a subspace estimated from a noisy Gram matrix shifts relative to its population analogue and analogously bounds the error of the projector we construct in §5.2. The bias of two-stage semi-/non-parametric plug-in estimators in general scales with the nuisance complexity [9, 32]; our contribution localises this to a per-direction amplification on the directions where two treatments are nearly indistinguishable.

Two-estimator combination. Combining estimators of different complexity to reduce risk goes back to James-Stein [25] and frequentist model averaging [19]. The B-Learner of [34] is a recent CATE-specific instance, combining intermediate estimators with quasi-oracle bounds under hidden confounding. Our combination is a directional special case: it shrinks toward the restricted estimator only on the contrast subspace identified by the hierarchy and leaves the rest of treatment space untouched.

Causal discovery for treatment hierarchies. The hierarchy is supplied by the practitioner in this paper, but the same input could in principle be learned from data. The latent-variable structure-learning literature offers several routes: [39] and [3] identify groups of observed variables that share a hidden cause and recent extensions handle nonlinear settings [1] and identifiability under soft interventions [48]. Combining these with our correction is a natural follow-up.

3 Problem Setup and Notation

We observe a sample of triples $\{(X_i, T_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^d$ is a vector of pre-treatment covariates, $T_i \in \mathbb{R}^p$ is a vector of p continuous treatments and $Y_i \in \mathbb{R}$ is the outcome. We adopt the partially linear structural model

$$Y_i = T_i^\top \tau(X_i) + g(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_i, T_i] = 0, \quad (1)$$

in which $\tau : \mathbb{R}^d \rightarrow \mathbb{R}^p$ is the heterogeneous per-unit effect and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is a baseline function of covariates. The estimand of interest is $\tau_j(x)$, the per-unit causal effect of treatment j at a query point x , defined as the partial derivative of the conditional outcome with respect to T_j while holding the other treatments fixed at their observed values [35]; under (1) this equals the j -th coordinate of $\tau(x)$. We assume the standard identifying conditions of unconfoundedness ($Y(t) \perp T | X$), overlap and consistency.

In addition to the data, we assume the practitioner can identify which treatments are likely to share an external cause and supply this information as a partition of the treatment indices.

DEFINITION 1 (TREATMENT HIERARCHY). A hierarchy is a partition of the treatment indices $\{1, \dots, p\}$ into groups $\mathcal{H} = \{G_1, \dots, G_K\}$,

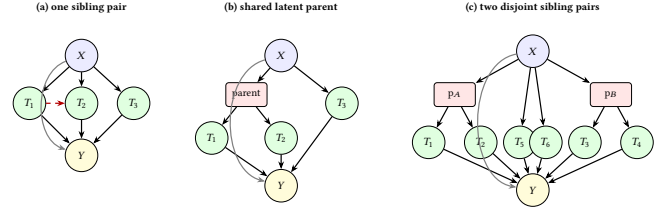


Figure 1: Three treatment graphs and their hierarchies. (a) $\mathcal{H} = \{\{1, 2\}\}$ with T_3 independent. (b) same hierarchy, but the sibling pair shares an unobserved parent rather than a direct correlation. (c) $\mathcal{H} = \{\{1, 2\}, \{3, 4\}\}$ with T_5, T_6 independent.

where each $G_k \subseteq \{1, \dots, p\}$ collects treatments that the practitioner expects to share an external cause. Groups of size one represent treatments that vary independently after conditioning on X .

The hierarchy is supplied as input to our method (§5); it is not estimated from data. Figure 1 illustrates three hierarchies that arise in our experiments.

We work extensively with the residualized variables

$$\tilde{Y}_i = Y_i - \mathbb{E}[Y_i | X_i], \quad \tilde{T}_i = T_i - \mathbb{E}[T_i | X_i]. \quad (2)$$

Estimates of the conditional expectations are denoted $\hat{g}(\cdot)$ and $\hat{m}(\cdot)$ and are referred to as the *nuisance learners*. The local residualized treatment covariance at x is the matrix $\Sigma(x) = \mathbb{E}[\tilde{T}\tilde{T}^\top | X = x] \in \mathbb{R}^{p \times p}$. We write its eigendecomposition $\Sigma(x) = \sum_{k=1}^p \lambda_k(x) v_k(x) v_k(x)^\top$, where the orthonormal eigenvectors $v_k(x) \in \mathbb{R}^p$ are the principal directions of variation of the residualised treatment vector at x and the eigenvalues $\lambda_k(x) \geq 0$ measure the variance of the residualised treatments along $v_k(x)$. We order them $\lambda_1(x) \geq \dots \geq \lambda_p(x) > 0$, so $\lambda_p(x)$ is the smallest eigenvalue and corresponds to the most poorly identified direction. For a hierarchy group G_k we write $\Sigma_{G_k}(x)$ for the principal submatrix of $\Sigma(x)$ on rows and columns indexed by G_k .

4 Spectral Diagnosis: Variance and Bias Inflation

4.1 The Established Variance Inflation

Under the partially linear model (1), the residualized variables satisfy $\tilde{Y}_i = \tilde{T}_i^\top \tau(X_i) + \varepsilon_i$, so the per-unit effect can be recovered by regressing $\tilde{Y}$ on $\tilde{T}$. Locally at x , this regression has covariance matrix $\Sigma(x)$ for the regressors and residual variance $\sigma^2(x) = \text{Var}(\varepsilon | X = x)$ for the response. The asymptotic variance of the per-unit coefficient on treatment j is given by the standard ordinary-least-squares formula in the residualized variables, which by the Frisch-Waugh-Lovell theorem [17, 30] equals

$$\text{Var}(\hat{\tau}_j(x)) = \frac{\sigma^2(x)}{n \text{Var}(\tilde{T}_j | X = x) (1 - R_j^2(x))}, \quad (3)$$

where $R_j^2(x)$ is the population R^2 from regressing $\tilde{T}_j$ on the other components of $\tilde{T}$ given $X = x$. The factor $1/(1 - R_j^2(x))$ is the variance inflation factor (VIF). When $\tilde{T}_j$ is perfectly predictable from the other residualized treatments, $R_j^2(x) \rightarrow 1$ and $\text{Var}(\hat{\tau}_j(x)) \rightarrow \infty$.

Diagonalizing $\Sigma(x) = \sum_k \lambda_k(x) v_k(x) v_k(x)^\top$ separates informative from uninformative directions in treatment space cleanly: each eigenvector $v_k(x)$ is a linear combination of treatments along which the residualized treatments vary by an amount $\sqrt{\lambda_k(x)}$ and the variances along orthogonal eigenvectors are independent. Standard regression algebra applied to the diagonalized residual covariance gives the variance of $\hat{\tau}$ projected onto v_k ,

$$\text{Var}(v_k(x)^\top \hat{\tau}(x)) = \frac{\sigma^2(x)}{n \lambda_k(x)}, \quad (4)$$

which diverges as $\lambda_k(x) \rightarrow 0$. Expression (4) is the eigenbasis form of the same statement as (3): the per-treatment VIF in (3) is the projection onto a single coordinate axis, while (4) reads off the variance along an arbitrary direction in treatment space. The eigenbasis is the only choice of axes that diagonalizes this decomposition, in the sense that an error in one v_k does not influence the variance of another and we will work in it for the rest of the paper.

When variation along some $v_k(x)$ is small the variance of $\hat{\tau}$ on that direction is large, the plug-in interval widens to match and coverage is asymptotically valid: the interval is wide for the right reason. What the variance formula does not say is anything about the *location* of the centre of that interval. The bias of $\hat{\tau}$, which arises from imperfect nuisance fits and is generic in finite samples, is invisible to (3) and (4); the next subsection shows it is amplified by the same eigenvalue.

