Proceedings of Machine Learning Research 323:1–34, 2026 5th Conference on Causal Learning and Reasoning

Retrospective Counterfactual Prediction by Conditioning on the Factual Outcome: A Cross-World Approach

Juraj Bodik JURAJ.BODIK@UNIL.CH
HEC Lausanne, University of Lausanne, Switzerland

Editor: Bijan Mazaheri and Niels Richard Hansen

Abstract

Retrospective causal questions ask what would have happened to an observed individual had they received a different treatment. We study the problem of estimating $\mu(x, y) = \mathbb{E}[Y(1) \mid X = x, Y(0) = y]$, the expected counterfactual outcome for an individual with covariates x and observed outcome y, and constructing valid prediction intervals under the Neyman-Rubin superpopulation model. This quantity is generally not identified without additional assumptions. To link the observed and unobserved potential outcomes, we work with a cross-world correlation $\rho(x) = \mathrm{cor}(Y(1), Y(0) \mid X = x)$; plausible bounds on $\rho(x)$ enable a principled approach to this otherwise unidentified problem. We introduce retrospective counterfactual estimators $\hat{\mu}_\rho(x, y)$ and prediction intervals $C_\rho(x, y)$ that asymptotically satisfy $P[Y(1) \in C_\rho(x, y) \mid X = x, Y(0) = y] \geq 1 - \alpha$ under standard causal assumptions. Many common baselines implicitly correspond to endpoint choices $\rho = 0$ or $\rho = 1$ (ignoring the factual outcome or treating the counterfactual as a shifted factual outcome). Interpolating between these cases through cross-world dependence yields substantial gains in both theory and practice.

Keywords: Causal inference; Counterfactual prediction; Cross-world assumptions; Prediction intervals

1. Introduction

At its core, causal inference pursues two goals: assessing what would have happened to an individual under an alternative treatment, and predicting how a new individual will benefit from treatment (Rubin, 2005). For answering the second goal, the literature focuses on average treatment effects (ATE) or conditional average treatment effects (CATE). However, estimating retrospective counterfactuals (first goal) is often more challenging, as it requires more untestable assumptions, connected to the Pearl's third ladder of causation (Pearl and Mackenzie, 2019). Estimates of counterfactuals are critical in many fields: in medicine, they enable evaluating how a patient might have responded to a different treatment (Imbens and Rubin, 2015); in criminal law, they underpin the "but-for" test of causation, which assesses liability based on whether harm would have occurred absent the defendant's action (Wright, 1985).

Consider a medical scenario in which a patient, Alex, arrives at a hospital with covariates $X = x$ (e.g., age, weight, and other characteristics), does not receive the treatment ($T = 0$), and experiences an outcome $Y(0) \in \mathbb{R}$. Estimating his retrospective counterfactual outcome $Y(1)$ is central to causal reasoning, when accountability demands a clear assessment of whether the treatment decision aligned with what would likely have led to a better outcome. In high-stakes settings such as healthcare, it is equally important to quantify the uncertainty in individual treatment effects (ITEs); that is, to construct a set $C \subseteq \mathbb{R}$ that contains $Y(1)$ with high probability.

© 2026 J. Bodik.

Existing methods primarily focus on estimating the CATE, defined as $\tau(x) = \mu_1(x) - \mu_0(x)$, where $\mu_t(x) = \mathbb{E}[Y(t) \mid X = x]$ for $t = 0, 1$ can be estimated via e.g. random forest (Wager and Athey, 2018). The missing counterfactual is often imputed either by $\hat{Y}(1) = Y(0) + \hat{\tau}(X)$, by $\hat{Y}(1) = \hat{\mu}_1(X)$, or through a matching-based approach. Some notable exceptions are presented in Section 2 and Appendix A.1.

Most existing approaches condition only on covariates X, overlooking the observed (factual) outcome $Y(0)$, which often contains valuable individual-specific information. For instance, if Alex left the hospital healthy after not receiving treatment ($Y(0)$ is large), it is highly likely that he would also be healthy under the counterfactual scenario in which he received treatment ($Y(1)$ is large). Incorporating the factual outcome alongside the covariates can therefore refine individual-level predictions and improve the accuracy of estimated counterfactuals.

In this work, we propose leveraging covariates *and* the factual outcome to enhance counterfactual prediction. Specifically, instead of estimating $\mathbb{E}[Y(1) \mid X = x]$, we aim to construct point estimates

$$\hat{\mu}_\rho(x, y) \quad \text{for} \quad \mathbb{E}\left[Y(1) \mid X = x, Y(0) = y\right], \tag{1}$$

and prediction intervals $C_\rho(x, y)$ for the counterfactuals satisfying (at least asymptotically):

$$P\left(Y(1) \in C_\rho(x, y) \mid X = x, Y(0) = y\right) \geq 1 - \alpha, \tag{2}$$

for $\alpha \in (0, 1)$ (typically $\alpha = 0.1$). By leveraging the factual outcome in conditioning, we obtain more informative and reliable individual-level counterfactual predictions than methods that rely solely on covariates.

However, conditioning on the factual outcome introduces a fundamental challenge: since both potential outcomes are never observed for the same individual, the joint distribution of $\big(Y(0), Y(1)\big)$ is unidentifiable without further assumptions. To address this, we adopt a class of assumptions known as cross-world assumptions.

Definition 1 (Bodik et al. (2025)) *In the Neyman–Rubin super-population model with i.i.d. units, the dependence structure (conditional correlation) between the potential outcomes $Y(1), Y(0)$, conditioned on the observed covariates X, is defined as:*

$$\rho(x) = \operatorname{cor}\big(Y(1), Y(0) \mid X = x\big).$$

We refer to assumptions about the dependence between $Y(1)$ and $Y(0)$ as cross-world assumptions.

Related notions have appeared in prior literature (see Section 2), often represented via an additive structural equation model:

$$Y(0) = \mu_0(X) + \varepsilon_0, \quad Y(1) = \mu_1(X) + \varepsilon_1, \quad \text{where } \operatorname{cor}(\varepsilon_1, \varepsilon_0) = \rho(X).$$

Although ρ is not identifiable from data, postulating plausible values or bounds from domain experts is often both feasible and well-aligned with how humans make judgments. Observing one potential outcome often conveys information about the other, beyond what is captured by covariates. In some settings, ρ can even be estimated or bounded more sharply when auxiliary variables are available, as these may reveal additional structure shared across the two potential outcomes.

While specifying ρ may be difficult in some applications, we show (Lemma 2) that point-identifying $\mathbb{E}[Y(1) \mid X = x, Y(0) = y]$ requires fully specifying the copula linking $Y(1)$ and $Y(0)$. For clarity, we adopt a Gaussian copula parametrized by ρ. This choice could be replaced by any fixed copula family, and the same reasoning applies, although doing so would require more detailed practitioner input regarding the exact form of cross-world dependence. Identifying a credible dependence structure is therefore the central and practically consequential challenge in retrospective counterfactual inference.

Our main contributions:

- We introduce cross-world dependence as a tool for retrospective counterfactual prediction.
- We provide the first systematic treatment of the estimands (1) and (2).
- Assuming a fully specified cross-world dependence, we propose Retrospective Counterfactual Prediction framework ($RCP(\rho)$): a consistent estimator of (1) and prediction intervals that asymptotically satisfy (2) under standard causal assumptions.
- We conduct an extensive empirical study showing that conditioning on the factual outcome can substantially improve accuracy and reliability relative to existing approaches.

For clarity, we mainly focus on the case where $Y(0)$ is observed ($T = 0$) and $Y(1)$ is the counterfactual; the opposite case is analogous.