4.2 Bias Inflation along the Eigenbasis

Let $q(x) = \mathbb{E}[\tilde{T}\tilde{Y} \mid X = x]$ denote the local cross-moment of residualised treatment and residualised outcome (distinct from the treatment-propensity nuisance $\hat{m}(\cdot)$). Multiplying the residualized regression equation $\tilde{Y} = \tilde{T}^\top \tau(x) + \varepsilon$ by $\tilde{T}$ and taking conditional expectation at $X = x$ gives $q(x) = \Sigma(x)\tau(x)$, so the population estimand can be written in closed form as

$$\tau(x) = \Sigma(x)^{-1} q(x). \quad (5)$$

The DML estimator is the sample analog $\hat{\tau}(x) = \hat{\Sigma}(x)^{-1} \hat{q}(x)$, where $\hat{\Sigma}(x)$, $\hat{q}(x)$ are computed from the cross-fit residuals. Because the nuisance learners are imperfect in finite samples, $\hat{\Sigma}(x)$ and $\hat{q}(x)$ each carry a small error relative to their population counterparts; we denote these errors by

$$\Delta_\Sigma(x) = \hat{\Sigma}(x) - \Sigma(x), \quad \Delta_q(x) = \hat{q}(x) - q(x).$$

Under the standard DML conditions on the nuisance learners, both perturbations vanish at rate $n^{-1/2}$ as the sample size grows, so they are small but nonzero in any finite sample. Substituting $\hat{\Sigma} = \Sigma + \Delta_\Sigma$, $\hat{q} = q + \Delta_q$ into the sample formula and Taylor-expanding the inverse to first order in Δ_Σ (the Neumann series for the perturbed inverse) yields

$$\hat{\tau}(x) - \tau(x) \approx \Sigma(x)^{-1} (\Delta_q(x) - \Delta_\Sigma(x) \tau(x)), \quad (6)$$

with a remainder of order $O_P(\|\Sigma^{-1} \Delta_\Sigma\|^2)$, asymptotically negligible whenever $\lambda_{\min}(\Sigma(x)) \gg n^{-1/2}$ (the regime in which the Neumann series converges; see remark below the proposition). Left-multiplying (6) by $v_k(x)^\top$ and using the eigen-identity $v_k(x)^\top \Sigma(x)^{-1} = \lambda_k(x)^{-1} v_k(x)^\top$ separates the deviation into one component per eigendirection of $\Sigma(x)$.

PROPOSITION 1 (BIAS AMPLIFICATION ALONG THE EIGENBASIS). *Under the standard DML regularity conditions [10], the leading-order bias of $\hat{\tau}(x)$ projected onto the eigenvector $v_k(x)$ of $\Sigma(x)$ is*

$$v_k(x)^\top \mathbb{E}[\hat{\tau}(x) - \tau(x)] \approx \frac{1}{\lambda_k(x)} v_k(x)^\top \mathbb{E}[\Delta_q(x) - \Delta_\Sigma(x) \tau(x)]. \quad (7)$$

Combined with the variance scaling (4), the bias-to-standard-deviation ratio along $v_k(x)$ is

$$\frac{|v_k(x)^\top \mathbb{E}[\hat{\tau}(x) - \tau(x)]|}{\sqrt{\text{Var}(v_k(x)^\top \hat{\tau}(x))}} \approx \frac{\sqrt{n} |v_k(x)^\top \mathbb{E}[\Delta_q(x) - \Delta_\Sigma(x) \tau(x)]|}{\sigma(x) \sqrt{\lambda_k(x)}}. \quad (8)$$

Regime of validity. Proposition 1 is a leading-order statement under $\lambda_{\min}(\Sigma(x)) \gg n^{-1/2}$, the regime where the Neumann expansion converges and quadratic remainders are negligible. In the extreme regime ($\rho \rightarrow 1$ with n small, e.g. $\rho = 0.95, n = 500$) this condition can fail and the proposition provides directional intuition rather than a rigorous bound; the empirical results in that regime (§8.4) confirm the qualitative prediction without relying on the exact $1/\sqrt{\lambda_k}$ rate. The same $\Sigma(x)^{-1}$ that amplifies the variance of $v_k(x)^\top \hat{\tau}(x)$ by a factor of $1/\lambda_k(x)$ amplifies its bias by $1/\lambda_k(x)$ as well. Dividing the bias by the standard deviation $\sigma(x)/\sqrt{n\lambda_k(x)}$ leaves a net factor of $1/\sqrt{\lambda_k(x)}$ in the standardized bias.

The standard plug-in confidence interval centred on $\hat{\tau}_j(x)$ uses the variance from (4) and is calibrated under the assumption that the bias is asymptotically negligible. Equation (8) shows when that assumption fails: the standardized bias along $v_k(x)$ scales as $1/\sqrt{\lambda_k(x)}$. When $\lambda_k(x)$ is bounded away from zero the standardized bias remains bounded and the interval covers asymptotically; when $\lambda_k(x) \rightarrow 0$ the standardized bias diverges and the interval miscovers, regardless of how well the nuisances are estimated. The interval is wide for the right reason on these directions, but its centre is in the wrong place.

The eigenvalues of $\Sigma(x)$ measure the conditional variance of the residualized treatments along each direction. When two treatments $\tilde{T}_a, \tilde{T}_b$ are conditionally independent given X the corresponding off-diagonal entry of $\Sigma(x)$ is zero and the eigenvalues remain comparable to the per-treatment variances. When two treatments share an external cause that the covariates do not fully explain, that off-diagonal entry is large and the eigenvalue corresponding to the eigenvector approximately aligned with the contrast between the two treatments is small. To make the geometry tangible, consider the simplest such configuration, $p = 2$ correlated treatments with conditional variances $\text{Var}(\tilde{T}_1 \mid X = x) = \text{Var}(\tilde{T}_2 \mid X = x) = \sigma_T^2(x)$ and conditional correlation $\text{Cor}(\tilde{T}_1, \tilde{T}_2 \mid X = x) = \rho(x)$, so that

$$\Sigma(x) = \sigma_T^2(x) \begin{pmatrix} 1 & \rho(x) \\ \rho(x) & 1 \end{pmatrix}.$$

The eigenvectors of $\Sigma(x)$ are $v_+ = \frac{1}{\sqrt{2}}(1, 1)^\top$ and $v_- = \frac{1}{\sqrt{2}}(1, -1)^\top$ with eigenvalues $\lambda_+(x) = \sigma_T^2(x)(1 + \rho(x))$ and $\lambda_-(x) = \sigma_T^2(x)(1 - \rho(x))$. The direction v_+ is well-identified: the data exhibits ample variation in the joint movement of $\tilde{T}_1$ and $\tilde{T}_2$. The direction v_- , which captures the difference of the two treatments, is poorly identified because both move together. As $\rho(x) \rightarrow 1$, $\lambda_-(x) \rightarrow 0$, the variance amplification (4) along v_- diverges and by Proposition 1 the standardized bias (8) along v_- also diverges as $1/\sqrt{\lambda_-(x)}$.

The same structure generalises to $p \geq 3$ treatments with a small-eigenvalue eigenvector approximately aligned with the contrast direction within each sibling group.

5 From Diagnosis to Correction

Two facts from Proposition 1 drive the construction below. The bias of the per-unit estimator is concentrated on the span of the eigenvectors of $\Sigma(x)$ with the smallest eigenvalues, which we call the *contrast subspace*: outside this subspace the standard plug-in interval is asymptotically valid, while inside it the interval is wide for the right reason on the variance side but centred at the wrong place on the bias side. The size of the bias on the contrast subspace is governed by the projection of the nuisance error onto it, which depends on the choice of nuisance class.

Three structural features of multi-treatment DML let us act on these two facts. The closed-form $\hat{\tau}(x) = \hat{\Sigma}(x)^{-1} \hat{q}(x)$ from (5) is what produces the spectral bias decomposition. The local $\hat{\Sigma}(x)$ is observable from cross-fit residuals, so the contrast subspace can be estimated without knowing τ . And two CATE estimators with different nuisance classes produce perturbations whose projections onto the contrast subspace can differ even when they agree elsewhere, giving us a lever. None of these is available for a generic black-box regressor of Y on (X, T) .

5.1 Two Estimators with Different Nuisance Classes

Suppose two treatments T_a, T_b share an external cause whose mean conditional on X is some $f(X)$ plus an unobserved component that drives their residual dependence. A flexible nuisance learner can absorb $f(X)$ in residualisation: the residuals along the contrast direction are then dominated by the unobserved component, the regression of $\tilde{Y}$ on $\tilde{T}$ there is poorly identified and the flexible-class bias projection onto that direction is large. A deliberately restricted learner (e.g. linear regression when f is non-linear) cannot absorb $f(X)$, leaves the parent-driven structure in the residuals and has a much smaller contrast-subspace bias. This argument relies on an approximate-symmetry condition: the parent’s conditional-mean contribution should be roughly comparable across siblings, so that the restricted learner’s residualisation error has small projection onto the contrast direction. When the condition is grossly violated (e.g. one sibling’s mean is highly non-linear and the other’s is nearly linear), the restricted learner’s error becomes asymmetric and its contrast projection can be non-trivial; we take the symmetry as plausible in practice for siblings driven by a common cause and verify the mechanism empirically. The two classes are not ranked overall: flexible has lower bias on well-identified directions, restricted has lower bias on the contrast subspace. We fit two CATE estimators that share the downstream pipeline but use these classes:

- $\hat{\tau}_b(x)$, *base*: gradient boosting for $\hat{g}_b, \hat{m}_b$; its residuals also yield $\hat{\Sigma}(x)$.
- $\hat{\tau}_a(x)$, *alternative*: linear regression for $\hat{g}_a, \hat{m}_a$, otherwise identical.

Both come with per-unit standard errors $s_b(x), s_a(x)$ from the standard DML formula. By Proposition 1, on well-identified directions both have bias of order $O_P(n^{-1/2})$ under the DML rate conditions

[10] and agree asymptotically; on the contrast subspace, the asymmetry above implies the perturbations of $\hat{\tau}_a$ are smaller than those of $\hat{\tau}_b$, so $\hat{\tau}_b(x) - \hat{\tau}_a(x)$ approximates the contrast-subspace bias of $\hat{\tau}_b(x)$ to leading order, which we exploit in (12) and verify in Figure 4.

5.2 Identifying the Contrast Subspace

The smallest eigenvector of the empirical $\hat{\Sigma}(x)$ is one estimate of the contrast subspace, but in finite samples it is noisy and may pick up low-variance directions unrelated to shared causes. The user-supplied hierarchy $\mathcal{H} = \{G_1, \dots, G_K\}$ narrows the search: for each sibling group G_k of size at least two, the small-eigenvalue direction we care about lies inside the block of $\hat{\Sigma}(x)$ indexed by G_k . We therefore construct a projector that targets exactly these blocks (Figure 2). For each G_k with $|G_k| \geq 2$, we compute the $|G_k| - 1$ smallest-eigenvalue eigenvectors $u^{(k,\ell)}(x) \in \mathbb{R}^{|G_k|}$ of $\hat{\Sigma}(x)[G_k, G_k]$ (the within-group contrast subspace has dimension $|G_k| - 1$: a single contrast for a pair, a two-dimensional subspace orthogonal to the “everyone moves together” direction for a triple). Each $u^{(k,\ell)}$ is embedded into $\mathbb{R}^p$ by placing its entries at the G_k indices and zeros elsewhere, giving $\tilde{u}^{(k,\ell)}(x)$. The *hierarchy-aware projector* is

$$P_c(x) = \sum_{k: |G_k| \geq 2} \sum_{\ell=1}^{|G_k|-1} \tilde{u}^{(k,\ell)}(x) \tilde{u}^{(k,\ell)}(x)^\top. \quad (9)$$

For disjoint sibling groups the embedded eigenvectors have disjoint supports, so $P_c(x)$ is an orthogonal projector onto the union of within-group contrast subspaces. Treatments outside any group of size ≥ 2 contribute zero rows and columns and are untouched by everything that follows. When the parent of a sibling group is itself an observed treatment (e.g. a marketing campaign with its own effect on Y) and is included in G_k , the largest-eigenvalue eigenvector of $\hat{\Sigma}(x)[G_k, G_k]$ aligns with the shared-movement direction (which loads heavily on the parent) and the $|G_k| - 1$ smallest eigenvectors are orthogonal to it: generically child–child contrasts that exclude the parent direction. The parent direction is therefore well-identified (large eigenvalue) and is left untouched by $P_c(x)$, while the contrast directions among the children, where the bias inflation lives, are the ones the projector targets.

5.3 Bias Correction by Combining the Two Estimators

By Proposition 1, the bias of $\hat{\tau}_b(x)$ along eigendirection $v_k(x)$ scales as $\lambda_k(x)^{-1} v_k(x)^\top \mathbb{E}[\Delta_q^{(b)}(x) - \Delta_\Sigma^{(b)}(x) \tau(x)]$ and the same statement holds for $\hat{\tau}_a(x)$ with its own perturbations $\Delta_q^{(a)}, \Delta_\Sigma^{(a)}$. The flexible-class perturbations are small on directions $v_k(x)$ where the data is informative and the small-bias projection is taken on by the flexible nuisance, leaving $\hat{\tau}_b$ with the smaller bias on the well-identified directions. The restricted-class perturbations are small on the contrast subspace because the restricted nuisance leaves the parent-driven structure in the residuals rather than over-fitting it, leaving $\hat{\tau}_a$ with the smaller bias there. The natural combination is to use each estimator on the directions where its bias is smaller. The projector $P_c(x)$ from §5.2 already separates the two subspaces,

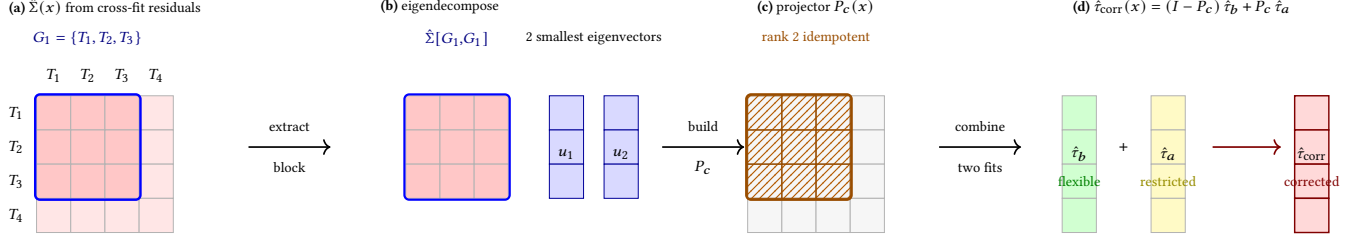


Figure 2: Building and applying the local projector $P_c(x)$ for a hierarchy with one sibling triple, $\mathcal{H} = \{\{T_1, T_2, T_3\}, \{T_4\}\}$. (a) Extract the 3×3 sibling block of $\hat{\Sigma}(x)$. (b) Take its two smallest-eigenvalue eigenvectors. (c) Embed them into $\mathbb{R}^4$ to form P_c (rank-2 idempotent; the orange hatch indicates a derived transformation, not a covariance). (d) Combine the flexible-class and restricted-class CATE estimates through P_c to produce the corrected per-unit estimate.

so the combination is

$$\hat{\tau}_{\text{corr}}(x) = (I - P_c(x)) \hat{\tau}_b(x) + P_c(x) \hat{\tau}_a(x). \quad (10)$$

On the orthogonal complement of $P_c(x)$ this is exactly $\hat{\tau}_b(x)$; on the contrast subspace it is exactly $\hat{\tau}_a(x)$. Treatments outside any sibling group of size ≥ 2 contribute zero to $P_c(x)$ and are therefore unchanged.