2. Preliminaries, related work and cross-world assumption

We adopt the Neyman-Rubin potential outcomes framework (Rubin, 2005), where each unit i has potential outcomes $Y_i(1)$ and $Y_i(0)$, covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, and treatment assignment $T_i \in \{0, 1\}$. The observed outcome is $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) \in \mathcal{Y} \subseteq \mathbb{R}$, while the $ITE_i = Y_i(1) - Y_i(0)$ remains unobservable. We assume $(Y_i(1), Y_i(0), T_i, X_i) \overset{\text{i.i.d.}}{\sim} (Y(1), Y(0), T, X)$, for a generic random vector $(Y(1), Y(0), T, X)$ with distribution P. The conditional average treatment effect (CATE) is defined as $\tau(x) = \mu_1(x) - \mu_0(x)$ with $\mu_t(x) = \mathbb{E}[Y(t) \mid X = x]$.

We impose **strong ignorability** and **overlap**, meaning $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$ and $0 < \pi(x) < 1$ for all $x \in \mathcal{X}$, where $\pi(x) = \mathbb{P}(T = 1 \mid X = x)$ denotes the propensity score. This ensures that treatment is as-if randomly assigned given covariates and that both treatments are feasible. Under these assumptions, CATE is identified via $\mu_t(x) = \mathbb{E}[Y \mid T = t, X = x]$; see Wager (2024).

We note that some authors use the terms "ITE" and "CATE" interchangeably, which can lead to confusion. Here, ITE is a latent, unit-specific quantity, while the CATE is an unknown function, defined as the conditional expectation of the ITE given covariates.

2.1. Related work: cross-world assumption

In the potential outcomes framework, the joint distribution of $\big(Y(1), Y(0)\big) \mid X$ is unidentifiable because only one potential outcome is observed per unit. While CATE can be identified without assumptions on this joint law, quantities such as variance, quantiles, or prediction intervals of ITE generally depend on the cross-world correlation $\rho(X) = \operatorname{cor}\big(Y(1), Y(0) \mid X\big)$ (Rubin, 1990; Ding et al., 2019). This has been studied in joint distribution modeling (Heckman et al., 1997;

Fan and Park, 2010), quantile treatment effect estimation (Firpo, 2007), nonparametric bounds using copulas (Zhang and Richardson, 2025a,b; Nelsen et al., 2001) and optimal transport (Dance and Bloem-Reddy, 2025). Andrews and Didelez (2021) highlight the implausibility of cross-world independence assumptions in mediation analysis. We complement these by parameterizing cross-world dependence via $\rho(x)$.

Bodik et al. (2025) and Cai et al. (2022) argue that in many real-world applications ρ is non-negative and often substantially positive due to shared latent factors affecting both potential outcomes. This is also supported empirically: across over 30 real-world datasets, the cross-world correlation was always positive and was typically above 0.5. Formally, consider a model where $Y(1) = \mu_1(X) + H + \tilde{\varepsilon}_1$ and $Y(0) = \mu_0(X) + H + \tilde{\varepsilon}_0$, where $X \in \mathbb{R}^d$ are observed covariates, $H \perp\!\!\!\perp (X, T)$ is an unobserved factor influencing both potential outcomes, and $\tilde{\varepsilon}_0 \perp\!\!\!\perp \tilde{\varepsilon}_1$ are idiosyncratic noise terms. Conditioning on X, it is easy to derive that $\rho(X) = \mathrm{cor}(Y(1), Y(0) \mid X) = \frac{\mathrm{var}(H)}{\sqrt{\mathrm{var}(\tilde{\varepsilon}_0)\,\mathrm{var}(\tilde{\varepsilon}_1)}} \geq 0$. The shared influence of H induces strictly positive correlation between $Y(1)$ and $Y(0)$, even after adjusting for X. Moreover, if the treatment has no or very small effect, then $Y(1) \approx Y(0)$ and hence $\rho \approx 1$.

Estimating ρ using auxiliary data. Bodik et al. (2025) showed that $\rho(x)$ can be estimated from data when richer auxiliary variables Z are available for a subset of units. For example, in a clinical setting, a subset of patients may have rich auxiliary measurements such as detailed medical history, lab panels, or genetic markers, whereas for most patients only a small set of routine covariates was measured. The presence of Z allows us to construct informative, data-driven *bounds* on $\rho(x)$. The authors proposed a consistent estimator for the bounds, and showed e.g. on the Twins dataset (we return to this dataset in Section 4), that the cross-world correlation lies in $\rho(x) \in (0.5, 0.7)$.

How to choose ρ without auxiliary data? If auxiliary variables are not available, $\rho(x)$ should be guided by domain knowledge and viewed as a *sensitivity parameter*. In this case, the goal is not to select a single "correct" value, but to examine how the conclusions vary over a range of plausible values of $\rho(x)$. Practitioners should therefore ask: how much residual dependence between $Y(0)$ and $Y(1)$, beyond what is explained by X, is scientifically credible in the application at hand? Equivalently, what values are plausible for $\frac{\mathrm{var(shared\ latent\ effects)}}{\mathrm{var(idiosyncratic\ noise)}}$? A useful strategy is to report results over a grid of values, or under lower and upper bounds, and assess whether the main qualitative conclusions remain stable. In this sense, $\rho(x)$ plays the same role as a standard sensitivity parameter: it encodes untestable cross-world dependence, and robustness is demonstrated when the substantive findings do not materially change across a reasonable range of specifications.

In many complex systems, it is reasonable to expect a substantial contribution from shared latent components, suggesting that $\rho(x)$ is often positive and sometimes fairly large. At the same time, $\rho(x)$ is rarely close to 1, since treatment effects generally exhibit heterogeneity even among individuals with the same observed covariates X. This is not a universal rule, but rather a practical guideline for constructing a plausible sensitivity range, grounded in the idea that latent common causes in many real-world systems influence both $Y(0)$ and $Y(1)$.

As an example, consider a clinical trial testing a new drug for reducing blood pressure, where treatment is randomly assigned. Let $Y_i(1)$ denote patient i's blood pressure after receiving the drug and $Y_i(0)$ after receiving a placebo. Even though baseline covariates such as age, weight, and pre-existing conditions are observed, unmeasured factors such as genetic predisposition can strongly influence both potential outcomes. A patient with naturally resilient cardiovascular health will likely exhibit relatively low blood pressure regardless of treatment, whereas a patient with

severe underlying issues will tend to have higher readings in both scenarios. These persistent latent traits induce positive dependence between $Y_i(1)$ and $Y_i(0)$ even after adjusting for a few observed covariates. Thus, in this setting, a sensitivity analysis centered on positive and moderately large values of $\rho(x)$ would often be scientifically more credible than one centered near zero.

2.2. Related work: retrospective counterfactuals for in-study units

Inferring individual counterfactual outcomes is fundamentally a missing data problem (Ding and Li, 2018). Many methods for counterfactual prediction use CATE-adjusted imputation $\hat{Y}_i(1) = Y_i(0) + \hat{\tau}(X_i)$, where $\hat{\tau}$ is estimated using doubly-robust estimator, random forests or S/T-learner (Wager, 2024; Künzel et al., 2019; Athey et al., 2019). Other approaches directly model the treated outcome as $\hat{Y}_i(1) = \hat{\mu}_1(X_i)$, thereby ignoring information contained in the observed outcome $Y_i(0)$ (possibly using control group only for the propensity estimation, Lei and Candès (2021)).

Classic counterfactual prediction methods target $\mathbb{E}[Y(T) \mid X]$ without conditioning on $Y(0)$. For instance, Kim et al. (2022) propose a doubly robust estimator for counterfactual classification that directly models the treated outcome distribution, and McClean et al. (2024) develop nonparametric estimators for conditional incremental effects (based on stochastic propensity interventions) with a similar goal of directly estimating $\mathbb{E}[Y(1) \mid X]$. More recently, Kim (2025) introduces a semiparametric counterfactual regression framework that likewise estimates $\mathbb{E}[Y(1) \mid X]$ using flexible machine learning and Ma et al. (2024) using diffusion models. Like many modern approaches, these methods aim to improve recovery of treatment effects from covariates X, but they do not explicitly exploit the realized factual outcome $Y_i(0)$ when predicting the missing counterfactual $Y_i(1)$. These approaches forego individual-level imputation using $Y(0)$, instead relying on robust modeling of the treated outcome. Most existing methods focus on minimizing the Precision in Estimation of Heterogeneous Effects (PEHE), defined as $\mathbb{E}_X\big(\hat{\tau}(X) - \tau(X)\big)^2$, which targets CATE recovery. However, optimizing PEHE is not well suited for inference about counterfactuals.