For the variance of the corrected estimator, we use the per-subspace plug-in approximation

$$\widehat{\text{Var}}(\hat{\tau}_{\text{corr}}(x)) = (I - P_c(x)) C_b(x) (I - P_c(x))^T + P_c(x) C_a(x) P_c(x)^T, \quad (11)$$

where $C_b(x), C_a(x) \in \mathbb{R}^{p \times p}$ are the per-unit asymptotic covariance matrices of $\hat{\tau}_b(x)$ and $\hat{\tau}_a(x)$. This is exact on each of the two subspaces individually: along any well-identified direction v with $P_c(x)v = 0$, $v^T \widehat{\text{Var}} v = v^T C_b(x) v$ recovers the variance of $\hat{\tau}_b(x)$ projected onto v ; along any contrast direction v with $P_c(x)v = v$, $v^T \widehat{\text{Var}} v = v^T C_a(x) v$ recovers the variance of $\hat{\tau}_a(x)$ projected onto v . The off-diagonal cross-subspace terms $\text{Cov}(\hat{\tau}_b(x), \hat{\tau}_a(x))$ between well-identified and contrast directions are dropped; this is justified to leading order because the small-bias projections of the two estimators are asymptotically uncorrelated (the flexible class controls error on the well-identified subspace and the restricted class controls it on the contrast subspace and the two error sources are governed by independent components of the nuisance perturbations). In our implementation we approximate each $C_\bullet(x)$ by the diagonal matrix of squared per-treatment standard errors returned by the underlying DML estimator (the same diagonals used to form its native plug-in interval). The two-sided $1 - \alpha$ confidence interval for treatment j is then the standard plug-in interval $\hat{\tau}_{\text{corr},j}(x) \pm z_{1-\alpha/2} \sqrt{[\widehat{\text{Var}}(\hat{\tau}_{\text{corr}}(x))]_{jj}}$. Empirically this restores coverage to near nominal across the heterogeneous DGPs, the IHDP benchmark and a multi-component treatment graph (§8). In extreme correlation regimes ($\rho(x) \rightarrow 1$), residual bias on the contrast subspace remains and the plug-in interval can dip a few points below nominal; the next subsection introduces a second correction that closes that residual gap.

5.4 Bias-Augmented Intervals for Extreme Correlation

When the smallest eigenvalue $\lambda_k(x)$ is very small, the bias-correction step cannot remove all of the contrast-subspace bias, because $\hat{\tau}_a(x)$

itself carries a residual contrast-subspace bias (smaller than $\hat{\tau}_b(x)$'s but non-zero). The asymmetry of §5.1 gives a proxy for that residual without further modelling: the projected disagreement between the two estimators,

$$\hat{\beta}(x) = P_c(x)(\hat{\tau}_b(x) - \hat{\tau}_a(x)), \quad (12)$$

which to leading order tracks the contrast-subspace bias of $\hat{\tau}_b(x)$ and serves as a conservative upper bound on the residual contrast-subspace bias of $\hat{\tau}_{\text{corr}}(x)$ (which inherits $\hat{\tau}_a(x)$'s smaller contrast-subspace bias). Treatments outside any sibling group of size ≥ 2 have $\hat{\beta}_j(x) \equiv 0$ by the construction of $P_c(x)$, so they receive no augmentation. The bias-augmented confidence interval is

$$\hat{\tau}_{\text{corr},j}(x) \pm z_{1-\alpha/2} \sqrt{[\widehat{\text{Var}}(\hat{\tau}_{\text{corr}}(x))]_{jj} + \hat{\beta}_j(x)^2}. \quad (13)$$

The construction is conservative: it reduces to the plug-in when $|\hat{\beta}_j(x)|$ is small relative to $\sqrt{[\widehat{\text{Var}}]_{jj}}$ and widens when the bias proxy dominates. The projected estimate $|v_k(x)^T \hat{\beta}(x)|$ scales as $\lambda_k(x)^{-1}$ on the contrast subspace and is zero on the orthogonal complement, so the augmented half-width grows as $\lambda_k(x)^{-1/2}$ exactly where (8) predicts the plug-in is unreliable and matches the plug-in on well-identified directions. The bias estimate uses only training-set quantities. The augmentation is a refinement for the regime where residual contrast-subspace bias is large enough that the plug-in dips below nominal; §5.3 alone suffices everywhere else.

5.5 End-to-End Procedure

Figure 3 shows the full procedure on top of a standard cross-fit DML pipeline [10]. The training sample and the user-supplied hierarchy enter on the left. Cross-fit residualisation runs twice in parallel with the two nuisance classes, producing the base estimator $\hat{\tau}_b$ and the alternative estimator $\hat{\tau}_a$. The flexible-class residuals also yield $\hat{\Sigma}(x)$, whose sibling-restricted eigenvectors form the projector $P_c(x)$. The projector combines the two estimators into the corrected centre $\hat{\tau}_{\text{corr}}$ via (10). The default output uses the plug-in variance from (11); for extreme-correlation regimes, the bias-augmented variance from (13) absorbs residual contrast-subspace bias.

The local treatment covariance $\hat{\Sigma}(x)$ is computed as a kernel-weighted average of the residual outer products $\tilde{T}_i^{(b)}(\tilde{T}_i^{(b)})^T$, with weights peaked at the query covariate x [14, 46]. Any reasonable choice of kernel works; we use a Gaussian.

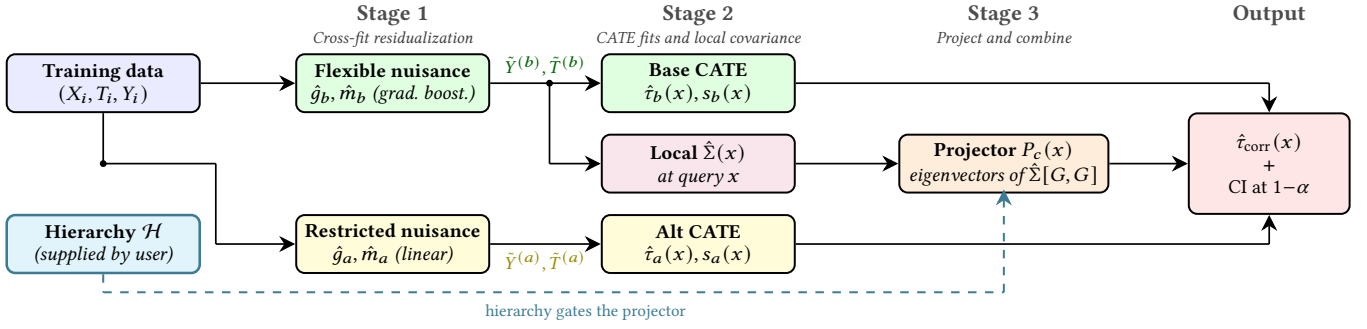


Figure 3: End-to-end procedure. Two cross-fit residualizations with different nuisance classes feed two CATE estimators (stages 1–2). The flexible-class residuals also yield the local covariance $\hat{\Sigma}(x)$, whose sibling-restricted eigenvectors form the projector $P_c(x)$ (stage 3). The user-supplied hierarchy gates which blocks of $\hat{\Sigma}$ enter P_c (dashed cyan). The two CATE estimators are combined through P_c to produce the corrected per-unit estimate with a bias-augmented confidence interval.

Under the standard DML conditions on both nuisance classes [10] and the assumption that the user-supplied hierarchy correctly identifies the contrast subspace (so the empirical projector $P_c(x)$ converges to its population analog), $\hat{\tau}_{\text{corr}}$ is $\sqrt{n}$ -consistent and asymptotically normal. We do not claim coverage robust to misspecification of $\mathcal{H}$, distribution-free finite-sample coverage, or coverage when (1) is wrong.

6 Synthetic Data Generation

To evaluate empirical coverage we need ground-truth per-unit effects, which are unavailable in observational data, so we construct DGPs that fix the partially linear structural model (1) and produce $\tau(x)$ by construction. Each DGP specifies a covariate distribution, a treatment map (with optional latent parents), the heterogeneous effect τ and a baseline g with noise. We vary four knobs across experiments: the treatment graph topology (number and sizes of sibling groups; coupling via direct residual dependence or a shared latent parent); the conditional correlation strength $\rho \in [0, 1]$, which scales the smallest eigenvalue of the relevant block of $\Sigma(x)$ linearly with $1 - \rho$; the smoothness of g (linear or non-linear); and the sample size n . Counterfactuals are obtained by fixing $T = t$ and keeping X, ε : $Y^{(T=t)}(x, \varepsilon) = t^\top \tau(x) + g(x) + \varepsilon$.

We instantiate the simulator at four synthetic settings and one with a realistic public covariate distribution. *LinearPair* and *Non-linearPair* have three treatments where T_1, T_2 are a correlated pair and T_3 is independent, differing in the smoothness of g . *Triple* has four treatments where T_1 is a shared cause with its own effect on Y , T_2, T_3 are its children at coupling strength ρ and T_4 is independent. *TwoTriples* extends this to two disjoint sibling triples plus two independent treatments.

We additionally use the IHDP public benchmark [21] with its original 25 covariates but a synthesised multi-treatment outcome, since IHDP itself is single-treatment. Three treatments are constructed from the standardised covariates: two are independent linear functions of distinct covariates plus Gaussian noise and the third is recoupled to the first at the desired conditional correlation strength using a convex combination weighted by that strength. The heterogeneous effect functions are linear in a single covariate for the correlated pair and constant for the third treatment, the

baseline outcome is a random linear combination of the covariates and the outcome is generated by the partially linear model with standard-normal noise. The hierarchy supplied to our method groups the two correlated treatments into a sibling pair and treats the third as independent. This construction preserves the realistic IHDP covariate distribution while giving us ground truth on the heterogeneous effects and a controllable knob on the conditional correlation strength.

7 Experimental Setup

We compare eight CATE estimators spanning the major two-stage families and one non-residualizing baseline. *S-learner* jointly regresses Y on (X, T) and recovers τ by finite differences. *DML-Constant* uses LinearDML [6, 10] with X entered as a confounder, producing a single homogeneous τ . *LinearDML-Hetero* augments LinearDML with a degree-2 polynomial featurization of X . *KernelDML* uses the Nyström-kernel implementation in EconML [6]. *DML-NN* uses a (64, 32)-hidden-layer MLP as the nuisance learner. *CausalForest* [6, 45] is the standard non-parametric heterogeneous estimator. *Calibrated-CF* pre-calibrates the gradient-boosted nuisances via isotonic regression, in the spirit of [43]. The proposed corrected estimator, *Ours*, is built on top of CausalForestDML: the base estimator $\hat{\tau}_b(x)$ uses gradient-boosted nuisances and is exactly the *CausalForest* baseline above, the alternative estimator $\hat{\tau}_a(x)$ uses linear-regression nuisances inside the same CausalForestDML pipeline and the two are combined through the hierarchy-aware projector of §5.

We report empirical coverage at the nominal level $1 - \alpha = 0.9$ (so $\alpha = 0.1$ is the two-sided miscoverage rate used in (13)), point-estimate mean squared error and mean interval width, computed per treatment on a held-out test set of 20% of the sampled dataset. Each cell of the experimental sweep is repeated for 10 random seeds and we report mean and standard error of the mean (SEM) across seeds. The simulator knobs are varied as follows: $\rho \in \{0.3, 0.7, 0.9\}$ for the heterogeneous synthetic DGPs and $\{0, 0.1, \dots, 0.99\}$ for the correlation-strength ablation, $n \in \{3000, 30000\}$ for the main heterogeneous tables and $n \in \{500, 1000\}$ for the extreme-correlation regime. All forest-based estimators (CausalForest, Calibrated-CF,

DR/R-learner, Ours) use 80 trees with minimum leaf size 20. Gradient-boosted nuisances (HistGradientBoosting) use 200 iterations and depth 5. The local Gaussian kernel bandwidth used to estimate $\hat{\Sigma}(x)$ in our method is fixed at $h = 1$ across all cells.

8 Results

The experiments below are organised around four research questions. **RQ1**: does the predicted spectral signature of the bias, concentration along the small-eigenvalue directions of the local treatment covariance, actually appear in finite samples and across estimators (§8.1)? **RQ2**: does the proposed correction restore near-nominal coverage on the correlated treatments while improving point-estimate accuracy and not over-widening the interval (§8.2)? **RQ3**: of the two corrections we develop, the projector-based bias correction and the bias-augmented variance term, what does each contribute on its own and when is each needed (§8.3)? **RQ4**: how does the method depend on the conditional correlation strength, on multi-component hierarchical graphs, on the correctness of the user-supplied hierarchy and on the extreme limit where treatments become almost indistinguishable (§8.4)?

8.1 Geometric Structure of the Bias

The theoretical analysis of §4 predicts a specific empirical signature: the per-unit bias of any two-stage CATE estimator should concentrate on the small-eigenvalue subspace of $\hat{\Sigma}(x)$. Figure 4 traces this signature on the Triple DGP at $\rho = 0.9$. The four-treatment local covariance has two well-identified directions (v_1, v_2) and a two-dimensional contrast subspace spanned by v_3, v_4 . Panel (a) shows the per-unit bias of our corrected estimator, projected onto (v_1, v_3, v_4) for 200 held-out test units. The cluster collapses near the origin: ours is the only method whose bias on the contrast subspace is comparable to its bias on the well-identified subspace. Panel (b) overlays all eight methods on the same axes (clipped at the 90th percentile so single outliers do not compress the rest of the cloud). Every residualizing baseline either trails along the contrast directions or is offset along them. Two methods sit closer to the origin than the rest: ours (green stars) and DML-NN (orange triangles), whose neural nuisance is the only baseline expressive enough to fit the non-linear parent function on this DGP.

DML-NN’s smaller bias on this figure does not translate into valid intervals: its mean absolute bias along v_3, v_4 is 38.3, roughly 9.7× ours (4.0), large enough that the standardized bias of (8) still exceeds the variance-only confidence bound and its coverage on the correlated treatments is 45.1% in Table 2. In the extreme regime ($\rho = 0.95, n = 500$) DML-NN covers 18.7% while bias correction alone covers 78.3%, because the neural nuisance overfits when data is scarce.

Panel (c) of Figure 4 aggregates per-axis bias magnitudes. CausalForest’s amplification ratio between the smallest and largest eigendirection is 30.9 (per-axis bias 3.7, 4.7, 113.2, 87.2); other residualizing two-stage estimators range from 7.6 (DML-NN) to 82.6 (DML-Constant); the non-residualising S-learner is at 6.9. Our corrected estimator achieves 1.4, the only method with comparable bias across axes. Proposition 1’s signature is thus visible per-unit, per-axis and as a coverage failure quantified next.

Table 1: Coverage, MSE, and width on the IHDP benchmark with synthesised multi-treatment outcomes, $\rho \in \{0.5, 0.7, 0.9\}$, 10 seeds. Mean (SEM); single best per column in bold.

Method	Correlated (T_1, T_2)			Independent (T_3)		
	Cov.	MSE	Width	Cov.	MSE	Width
S-learner	0.892 (0.006)	9.96 (0.25)	11.9 (0.2)	0.968 (0.003)	0.33 (0.02)	2.96 (0.05)
DML-Constant	0.073 (0.004)	0.84 (0.02)	0.19 (0.01)	0.700 (0.084)	0.00 (0.00)	0.11 (0.00)
LinearDML-Hetero	0.816 (0.005)	2.05 (0.15)	3.31 (0.12)	0.830 (0.006)	0.63 (0.03)	2.04 (0.02)
KernelDML	0.127 (0.005)	0.93 (0.03)	0.32 (0.01)	0.569 (0.019)	0.01 (0.00)	0.26 (0.01)
DML-NN	0.312 (0.025)	0.75 (0.10)	0.88 (0.02)	0.897 (0.014)	0.02 (0.00)	0.51 (0.01)
CausalForest	0.757 (0.012)	0.12 (0.01)	0.77 (0.02)	0.891 (0.008)	0.02 (0.00)	0.44 (0.00)
Calibrated-CF	0.329 (0.027)	1.17 (0.14)	1.02 (0.02)	0.912 (0.005)	0.03 (0.00)	0.62 (0.01)
Ours (bias only)	0.911 (0.007)	0.04 (0.00)	0.73 (0.02)	0.891 (0.008)	0.02 (0.00)	0.44 (0.00)
Ours (bias + aug)	0.957 (0.005)	0.04 (0.00)	1.11 (0.06)	0.891 (0.008)	0.02 (0.00)	0.44 (0.00)

Table 2: Coverage, MSE, and width on the heterogeneous DGPs (LinearPair, NonlinearPair, Triple), $\rho \in \{0.3, 0.7, 0.9\}$, $n \in \{3000, 30000\}$, 10 seeds. “Correlated” = in-group treatments, “Independent” = singletons. Mean (SEM); single best per column in bold.