There are a few notable exceptions where the construction of $\hat{Y}_i(1)$ follows a different principle. **Adversarial approaches**: Yoon et al. (2018) introduce GANITE, which employs adversarial training to generate $\hat{Y}_i(1)$. Although GANITE innovatively bypasses strict model assumptions, it focuses on PEHE and relies on black-box adversarial neural networks without explicitly modeling the joint distribution of potential outcomes. It typically performs well with large dimensions but poorly with small ones. **Bayesian causal inference**: Missing counterfactuals are treated as latent variables, and uncertainty is integrated through the posterior distribution. For example, Alaa and van der Schaar (2017) propose a Bayesian multitask Gaussian process to jointly model $\big(Y(1), Y(0)\big) \mid X$, producing posterior distributions over the potential outcomes. While Bayesian methods offer coherent uncertainty quantification, they rely on strong modeling assumptions and can be sensitive to prior specifications (Li et al., 2022). Moreover, they can be restrictive when aiming to leverage flexible modern machine learning techniques. **Matching methods**: Matching-based approaches (Hur and Liang, 2024) estimate counterfactual outcomes by pairing individuals i, j with similar covariates but different treatments, and approximating the ITE as $Y_j(1) - Y_i(0)$. However, this construction implicitly assumes independence between the potential outcomes ($\rho = 0$). To our knowledge, existing matching methods do not incorporate matching mechanisms that depend directly on the value of Y_i.

More detailed literature review can be found in Appendix A.1.

3. Constructing Retrospective Counterfactual Estimates under Cross-World Assumptions

Our goal is to construct a point estimate and prediction interval for the counterfactual outcome. If both $Y_i(1)$ and $Y_i(0)$ were observable for some individuals, the problem would reduce to classical regression with the factual outcome as an additional covariate. Since this is not possible, inferring counterfactual outcomes remains fundamentally challenging.

Lemma 2 (Non-identifiability of $\mu(x, y)$) *Consider two non-degenerate distributions P and P' satisfying strong ignorability and overlap, that share the same marginal laws of $Y(0) \mid X$ and $Y(1) \mid X$ but differ in their cross-world dependence structure; that is, $\rho(x) \neq \rho'(x)$.*

Then P and P' induce the same observable distribution of (X, T, Y), yet their expected counterfactual outcomes can differ:

$$\mathbb{E}_P[Y(1) \mid X = x, Y(0) = y] \neq \mathbb{E}_{P'}[Y(1) \mid X = x, Y(0) = y]$$

for some (x, y). Therefore $\mu(x, y)$ is not identified from the observed data and its value depends on the unobserved cross-world dependence structure.

Lemma 2 implies that counterfactual prediction cannot be based solely on the observed laws of $Y(0) \mid X$ and $Y(1) \mid X$. Different cross-world dependencies can produce identical observed data but yield different counterfactual predictions. Thus, any meaningful inference requires explicitly specifying a full cross-world dependence structure (unless one is willing to settle for partially identified bounds (Zhang and Richardson, 2025b)).

3.1. Construction under known cross-world dependence

In order to construct an estimate for the counterfactual outcome, a natural starting point is to *separately* construct point estimates and prediction intervals for the treated group and the control group. For point prediction, any machine learning method, such as random forests or neural networks, can be used. For interval estimation, one may use $\frac{\alpha}{2}$- and $(1 - \frac{\alpha}{2})$-quantile regression, conformal prediction, or any other uncertainty-quantification method. We refer to Appendix A.2 for details on classical methods and their properties. Suppose their form is as follows:

$$\begin{gathered}\hat{\mu}_0(x) \text{ and } \hat{\mu}_1(x) \text{ are estimates of } \mu_0(x) \text{ and } \mu_1(x), \text{ respectively, and} \\ \tilde{C}_0(x) = [\hat{\mu}_0(x) - l_0(x),\ \hat{\mu}_0(x) + u_0(x)], \qquad \tilde{C}_1(x) = [\hat{\mu}_1(x) - l_1(x),\ \hat{\mu}_1(x) + u_1(x)],\end{gathered} \tag{3}$$

where $l_t, u_t \geq 0$ are the (lower and upper) widths of prediction intervals for $Y(t), t = 0, 1$. This is visualized in Figure 1. Ideally, $\tilde{C}_t$ satisfy either marginal or conditional coverage:

$$P\big(Y(t) \in \tilde{C}_t(X)\big) \geq 0.9, \quad \text{or} \quad P\big(Y(t) \in \tilde{C}_t(x) \mid X = x\big) \geq 0.9,$$

where marginal coverage is automatically satisfied for conformal methods, while conditional coverage typically requires large sample sizes or strong assumptions in case of high-dimensional X. We combine these quantities to construct a point estimate $\hat{\mu}_\rho$ and a prediction interval C_ρ as follows:

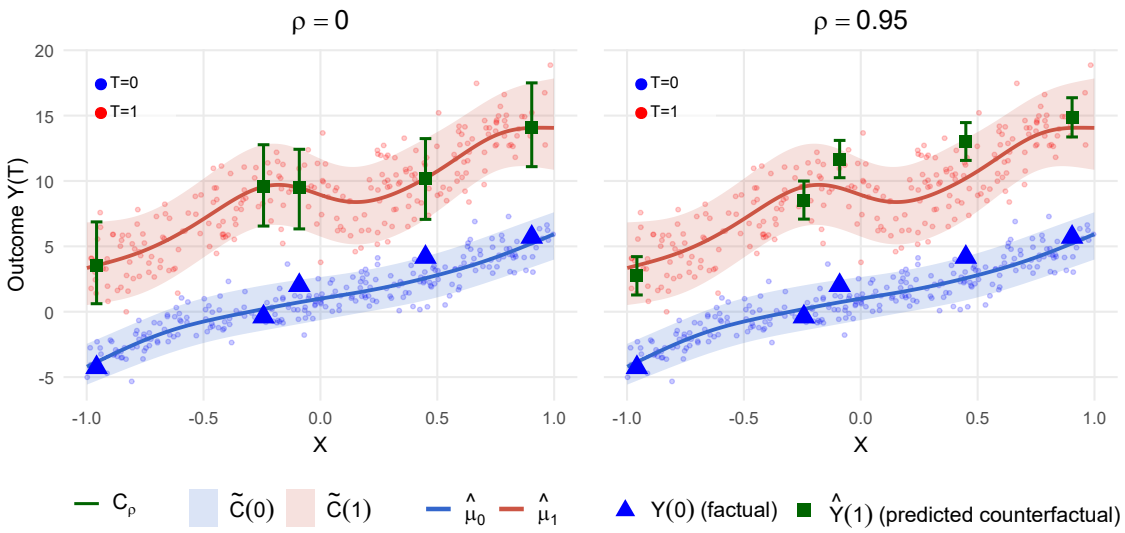

Figure 1: Proposed $\mathrm{RCP}(\rho)$ estimator $\hat{Y}(1) := \hat{\mu}_\rho(x, y)$ and prediction interval $C_\rho(x, y)$, combining baseline predictions with cross-world dependence, shown for two choices of ρ. Across panels, the underlying data and baseline models are identical; increasing ρ only shifts the counterfactual prediction in the direction implied by the factual residual and shrinks the interval. We show $\rho = 0$ (no outcome conditioning) and $\rho = 0.95$ (strong dependence) on five highlighted units.