Method	Correlated			Independent		
	Cov.	MSE	Width	Cov.	MSE	Width
S-learner	0.881 (0.006)	66948 (2090)	895.2 (15.0)	0.909 (0.004)	1084 (41)	160.6 (1.9)
DML-Constant	0.051 (0.002)	20359 (1050)	22.9 (0.8)	0.839 (0.027)	37.2 (5.0)	13.5 (0.7)
LinearDML-Hetero	0.432 (0.013)	16514 (1211)	139.2 (5.1)	0.889 (0.003)	1410 (118)	86.3 (4.8)
KernelDML	0.000 (0.000)	356185 (9303)	11.8 (0.5)	0.092 (0.016)	589.4 (33.3)	11.3 (0.6)
DML-NN	0.451 (0.015)	6055 (848)	76.2 (1.8)	0.895 (0.008)	232.7 (45.0)	37.6 (1.2)
CausalForest	0.303 (0.014)	13942 (1081)	79.5 (1.6)	0.937 (0.006)	142.6 (12.0)	43.1 (1.3)
Calibrated-CF	0.168 (0.010)	35593 (2308)	112.4 (2.1)	0.940 (0.004)	273.3 (19.2)	62.3 (1.5)
Ours (bias only)	0.942 (0.003)	271.6 (12.3)	55.3 (1.0)	0.937 (0.005)	143.8 (12.0)	43.1 (1.3)
Ours (bias + aug)	0.993 (0.001)	271.6 (12.3)	287.3 (12.0)	0.937 (0.005)	143.8 (12.0)	43.1 (1.3)

8.2 Coverage on Heterogeneous DGPs

Table 2 aggregates empirical coverage, MSE and interval width over the heterogeneous DGPs (LinearPair, NonlinearPair, Triple), across all knob settings and 10 seeds, separately for the correlated treatments and the independent treatment(s). Cells report mean and standard error across seeds. Every residualizing baseline covers 45.1% or less of the correlated set: DML-NN 45.1%, LinearDML-Hetero 43.2%, CausalForest 30.3%, DR/R-learner 24.7%, Calibrated-CF 16.8%, DML-Constant 5.1%, KernelDML 0%. The non-residualizing S-learner reaches 88.1% but at 4.8× higher MSE than CausalForest and 246× higher MSE than ours. On the independent treatment, all baselines cover between 83.9% (DML-Constant) and 94.0% (Calibrated-CF), with the exception of KernelDML at 9.2% (it is misspecified for this DGP). The bias correction alone covers 94.2% of the correlated set while holding 93.7% on the independent treatment, reduces the MSE on the correlated set by a factor of 51 relative to CausalForest and produces 30% narrower intervals (mean width 55.3 vs 79.5). Adding the bias augmentation lifts coverage further to 99.3% at 5.2× the bias-only width (mean width 287.3 vs 55.3). These aggregates fold together $\rho \in \{0.3, 0.7, 0.9\}$; §8.4 disaggregates by ρ and shows that bias-only coverage stays in the 96–98% range for $\rho \in [0.1, 0.99]$ while vanilla DML decays monotonically.

The same pattern holds on the IHDP public benchmark (Table 1). CausalForest covers 75.7% of the correlated pair; Calibrated-CF drops to 32.9% because its calibration step does not address the geometric source of the bias; the S-learner reaches 89.3% but at MSE 9.96, two orders of magnitude above CausalForest’s 0.12 and our 0.04. Bias correction alone covers 91.1% at 0.95× the width of CausalForest; with augmentation, coverage rises to 95.7% at 1.44×

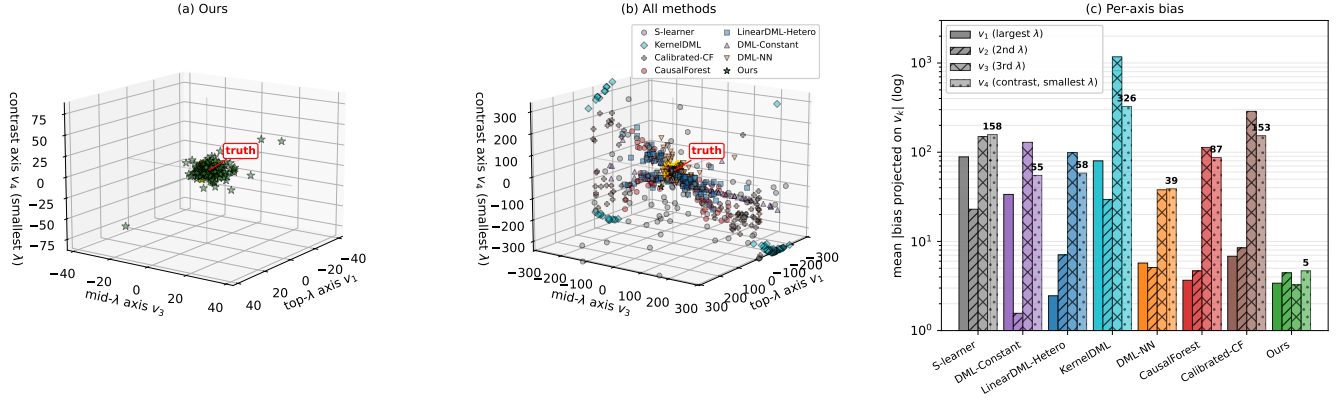


Figure 4: Per-unit bias projected onto the eigenbasis of $\hat{\Sigma}(x)$ on the Triple DGP at $\rho = 0.9$, 200 test units. (a) Ours: cluster collapses around the truth. (b) All methods on the same axes (clipped to 90th percentile); residualizing baselines trail along the contrast direction. (c) Mean per-axis bias magnitude (log).

the width. Coverage on the independent treatment is comparable across methods.

8.3 Decomposing the Contribution of Each Correction

Figure 5 traces what each correction contributes on the LinearPair DGP at $\rho = 0.95$, $n = 500$, where vanilla DML covers below 1% on the correlated pair. The vanilla intervals in (a) are tight, the point estimates are systematically offset from the truth and none of the intervals contain the right answer; each additional data point shrinks the interval at the standard $1/\sqrt{n}$ rate without moving its centre.

Applying bias correction (panel (b)) moves the centre from $\hat{\tau}_b$ to $\hat{\tau}_a$ on the contrast subspace under the plug-in variance (11); point estimates land near the truth, intervals stay tight and coverage rises to 78.3% in this extreme regime where residual contrast-subspace bias still misses $\approx 22\%$ of units. On the easier heterogeneous DGPs the same correction reaches 94.2% with intervals 30% narrower than vanilla DML, because the projector replaces the noise-amplified $1/\lambda_k$ scaling on the contrast subspace with the alt-class native variance. Adding the CI augmentation (c) keeps the same point estimate but widens the interval by the empirical estimator disagreement, lifting extreme-regime coverage to 100% and heterogeneous-DGP coverage to 99.3% at 5.2 $\times$ the bias-only width. Table 3 reports the four-row decomposition (vanilla / bias only / CI augmentation only / both) on the heterogeneous DGPs. Coverage on the independent treatment is essentially unchanged throughout, confirming both corrections touch only the declared sibling subspaces.

8.4 Ablations

We close with four ablations that probe how each correction depends on ρ , the multi-component structure, hierarchy correctness and the limit $\rho \rightarrow 1$.

Figure 6 sweeps $\rho \in \{0, 0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 0.95, 0.99\}$ on LinearPair at $n = 10,000$, 10 seeds. Vanilla CausalForest coverage decays monotonically from 60.3% at $\rho = 0$ to 0% at $\rho \geq 0.95$;

Table 3: Decomposing the contribution of each correction on the heterogeneous DGPs (LinearPair, NonlinearPair, Triple), $\rho \in \{0.3, 0.7, 0.9\}$, $n \in \{3000, 30000\}$, 10 seeds. Vanilla DML is the baseline. Bias correction only (the headline contribution) replaces the point estimate with $\hat{\tau}_{\text{corr}}$ and uses the standard plug-in variance. CI augmentation only keeps the vanilla point estimate but adds $\hat{\beta}^2$ to the variance. Both applies both corrections. Mean (SEM); single best per column in bold.

Setting	Correlated treatments			Independent treatments		
	Cov.	MSE	Width	Cov.	MSE	Width
Vanilla DML	0.303 (.014)	13942 (1081)	79.5 (1.6)	0.937 (.005)	143 (12)	43.1 (1.3)
Bias correction only	0.942 (.003)	272 (12)	55.3 (1.0)	0.937 (.005)	144 (12)	43.1 (1.3)
CI augmentation only	0.989 (.001)	13942 (1081)	296.7 (11.9)	0.937 (.005)	143 (12)	43.1 (1.3)
Both	0.993 (.001)	272 (12)	287.3 (12.0)	0.937 (.005)	144 (12)	43.1 (1.3)

LinearDML-Hetero and Calibrated-CF follow the same trajectory. Bias correction alone reaches $\geq 92\%$ from $\rho = 0.1$ and holds at 96–98% through $\rho = 0.99$; the augmentation lifts it to $\geq 99\%$ across the whole sweep. Vanilla CF MSE grows 240 $\times$ from $\rho = 0$ to $\rho = 0.99$, while bias-corrected MSE stays in the 138–568 range.

The projector composes correctly when the hierarchy contains multiple disjoint sibling groups. Table 4 reports per-treatment coverage on the TwoTriples DGP with hierarchy $\mathcal{H} = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7\}, \{8\}\}$. Every residualizing baseline covers below 90% on at least five of the six correlated treatments, with KernelDML at 0% across the board and Calibrated-CF below 40% on four of six. Bias correction alone covers between 81.7% (T_5) and 98.2% (T_3) on every correlated treatment, with mean 94.0% and matches the baselines on the two independent treatments (94.7%). With augmentation, coverage is between 97.5% and 99.9% on every correlated treatment.

The dependence of the full method on hierarchy correctness, evaluated on the Triple DGP, is what one would hope: with the correct hierarchy $\{\{1, 2, 3\}, \{4\}\}$, the corrected estimator covers 99.9% of the correlated treatments at MSE 105. With an incorrect hierarchy that breaks the parent away from one of its children,

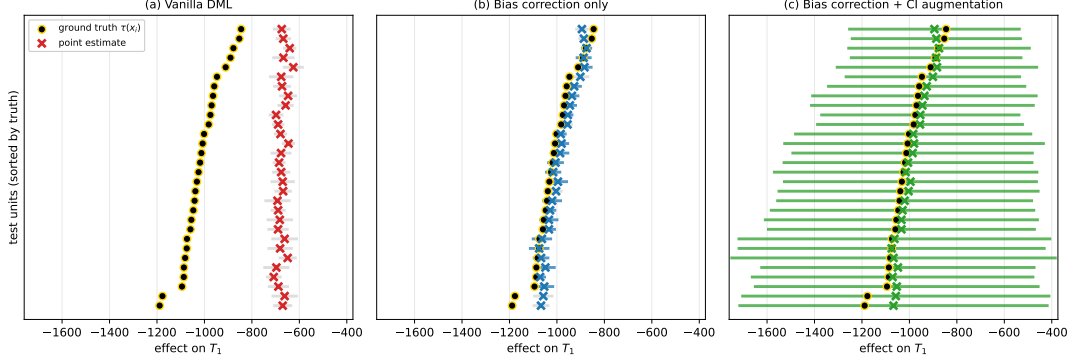


Figure 5: Per-unit 90% intervals on LinearPair at $\rho = 0.95$, $n = 500$, single seed; 30 units stratified to match each method's table-level coverage. Black dots are truth, crosses are point estimates, bars are intervals (gray = miss). Table-level coverage: (a) vanilla DML 0.8%; (b) bias correction alone 78.3% (residual contrast-subspace bias too large for the plug-in interval); (c) with augmentation 100%.

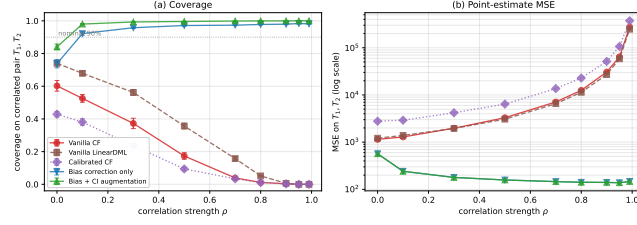


Figure 6: Coverage and point-estimate MSE on the correlated pair T_1, T_2 versus ρ , LinearPair, $n = 10,000$, 10 seeds. Vanilla DML (red, brown, purple) collapses; bias correction (blue) and bias + augmentation (green) hold near-nominal coverage and orders-of-magnitude lower MSE.

coverage drops to 79.8% and MSE rises to 3874. With no hierarchy, coverage collapses to 15.5% and MSE to 19045, matching vanilla DML. Misspecifying the sibling structure entirely is no better than vanilla DML; partially wrong hierarchies degrade gracefully.

The final ablation probes the limit $\rho \rightarrow 1$ on LinearPair at $\rho \in \{0.95, 0.99\}$, $n \in \{500, 1000\}$, where Proposition 1 predicts the standardised bias of every baseline diverges. CausalForest, DML-Constant, KernelDML, Calibrated-CF and DR/R-learner all collapse to below 1% coverage at $\rho = 0.95$ and to exactly 0% at $\rho = 0.99$; LinearDML-Hetero retains 24.8–57.7% at $\rho = 0.95$ and 6.3–33.0% at $\rho = 0.99$, but never reaches nominal. Bias correction alone covers 78.3–89.4% across the four cells; the residual contrast-subspace bias is too large for the plug-in interval to fully absorb. Adding the CI augmentation lifts coverage to between 99.97% and 100.00%, at the cost of 10–20 $\times$ wider intervals than the bias-only baseline. The DGP-implied variance inflation factor $1/(1 - \rho^2)$ ranges from ≈ 10 at $\rho = 0.95$ to ≈ 50 at $\rho = 0.99$ and the bias inflation tracks it as Proposition 1 predicts.

Table 4: Per-treatment coverage on the TwoTriples DGP (two disjoint sibling triples $\{T_1 - T_3\}, \{T_4 - T_6\}$; T_7, T_8 independent), $\rho \in \{0.3, 0.7, 0.9\}$, $n \in \{3000, 30000\}$, 10 seeds. Mean (SEM); single best per column in bold; leading zeros omitted.

Method	Group A			Group B			Indep.	
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
S-learner	.625(.03)	.924(.01)	.748(.02)	.532(.04)	.592(.04)	.838(.02)	.840(.01)	.873(.01)
DML-Constant	.084(.01)	.077(.01)	.039(.01)	.100(.01)	.077(.01)	.385(.03)	.767(.06)	.767(.06)
LinearDML-Hetero	.767(.01)	.728(.02)	.465(.03)	.787(.01)	.686(.02)	.879(.01)	.884(.01)	.868(.01)
KernelDML	.000(.00)	.000(.00)	.000(.00)	.000(.00)	.000(.00)	.000(.00)	.033(.02)	.000(.00)
DML-NN	.839(.02)	.501(.03)	.638(.03)	.792(.01)	.684(.01)	.848(.01)	.850(.02)	.665(.03)
CausalForest	.503(.04)	.529(.04)	.195(.03)	.539(.04)	.388(.03)	.896(.01)	.947(.01)	.947(.01)
Calibrated-CF	.395(.03)	.292(.03)	.078(.02)	.386(.03)	.242(.03)	.882(.01)	.938(.01)	.917(.01)
Ours (bias only)	.978(.01)	.970(.01)	.982(.00)	.929(.00)	.817(.01)	.960(.00)	.947(.01)	.947(.01)
Ours (bias + aug)	.998(.00)	.995(.00)	.999(.00)	.985(.00)	.975(.00)	.988(.00)	.947(.01)	.947(.01)

9 Discussion and Conclusion

The empirical failure of multi-treatment DML confidence intervals under correlated treatments has a single clean source. Beyond the established variance-inflation result for OLS, the bias of the per-unit estimator is also amplified along the small-eigenvalue directions of the local treatment covariance and the bias-to-standard-deviation ratio diverges along the contrast direction of any pair of treatments that share an external cause. As the sample size grows, the plug-in interval narrows at the usual $1/\sqrt{n}$ rate on the variance term while the bias term along the contrast direction does not vanish, so the practitioner ends up confidently centred on the wrong value. The proposed hierarchy-aware bias correction repairs the point estimate by combining two CATE estimators of different complexity through a projector built from a user-supplied grouping of treatments. A CI augmentation that absorbs the empirical estimator gap as a second variance term lifts coverage to 99.3% at the cost of wider intervals; this augmentation is useful primarily in extreme-correlation regimes where residual contrast-subspace bias persists. Both corrections leave treatments outside the supplied grouping unchanged. Several extensions are natural and we leave them to future work: data-driven estimation of the hierarchy, deeper hierarchies that include mediation and adaptive nuisance-class selection.