Definition 3 *Let $\rho \in [-1, 1]$. Consider baseline estimates in the form (3). We define the point predictors:*

$$\hat{\mu}^t_\rho(x, y) = \begin{cases} \hat{\mu}_1(x) + \rho \cdot \lambda(x) \cdot (y - \hat{\mu}_0(x)), & \text{if } t = 0, \\ \hat{\mu}_0(x) + \rho \cdot \dfrac{1}{\lambda(x)} \cdot (y - \hat{\mu}_1(x)), & \text{if } t = 1, \end{cases}$$

where $\lambda(x) = \frac{l_1(x)+u_1(x)}{l_0(x)+u_0(x)}$ is the relative width of the baseline prediction intervals. Given these point predictors, we define the C_ρ intervals by

$$C^t_\rho(x, y) = \begin{cases} \left[\hat{\mu}^t_\rho(x, y) - \sqrt{1-\rho^2} \cdot l_1(x),\ \ \hat{\mu}^t_\rho(x, y) + \sqrt{1-\rho^2} \cdot u_1(x)\right], & \text{if } t = 0, \\ \left[\hat{\mu}^t_\rho(x, y) - \sqrt{1-\rho^2} \cdot l_0(x),\ \ \hat{\mu}^t_\rho(x, y) + \sqrt{1-\rho^2} \cdot u_0(x)\right], & \text{if } t = 1. \end{cases}$$

For notational simplicity, we omit the superscript and write $C_\rho(x, y) = C^t_\rho(x, y)$ and $\hat{\mu}_\rho(x, y) = \hat{\mu}^t_\rho(x, y)$ when evident from context (typically when $t = 0$ and the counterfactual $Y(1)$ is of interest). We refer to the resulting procedure as Retrospective Counterfactual Prediction*, denoted* $\mathrm{RCP}(\rho)$.

The intuition behind the choices for $\hat{\mu}_\rho$ and C_ρ is simple: the larger ρ, the more weight is put on the (centered) factual outcome. The role of $\lambda(x)$ is to adjust for potential differences in variance between treated and untreated groups; in settings where equal variances across groups can be reasonably assumed, one may simply set $\lambda(x) = 1$.

3.2. Consistency and optimality

Under correctly specified cross-world dependence between between $Y(1)$ and $Y(0)$, our estimators are consistent and even asymptotically optimal:

Theorem 4 *Let $x \in \mathcal{X}$ and suppose $\big(Y(1), Y(0)\big) \mid X = x$ is Gaussian with $\rho(x) = \mathrm{cor}\big(Y(1), Y(0) \mid X = x\big) \in [-1, 1]$. Let $\hat{\mu}_t(x)$ be consistent estimators of $\mu_t(x)$, and assume the prediction interval widths $l_t(x), u_t(x)$ are asymptotically conditionally valid*[1].

Then, for any fixed $y \in \mathbb{R}$, the $\hat{\mu}_\rho(x, y)$ is a consistent estimator of the conditional mean,

$$\hat{\mu}_\rho(x, y) \xrightarrow{p} \mathbb{E}\big[Y(1) \mid X = x, Y(0) = y\big], \quad \text{as } n \to \infty.$$

The C_ρ prediction intervals achieve asymptotically perfect conditional coverage,

$$\lim_{n\to\infty} \mathbb{P}\big(Y(1) \in C_\rho(X, Y(0)) \mid X = x, Y(0) = y\big) = 0.9. \tag{4}$$

*Moreover, C_ρ prediction intervals are **asymptotically optimal** in a sense that it is the smallest set satisfying (4) as $n \to \infty$.*

The assumption of Gaussianity and $\rho(x)$ are both modeling assumptions about how $Y(1)$ and $Y(0)$ relate, and neither can be learned from data. The Gaussian copula simply translates a chosen value of $\rho(x)$ into a fully specified cross-world distribution, and any other copula could serve the same role. This highlights the central challenge of retrospective counterfactual prediction: a full dependence structure between the two potential outcomes must be specified, not estimated. Analogous consistency and optimality results to Theorem 4 can be straightforwardly derived under any alternative cross-world dependence structure, under appropriate modifications.

3.3. Special cases: $\rho = 0$ and $\rho = 1$

When $\rho = 0$, our predictions do not depend on y: $\mu_\rho(x, y) = \hat{\mu}_1(x)$ and $C_\rho(x, y) = \tilde{C}_1(x)$, as the factual outcome $Y_i(0)$ provides no information about the missing potential outcome. The problem then reduces to a standard regression setting, as discussed e.g. in Lei and Candès (2021). Under $Y(1) \perp\!\!\!\perp Y(0) \mid X$, our C_ρ intervals inherit the validity of the baseline $\tilde{C}_1$ interval:

$$\mathbb{P}(Y(1) \in \tilde{C}_1(X) \mid X = x) \geq 0.9 \implies \mathbb{P}(Y(1) \in C_\rho(X, Y(0)) \mid X = x, Y(0) = y) \geq 0.9, \tag{5}$$

$$\mathbb{P}(Y(1) \in \tilde{C}_1(X)) \geq 0.9 \implies \mathbb{P}(Y(1) \in C_\rho(X, Y(0))) \geq 0.9. \tag{6}$$

When $\rho = \pm 1$ and $\lambda(x) = 1$, we have $\hat{\mu}_\rho(x, y) = y + \hat{\tau}(x)$ and $C_\rho(x, y) = \{\hat{\mu}_\rho(x, y)\}$, corresponding to a constant treatment effect:

$$\mu_\rho(x, y_0) = \hat{\mu}_\rho(x, y_0) \implies \mathbb{P}(Y(1) \in C_\rho(X, Y(0)) \mid X = x, Y(0) = y) = 1. \tag{7}$$

In practice, however, $\mu_\rho(x, y_0)$ is unknown and must be estimated, introducing bias and potentially non-valid prediction intervals. Section 3.4 discusses how to extend C_ρ intervals to account for the additional uncertainty from this estimation.

1. This holds for many nonparametric estimators under ignorability, overlap and mild smoothness assumptions, including random forests for estimating $\hat{\mu}_t(x)$ (Wager, 2024; Scornet et al., 2015) and (conformalized) quantile random forests for prediction intervals (Meinshausen and Ridgeway, 2006).

3.4. Finite sample bias correction: introducing C_ρ^{+CI} prediction intervals

We enlarge C_ρ *prediction intervals* by adding *confidence intervals* for μ_ρ, estimated for instance via bootstrapping.

Definition 5 *Let $\rho \in [-1, 1]$. Consider prediction intervals for $Y(1)$ and $Y(0)$ of the form (3), and suppose we have confidence intervals $CI(x, y) = [\hat{\mu}_\rho(x, y) - r_l(x, y),\ \hat{\mu}_\rho(x, y) + r_u(x, y)]$. We define the bias-corrected C_ρ^{+CI} intervals as*

$$C_\rho^{+CI}(x, y) = \left[\hat{\mu}_\rho(x, y) - c \cdot r_l(x, y) - \sqrt{1-\rho^2} \cdot l_{1-T_i}(x),\ \ \hat{\mu}_\rho(x, y) + c \cdot r_u(x, y) + \sqrt{1-\rho^2} \cdot u_{1-T_i}(x)\right],$$

where $l_{1-T_i}(x)$ and $u_{1-T_i}(x)$ select the appropriate prediction bounds depending on treatment status T_i, and $c \in [0, 1]$ is a hyperparameter. In simple terms, C_ρ^{+CI} extends C_ρ by adding a scaled confidence interval around $\hat{\mu}_\rho(x, y)$, with scaling factor c. We consider the choice $c = \rho^2$ following the same argument as in (Bodik et al., 2025).

Following (5) and (7), when $\rho = 0$, no adjustment is needed, while for $\rho = \pm 1$, full confidence intervals must be incorporated to guarantee correct coverage. This motivates the choice $c = \rho^2$, ensuring that C_ρ^{+CI} smoothly interpolates between no correction ($\rho = 0$) and full correction ($\rho = 1$). For this choice, we also have the following guarantee.