GenAI Usage Disclosure

Popular SOTA generative AI tools were used for minor rephrasing and editing of the text to improve the language and structure of the publication. The authors take full responsibility and ownership of the method proposed and its results.

References

- [1] Raj Agrawal, Chandler Squires, Neha Prasad, and Caroline Uhler. 2023. The DeCAMFounder: Nonlinear Causal Discovery in the Presence of Hidden Variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 85, 5 (2023), 1639–1667. doi:10.1093/jrssb/bkac071
- [2] Ahmed Alaa, Zaid Ahmad, and Mark van der Laan. 2023. Conformal Meta-learners for Predictive Inference of Individual Treatment Effects. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.
- [3] Animashree Anandkumar, Daniel Hsu, Adel Javanmard, and Sham Kakade. 2013. Learning Linear Bayesian Networks with Latent Variables. In *Proceedings of the 30th International Conference on Machine Learning (ICML) (PMLR, Vol. 28)*. 249–257.
- [4] Timothy B. Armstrong and Michal Kolesár. 2021. Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness. *Econometrica* 89, 3 (2021), 1141–1177. doi:10.3982/ecta16907
- [5] Susan Athey, Niall Keleher, and Jann Spiess. 2023. Machine Learning Who to Nudge: Causal vs Predictive Targeting in a Field Experiment on Student Financial Aid Renewal. *arXiv preprint arXiv:2310.08672* (2023).
- [6] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2019. EconML: A Python package for ML-based heterogeneous treatment effects estimation. <https://github.com/py-why/EconML>. Version 0.16.0.
- [7] David A. Belsley, Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley.
- [8] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. 2014. Visual causal feature learning. *arXiv preprint arXiv:1412.2309* (2014).
- [9] Xiaohong Chen. 2007. Large Sample Sieve Estimation of Semi-Nonparametric Models. In *Handbook of Econometrics*, James J. Heckman and Edward E. Leamer (Eds.), Vol. 6B. Elsevier, Chapter 76, 5549–5632.
- [10] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (01 2018), C1–C68. arXiv:https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf doi:10.1111/ectj.12097
- [11] John G. Cragg and Stephen G. Donald. 1993. Testing Identifiability and Specification in Instrumental Variable Models. *Econometric Theory* 9, 2 (1993), 222–240.
- [12] Enes Diler and Colin Gray. 2025. Shrinkage-Based Regressions with Many Related Treatments. *arXiv preprint arXiv:2507.01202* (2025).
- [13] Jacob Dorn, Kevin Guo, and Nathan Kallus. 2024. Doubly-Valid/Doubly-Sharp Sensitivity Analysis for Causal Inference with Unmeasured Confounding. *J. Amer. Statist. Assoc.* (2024). doi:10.1080/01621459.2024.2335588
- [14] Jianqing Fan and Irène Gijbels. 1996. *Local Polynomial Modelling and Its Applications*. Chapman and Hall/CRC.
- [15] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. 2024. Causal Machine Learning for Predicting Treatment Outcomes. *Nature Medicine* 30 (2024), 958–968. doi:10.1038/s41591-024-02902-1
- [16] Dylan J. Foster and Vasilis Syrgkanis. 2023. Orthogonal Statistical Learning. *The Annals of Statistics* 51, 3 (2023), 879–908. doi:10.1214/23-aos2258
- [17] Ragnar Frisch and Frederick V Waugh. 1933. Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society* (1933), 387–401.
- [18] Christopher Frye, Colin Rowat, and Ilya Feige. 2020. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in neural information processing systems* 33 (2020), 1229–1239.
- [19] Bruce E. Hansen. 2007. Least Squares Model Averaging. *Econometrica* 75, 4 (2007), 1175–1189.
- [20] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems* 33 (2020), 4778–4789.
- [21] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.
- [22] Oliver Hines, Karla Diaz-Ordaz, and Stijn Vansteelandt. 2026. Parameterising the Effect of a Continuous Treatment Using Average Derivative Effects. *Biometrika* (2026). doi:10.1093/biomet/asag012
- [23] Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12, 1 (1970), 55–67.
- [24] Kosuke Imai, Luke Keele, and Teppei Yamamoto. 2010. Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* 25, 1 (2010), 51–71. doi:10.1214/10-STS321
- [25] W. James and Charles Stein. 1961. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*. University of California Press, 361–379.
- [26] Jikai Jin and Vasilis Syrgkanis. 2024. Structure-Agnostic Optimality of Doubly Robust Learning for Treatment Effect Estimation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
- [27] Edward H Kennedy. 2024. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of statistical methods for precision medicine* (2024), 207–236.
- [28] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.
- [29] Lihua Lei and Emmanuel J. Candès. 2021. Conformal Inference of Counterfactuals and Individual Treatment Effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83, 5 (2021), 911–938.
- [30] Michael C Lovell. 1963. Seasonal adjustment of economic time series and multiple regression analysis. *J. Amer. Statist. Assoc.* 58, 304 (1963), 993–1010.
- [31] Donald W. Marquardt. 1970. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics* 12, 3 (1970), 591–612.
- [32] Whitney K. Newey. 1997. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79, 1 (1997), 147–168.
- [33] Xinkun Nie and Stefan Wager. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108, 2 (2021), 299–319.
- [34] Miruna Oprescu, Jacob Dorn, Marah Ghoumnia, Andrew Jesson, Nathan Kallus, and Uri Shalit. 2023. B-Learner: Quasi-Oracle Bounds on Heterogeneous Causal Effects Under Hidden Confounding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.
- [35] Judea Pearl. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann, 411–420.
- [36] Victor Quintas-Martinez, Mohammad Taha Bahadori, Eduardo Santiago, Jeff Mu, Dominik Janzing, and David Heckerman. 2024. Multiply-robust causal change attribution. *arXiv preprint arXiv:2404.08839* (2024).
- [37] Peter M Robinson. 1988. Root-N-consistent semiparametric regression. *Econometrica: journal of the Econometric Society* (1988), 931–954.
- [38] Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. 2017. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819* (2017).
- [39] Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. 2006. Learning the Structure of Linear Latent Variable Models. *Journal of Machine Learning Research* 7 (2006), 191–246.
- [40] James H. Stock and Motohiro Yogo. 2002. *Testing for Weak Instruments in Linear IV Regression*. NBER Technical Working Paper 284. National Bureau of Economic Research.
- [41] Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. 2019. Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [42] Lars van der Laan and Ahmed M. Alaa. 2024. Self-Calibrating Conformal Prediction. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*.
- [43] Lars van der Laan, Alex Luedtke, and Marco Carone. 2024. Automatic doubly robust inference for linear functionals via calibrated debiased machine learning. *arXiv preprint arXiv:2411.02771* (2024).
- [44] Tyler VanderWeele and Stijn Vansteelandt. 2014. Mediation analysis with multiple mediators. *Epidemiologic methods* 2, 1 (2014), 95–115.
- [45] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.
- [46] M. P. Wand and M. C. Jones. 1995. *Kernel Smoothing*. Chapman and Hall/CRC.
- [47] Per-Ake Wedin. 1972. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics* 12, 1 (1972), 99–111.
- [48] Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. 2023. Identifiability Guarantees for Causal Disentanglement from Soft Interventions. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.