Consequence 6 *If $\rho = \pm 1$ and confidence intervals satisfy $\mathbb{P}(\mu(x, y_0) \in \hat{\mu}_\rho(x, y_0) \pm r(x, y_0)) \geq 1 - \alpha$, then $\mathbb{P}(Y(1) \in C_\rho^{+CI}(X, Y(0)) \mid X = x, Y(0) = y) \geq 1 - \alpha$.*

4. Experiments

We evaluate our method on synthetic, semi-synthetic, and real datasets using both point estimation and prediction interval metrics, comparing against four baselines under varying cross-world correlation ρ. A user-friendly implementation of our methods in both `R` and `Python`, along with scripts to reproduce all experiments, is available at: https://github.com/jurobodik/Counterfactual_prediction.git

4.1. Details

Datasets: We consider a variety of data-generating processes commonly used in the related literature; full details are provided in Appendix C.1. The **synthetic** datasets feature non-constant propensity scores and randomly generated CATE functions based on smooth random polynomials. These settings allow us to vary the dimensionality $d = \dim(\mathbf{X})$ and the cross-world correlation parameter ρ, thus controlling both complexity and treatment-effect heterogeneity. In addition, we include the **IHDP** dataset, which uses real covariates from a randomized trial and simulated counterfactual outcomes, providing a semi-synthetic benchmark. The **Twins** dataset contains real covariates and real paired outcomes corresponding to different treatment assignments, enabling the construction of both factual and counterfactual outcomes for each unit. In this dataset, we consider $\rho = 0.5$ as a reference choice following Bodik et al. (2025).

Implementation details: To better reflect real-world scenarios where ρ is unknown, we report both i) $\rho_{used} = \rho_{true}$ and ii) $\rho_{used} = \rho_{true} + Unif(-0.5, 0.5)$ capped at $[-1, 1]$.

In our RCP(ρ) framework, we estimate conditional means and quantiles: in low dimensions ($d \leq 5$) we use GAM for the mean and qGAM (Fasiolo et al., 2017) for quantiles, while in higher dimensions ($d > 5$) we switch to random forests for the mean and quantile random forests (Meinshausen and Ridgeway, 2006) for quantiles, trading some low-dimensional efficiency for scalability. TabPFN (Hollmann et al., 2023) is a good potential alternative.

To construct the proposed C_ρ and C_ρ^{+CI} intervals, we use CQR (see Appendix A.2) to produce the base intervals in (3). While more advanced methods often achieve better empirical results, we adopt CQR as a simple, well-established baseline, following Lei and Candès (2021); Alaa et al. (2023), and Bodik et al. (2025).

Baseline methods: In Appendix A.1, we provide details of the existing methods used to estimate counterfactuals. We consider four representative approaches. First, **Direct Outcome (DO) modeling** estimates the counterfactuals by fitting the treatment-specific regression $\hat{Y}_i(1) := \mu_1(X_i)$ using Random Forests (Wager and Athey, 2018) or Generalized Additive Models (Fasiolo et al., 2017) (using the same choices as in C_ρ). Second, **CATE-adjusted imputation** estimates the CATE via a T-learner (Künzel et al., 2019), DR-learner (doubly robust, Dukes et al. (2024)) or Generalized Random Forest (Athey et al., 2019), and adjusts the observed outcome using $\hat{Y}_i(1) = Y_i(0) + \hat{\tau}(X_i)$. We only report the T-learner as it yielded the best results on the considered datasets. Note that while many other CATE estimators exist, the goal is to illustrate the core imputation approach, which remains fundamentally limited even with perfectly estimated CATE. Third, **Matching-based imputation** uses nearest-neighbor matching with Mahalanobis distance to impute the missing potential outcome from similar units in the opposite treatment group. Fourth, **adversarial generative modeling** employs GANITE (Yoon et al., 2018), a two-stage generative adversarial network that imputes and refines counterfactual predictions, typically suitable only in high-dimensional, nonlinear settings.

Setup: We conducted experiments on datasets: synthetic ($n = 1000$), IHDP ($n = 747$), and Twins ($n = 11{,}983$). Each synthetic and IHDP experiment was repeated 50 times to reduce Monte Carlo variability, while the Twins dataset was analyzed once using the full sample. All methods used an 80/20 train–calibration split for CQR and prediction intervals at level $\alpha = 0.1$. Computing μ_ρ and C_ρ is fast, as the main cost lies in fitting four quantile regressions; however, C_ρ^{+CI} requires computing bootstrap confidence intervals (we used 100 bootstraps), which is computationally more intensive; running all datasets and repetitions took approximately four days on an Intel Core i5-6300U (2.5 GHz, 16 GB RAM).

Metrics: To assess performance, we use MSE for point predictions and the Interval Score (metric that combines coverage and width) for prediction intervals:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i^{cf} - Y_i^{cf})^2, \qquad \text{IS}_\alpha = \frac{1}{n}\sum_{i=1}^{n}(U_i - L_i) + \frac{2}{\alpha}\big[(L_i - Y_i^{cf})_+ + (Y_i^{cf} - U_i)_+\big],$$

where $[L_i, U_i]$ are the estimated prediction intervals at level $1-\alpha$ and $z_+ = \max(z, 0)$.

4.2. Results of the experiments

Figure 2 reports the MSE of point predictions; Figure 5 in Appendix C.2 reports the interval scores. Both RCP(ρ) variants (correctly specified and misspecified) outperform the baselines in the regimes where conditioning on the factual outcome is informative, i.e., when $0 < \rho < 1$.

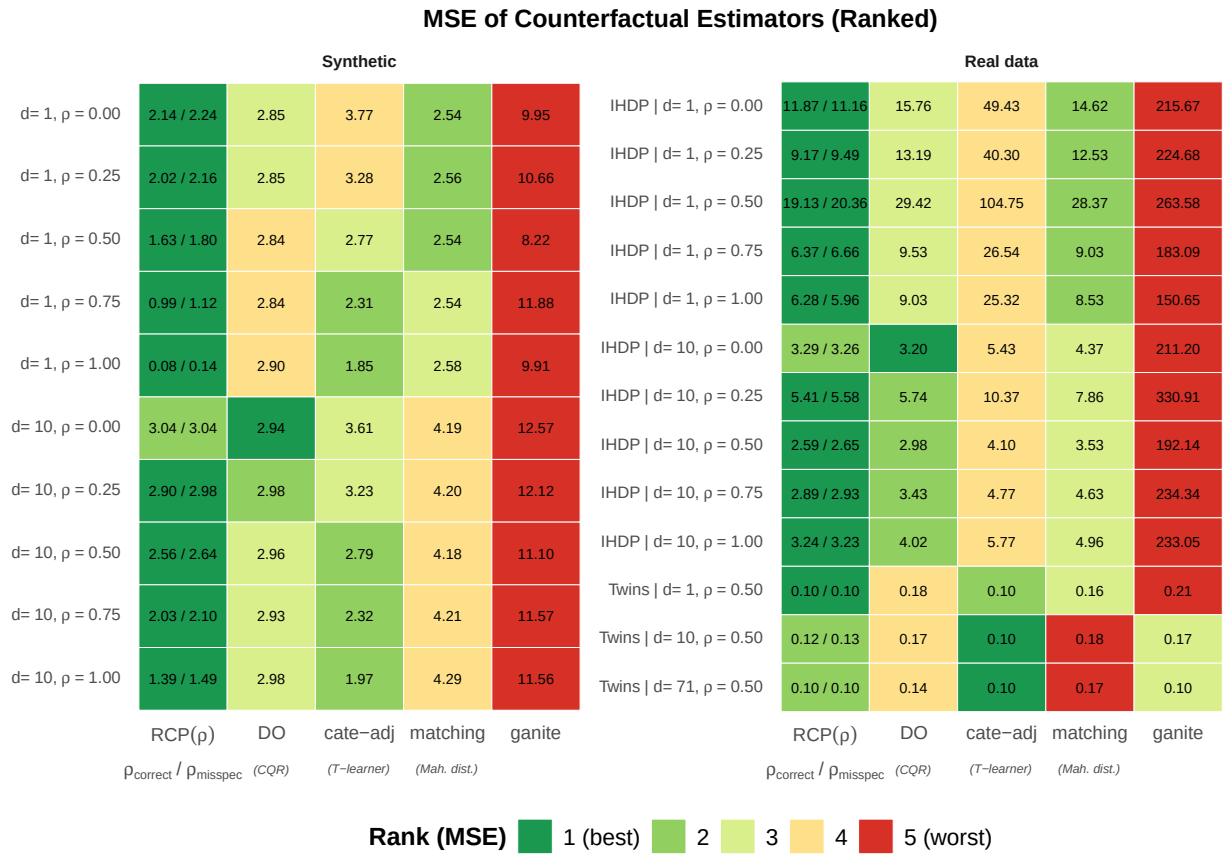

Figure 2: Mean squared error of different estimators across different datasets, averaged over 50 repetitions. In μ_ρ, we use either $\rho = \rho_{true}$, or mimic misspecification by using $\rho = \rho_{true} + Unif(-0.5, 0.5)$. Standard deviations for each entry can be found in Figure 7 located in Appendix C.

When the cross-world dependence is known, conditioning on $Y(0)$ reduces the conditional variance of the counterfactual from $\mathrm{Var}(Y(1) \mid X = x)$ to $\mathrm{Var}(Y(1) \mid X = x, Y(0) = y)$. In our synthetic experiments, $(Y(0), Y(1))$ are generated via a Gaussian copula with correlation ρ, and the resulting variance reduction has a closed-form scaling: $\frac{\mathrm{Var}(Y(1)|X=x,Y(0)=y)}{\mathrm{Var}(Y(1)|X=x)} = 1 - \rho^2$. Thus, as ρ increases, the factual $Y(0)$ becomes more informative about $Y(1)$, so the best achievable MSE decreases roughly proportionally to $(1 - \rho^2)$. This explains the monotone MSE decrease with ρ in Figure 2, and also why the $\rho = 0$ case performs similarly as DO (the factual outcome provides no additional signal beyond X).

In finite samples, the observed MSE is not only driven by this dependence-induced variance reduction, but also by error from estimating nuisance components. This contribution grows with dimension and can blur the clean $(1 - \rho^2)$ scaling, making improvements dramatic in low dimension but more muted in higher dimension. While GANITE performs poorly in our settings, it was designed for higher-dimensional regimes and typically benefits from larger d and n. Our method is primarily targeted to lower dimensions where $Y(T)$ contains substantial information beyond X.

In a few real-world datasets, ρ_{misspec} yields slightly better performance than ρ_{correct}; this is a consequence of Monte Carlo variability. As shown in Figure 7, the corresponding confidence intervals

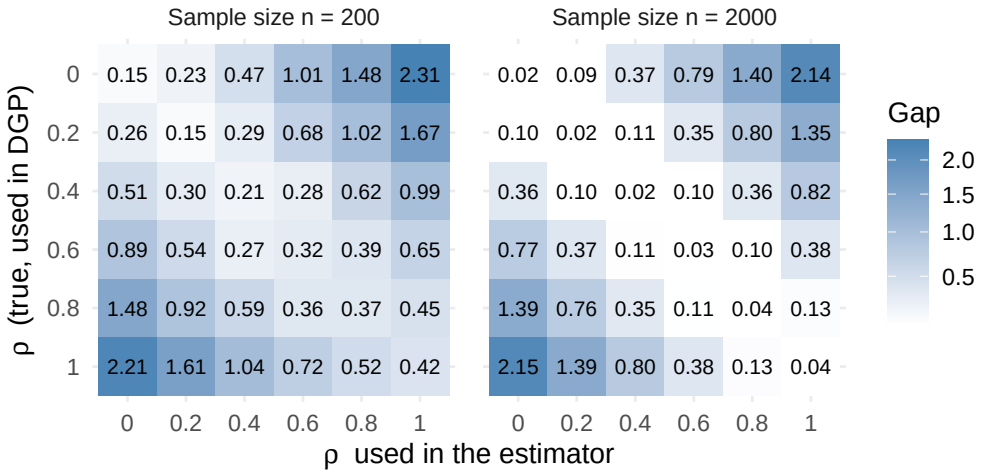

Figure 3: Gap $= \text{MSE}_{\text{RCP}} - \text{MSE}_{\text{oracle}}$ calculated across different misspecifications of ρ. Bias persists when ρ is far from the truth, but vanishes asymptotically if ρ is specified correctly. This demonstrates that incorporating even approximate knowledge of cross-world dependence improves counterfactual predictions.

are very large in these cases, and resolving these differences would require much more repetitions to reduce simulation noise.

4.3. Additional experiments: misspecified ρ and non-Gaussianity

We conduct two additional experiments, evaluating the Gap $= \text{MSE}_{\text{RCP}} - \text{MSE}_{\text{oracle}}$, where the oracle estimator is equal to the true $\mathbb{E}[Y^{cf} \mid X, Y^{obs}, T]$. All details can be found in Appendix B.

- **(Misspecifying ρ).** Figure 3 reports experiments on synthetic data varying the *true* correlation ρ_{true} in the data-generating process (DGP) and the *assumed* value ρ_{est} in our estimator. Bias grows with misspecification $|\rho_{\text{est}} - \rho_{\text{true}}|$, and vanishes with larger n only when ρ_{est} is close to ρ_{true}; otherwise, it persists even asymptotically. This shows that even rough knowledge of ρ yields large gains over ignoring the factual outcome.
- **(Robustness to non-Gaussianity).** Appendix B.1 (Figure 4) contains experiments with non-Gaussian outcome distributions $(Y(0), Y(1))$. In all cases the gap vanishes with n, though convergence is slower under non-Gaussian noise. Discrepancies are most visible at $\rho = 1$.

5. Conclusion, Limitations and Future Research

The factual outcome carries valuable individual-level information that should not be ignored in retrospective counterfactual prediction. We formalize this importance through the cross-world correlation parameter ρ, which determines how strongly observed and unobserved outcomes are linked. By treating ρ as an explicit modeling choice, our approach interpolates between classical extremes, with $\rho = 0$ discarding the factual outcome and $\rho = 1$ assuming constant effects, and delivers predictions that are theoretically well motivated and empirically effective whenever even approximate knowledge of ρ is available.

The biggest practical limitation of our approach is the necessity of specifying the cross-world dependence. This requirement reflects a fundamental barrier rather than a methodological issue: the

relationship between $Y(1)$ and $Y(0)$ is inherently unidentifiable. However, *every existing method already makes a fixed, implicit assumption about* ρ. Our contribution is to make this dependence explicit, enabling practitioners to incorporate domain knowledge, auxiliary data or sensitivity analysis into retrospective counterfactual inference. This transparency clarifies the assumptions needed for inference and opens new possibilities for research.

Future work should investigate how the finite-sample behavior depends on the strength and smoothness of dependence, aiming to better understand rates of convergence when ρ is known or restricted. A second direction is to explore richer, non-Gaussian dependence structures that may be more appropriate in some domain applications. Another promising direction is to extend the methodology to continuous treatments (Bodik, 2024) or dynamic settings such as time series (Li et al., 2026; Bodik et al., 2024), where cross-world assumptions could provide structure for dose–response counterfactual curves. Finally, selecting such structures in practice remains an open challenge; developing tools or diagnostics to guide these choices would broaden the applicability of cross-world inference.

Reproducibility Statement and Usage of Large Language Models

All code and datasets used in this work are provided in the supplementary material to ensure full reproducibility of our results. We declare that we used a large language model for grammar and language polishing, as well as for limited coding assistance (e.g., boilerplate code and debugging). All conceptual and theoretical contributions, experimental designs, and conclusions are our own.

References

A. Abadie and G. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 2006.

A. Agarwal, M. Xiao, R. Barter, O. Ronen, B. Fan, and B. Yu. PCS-UQ: Uncertainty quantification via the predictability-computability-stability framework, 2025. URL https://arxiv.org/abs/2505.08784.

A. Alaa and M. van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6a508a60aa3bf9510ea6acb021c94b48-Paper.pdf.

A. Alaa, Z. Ahmad, and M. van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=IwnINorSZ5.

R. M. Andrews and V. Didelez. Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology*, 32(2):209–219, 2021. doi: 10.1097/EDE.0000000000001313.

A. N. Angelopoulos, R. F. Barber, and S. Bates. Theoretical foundations of conformal prediction, 2024. URL https://arxiv.org/abs/2411.11824.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2): 1148–1178, 2019.

I. Azizi, J. Bodik, J. Heiss, and B. Yu. Clear: Calibrated learning for epistemic and aleatoric risk, 2025. URL https://arxiv.org/abs/2507.08150.

K. Bairaktari, R. Izbicki, and E. J. Candès. Kandinsky conformal prediction: Beyond class- and covariate-conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.17264.

R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference, 2020. URL https://arxiv.org/abs/1903.04684.

I Bica, J Jordon, and M van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16434–16445. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/bea5955b308361a1b07bc55042e25e54-Paper.pdf.

J. Bodik. Extreme treatment effect: Extrapolating dose-response function into extreme treatment domain. *Mathematics*, 12(10), 2024. doi: 10.3390/math12101556.

J. Bodik and V. Chavez-Demoulin. Structural restrictions in local causal discovery: identifying direct causes of a target variable. *Biometrika*, 2025. URL https://arxiv.org/abs/2307.16048.

J. Bodik, M. Paluš, and Z. Pawlas. Causality in extremes of time series. *Extremes*, 27:67–121, 2024. doi: 10.1007/s10687-023-00479-5.

J. Bodik, Y. Huang, and B. Yu. Cross-world assumption and refining prediction intervals for individual treatment effects. *ArXiv preprint ArXiv:2507.12581*, 2025. URL https://arxiv.org/abs/2507.12581.

M. Cai, S. Buuren, and V. Gerko. How to relate potential outcomes: Estimating individual treatment effects under a given specified partial correlation, 2022. URL https://arxiv.org/abs/2208.12931.

K. E. Colson et al. Optimizing matching and analysis combinations for estimating causal effects. *Scientific Reports*, 2016.

H. Dance and B. Bloem-Reddy. Counterfactual cocycles: A framework for robust and coherent counterfactual transports, 2025. URL https://arxiv.org/abs/2405.13844.

A. Dieng et al. Interpretable almost-exact matching for causal inference. *Journal of Causal Inference*, 2019.

P. Ding and F. Li. Causal inference: A missing data perspective. *Statistical Science*, 33(2):214–237, 2018. doi: 10.1214/18-STS645. URL https://doi.org/10.1214/18-STS645.

P. Ding, A. Feller, and L. Miratrix. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.

O. Dukes, S. Vansteelandt, and D. Whitney. On doubly robust inference for double machine learning in semiparametric regression. *Journal of Machine Learning Research*, 25(279):1–46, 2024. URL http://jmlr.org/papers/v25/22-1233.html.

Y. Fan and S. S. Park. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951, 2010. doi: 10.1017/S0266466609990168.

M. Fasiolo, Y. Goude, R. Nedellec, and S. Wood. Fast calibrated additive quantile regression, 2017. Available at https://arxiv.org/abs/1707.03307.

S. P. Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1): 259–276, 2007.

I. Gibbs, J. J. Cherian, and E. J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf008, 03 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf008. URL https://doi.org/10.1093/jrsssb/qkaf008.

J. J. Heckman, J. Smith, and N. Clements. Making the most out of program evaluations and social experiments: Accounting for heterogeneity in program impacts. *Review of Economic Studies*, 64 (4):487–535, 1997.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second, 2023. URL https://arxiv.org/abs/2207.01848.

Y. Hur and T. Liang. A convexified matching approach to imputation and individualized inference, 2024. URL https://arxiv.org/abs/2407.05372.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015. doi: 10.1017/CBO9781139025751.

F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3020–3029, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/johansson16.html.

J. Jonkers, J. Verhaeghe, G. Wallendael, L. Duchateau, and S. Hoecke. Conformal convolution and monte carlo meta-learners for predictive inference of individual treatment effects, 2024. URL https://arxiv.org/abs/2402.04906.

S. Joshi, A. Korba, T. Trogdon, and E. Candès. Conformal inference under high-dimensional covariate shifts via likelihood-ratio regularization. *arXiv preprint arXiv:2502.13030*, 2025. URL https://arxiv.org/abs/2502.13030.

N. Kallus. A Framework for Optimal Matching for Causal Inference. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 372–381. PMLR, 20–22 Apr 2017.

K. Kasa, Z. Zhang, H. Yang, and G. W. Taylor. Adapting prediction sets to distribution shifts without labels. In *Uncertainty in Artificial Intelligence (UAI)*, 2025. URL https://proceedings.mlr.press/v286/kasa25a.html.

K. Kim. Semiparametric counterfactual regression. *arXiv preprint arXiv:2504.02694*, 2025.

K. Kim, E. H. Kennedy, and J. R. Zubizarreta. Doubly robust counterfactual classification. In *Advances in Neural Information Processing Systems*, volume 35, pages 34831–34845, 2022.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116 (10):4156–4165, 2019.

A. Lacombe and M. Sebag. Asymmetrical latent representation for individual treatment effect modeling, 2025. URL https://arxiv.org/abs/2501.14006.

L. Lei and E. J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, November 2021. doi: 10.1111/rssb.12445.

F. Li, P. Ding, and F. Mealli. Bayesian causal inference: A critical review, 2022. URL https://arxiv.org/abs/2206.15460.

M. M Li, K. Li, Y. Ektefaie, Y. Jin, Y. Huang, S. Messica, T. Cai, and M. Zitnik. Controllable sequence editing for biological and clinical trajectories. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dvB0QDmnry.

C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6449–6459, 2017.

Y. Ma, V. Melnychuk, J. Schweisthal, and S. Feuerriegel. Diffpo: A causal diffusion model for learning distributions of potential outcomes. In *Advances in Neural Information Processing Systems*, volume 37, pages 43663–43692. Curran Associates, Inc., 2024. doi: 10.52202/079017-1384.

A. McClean, Z. Branson, and E. H. Kennedy. Nonparametric estimation of conditional incremental effects. *Journal of Causal Inference*, 12(1):20230024, 2024. doi: 10.1515/jci-2023-0024.

N. Meinshausen and G. Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 2006.

R. B. Nelsen, J. Quesada-Molina, J. Antonio Rodriiguez-Lallena, and M. Úbeda-Flores. Bounds on bivariate distribution functions with given margins and measures of association. *Communications in Statistics - Theory and Methods*, 30(6):1055–1062, 2001. doi: 10.1081/STA-100104355.

J. Pearl and D. Mackenzie. *The Book of Why*. Penguin Books, 2019. URL http://bayes.cs.ucla.edu/WHY/.

K. Perlin. An image synthesizer. *Siggraph Comput. Graph.*, 19(0097-8930):287–296, 1985.

V. Plassier, O. Bouhali, and N. El Karoui. Rectifying conformity scores for better conditional coverage. *arXiv preprint arXiv:2502.16336*, 2025. URL https://arxiv.org/abs/2502.16336.

Y. Romano, E. Patterson, and E. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pages 3538–3548, 2019.

D. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. doi: 10.1198/016214504000001880.

D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, November 1990. doi: 10.1214/ss/1177012032.

E. Scornet, G. Biau, and J. Vert. Consistency of random forests. *The Annals of Statistics*, 43(4), 2015. doi: 10.1214/15-aos1321.

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 2010.

R. J. Tibshirani, R. F. Barber, E. Candès, and A. Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, pages 2530–2540, 2019.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

S. Wager. Causal inference: A statistical learning approach, 2024. URL https://web.stanford.edu/~swager/causal_inf_book.pdf. Stanford University.

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

B. Wang and X. Qiao. Conformal prediction under generalized covariate shift with posterior drift. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025. URL https://arxiv.org/abs/2502.17744.

R. W. Wright. Causation in tort law. *California Law Review*, 73(6):1735–1828, 1985. doi: 10.2307/3480056.

L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a50abba8132a77191791390c3eb19fe7-Paper.pdf.

J. Yoon, J. Jordon, and M. van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByKWUeWA-.

B. Yu and R. L. Barter. *Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making*. Adaptive Computation and Machine Learning Series. MIT Press, 2024. ISBN 0262379708, 9780262379700.

Z. Zhang and T. S. Richardson. Bounds on the distribution of a sum of two random variables: Revisiting a problem of kolmogorov with application to individual treatment effects, 2025a. URL https://arxiv.org/abs/2405.08806.

Z. Zhang and T. S. Richardson. Individual treatment effect: Prediction intervals and sharp bounds, 2025b. URL https://arxiv.org/abs/2506.07469.

Appendix

Appendix A. Literature review: details

A.1. Counterfactual estimation methods: other approaches

We consider four classes of approaches for estimating the unobserved potential outcome $Y_i(1)$ for units with $T_i = 0$ (and analogously $Y_i(0)$ for $T_i = 1$).

Direct outcome modeling (*DO*). This approach estimates the treatment-specific regression function $\mu_1(x) = \mathbb{E}[Y \mid X = x, T = 1]$ directly from the treated sample and use $\hat{Y}_i(1) = \hat{\mu}_1(X_i)$ for counterfactual prediction. We consider two implementations: Random Forests (RF) (Wager and Athey, 2018) and Generalized Additive Models (GAM) (Fasiolo et al., 2017). Unlike the CATE-adjusted approach, these methods do not require access to the observed control outcome $Y_i(0)$ for the unit, relying entirely on model-based extrapolation from treated units. To quantify uncertainty, we use the same prediction intervals as in (3).

There is also a large number of similar approaches besides RF and GAM, also adjusting for the distribution shift between the treated/untreated groups. Yao et al. (2018) employ deep representation learning to estimate $\hat{Y}_i(1-T) = g(f(X_i), T_i)$ where f, g are neural networks based preserving local similarity between the treated groups.

CATE-adjusted imputation (*CATE-adj*). This approach first estimates CATE $\tau(X_i)$ and then shifts the observed control outcome by this estimated effect:

$$\hat{Y}_i(1) = Y_i(0) + \hat{\tau}(X_i).$$

We use three alternative CATE estimators: the T-learner (Künzel et al., 2019), the Generalized Random Forest (GRF) (Athey et al., 2019), and a doubly robust (DR) estimator (Dukes et al., 2024). Closely related meta-learners include the S-learner, which fits a single model with treatment as an input feature, and the X-learner, which augments the T-learner with imputed treatment effects for the opposite treatment group and often performs well under treatment imbalance (Künzel et al., 2019). These alternative meta-learners share the same conceptual foundation. Johansson et al. (2016); Lacombe and Sebag (2025) use deep learning alternatives; balancing counterfactual regression or adding assymetrical latend represnetation.

To quantify uncertainty, confidence intervals are computed using standard procedures, obtaining prediction intervals in a form $\hat{Y}_i(1) = Y_i(0) + \hat{\tau}(X_i) \pm conf.int(\hat{\tau}(X_i))$. In our experiments, we only considered T-learner, GRF and DR estimators for CATE-adjusted imputation, as other approaches are typically significantly more performative only in high-dimensional datasets or when treated and untreated units differ substantially, which is not the case in our datasets.

Matching-based imputation (*Matching*). This approach imputes missing potential outcomes using outcomes from similar units in the opposite treatment group, selected via a distance metric in covariate space (Stuart, 2010; Abadie and Imbens, 2006). Beyond nearest-neighbor and optimal matching, advances include kernel-based matching to minimize estimation error (Kallus, 2017) and full or genetic matching combined with double-robust analysis for improved bias and efficiency (Colson et al., 2016). For high-dimensional or categorical data, algorithms like DAME prioritize relevant covariates (Dieng et al., 2019). Similar ideology was also used in ALRITE (Lacombe and

Sebag, 2025), where the authors imputed counterfactuals based on the closest distance in a latent space, in order to improve CATE estimation.

We implemented nearest-neighbor matching with a uniform kernel and optional replacement, using either the Mahalanobis distance between standardized covariates or the absolute difference in logit propensity scores (the former led to better results so we only report that). The propensity scores is estimated by standard classification forest. For a treated unit, the counterfactual $\hat{Y}_i(0)$ is the average outcome among its matched controls, and vice versa for control units. This nonparametric approach relies on local overlap in covariates and assumes conditional independence of potential outcomes and treatment given covariates. To quantify uncertainty, we construct unit-level prediction intervals for the counterfactuals using the empirical variance of the donor outcomes: for a unit with $K \geq 2$ matches, the half-width is given by $t1-\alpha/2, K-1 \cdot s/\sqrt{K}$, where s is the sample standard deviation of the matched donor outcomes, yielding ($\hat{Y}_i^{\text{cf}} \pm$ half-width); if $K = 1$, the half-width is zero. This approach implicitly assumes conditional independence of potential outcomes ($\rho = 0$, similarly to DO) and independent treatment given covariates.

Adversarial generative modeling (*GANITE*). (Yoon et al., 2018) employs a two-stage generative adversarial network (GAN) framework tailored to causal inference. In the first stage, a generator–discriminator pair is trained to impute the missing counterfactual outcomes by making the generated outcomes indistinguishable from observed ones given covariates and treatment assignment. In the second stage, a separate adversarial network refines these predictions to improve estimation of individualized treatment effects, encouraging accurate recovery of both potential outcomes simultaneously. This approach is particularly suited to high-dimensional, nonlinear settings. Some extentions were also proposed that work better under some alternative scenarios (e.g. SCIGAN-ITE by Bica et al. (2020)).

Other approaches. Some other approaches exist, such as **Bayesian causal inference**, where the missing counterfactuals are treated as latent variables, and uncertainty is integrated through the posterior distribution. For example, Alaa and van der Schaar (2017) propose a Bayesian multitask Gaussian process to jointly model $\big(Y(1), Y(0)\big) \mid X$, producing posterior distributions over the potential outcomes. While Bayesian methods offer coherent uncertainty quantification, they often rely on strong modeling assumptions and can be sensitive to prior specifications (Li et al., 2022). Moreover, they can be restrictive when aiming to leverage flexible modern machine learning techniques.

A.2. Uncertainty quantification and prediction intervals in classical regression

In a standard regression framework, we observe data $(X_i, Y_i) \sim P_X \times P_{Y|X}$ for $i = 1, \ldots, n$, and seek a prediction set $C(X)$ for future responses that satisfies a coverage property. Two common notions of coverage are:

$$\mathbb{P}\big(Y_{n+1} \in C(X_{n+1})\big) \geq 1-\alpha \qquad \text{(marginal coverage)},$$
$$\mathbb{P}\big(Y_{n+1} \in C(X_{n+1}) \mid X_{n+1} = x\big) \geq 1-\alpha \qquad \text{(conditional coverage)}.$$

Conditional coverage is a stronger requirement but is generally unattainable in a distribution-free, finite-sample setting without strong assumptions or asymptotics (Barber et al., 2020). By contrast, marginal coverage can be attained without modeling assumptions via conformal prediction (Angelopoulos et al., 2024). Recent work has also explored data-driven techniques to improve

conditional coverage, such as combining epistemic+aleatoric sources of uncertainty (Azizi et al., 2025), rectifying conformity scores (Plassier et al., 2025), or optimizing subgroup-conditional guarantees through flexible frameworks like Kandinsky conformal prediction (Bairaktari et al., 2025). These developments are consistent with the broader principles of Predictability, Computability, and Stability (PCS) advocated for trustworthy data science (Agarwal et al., 2025; Yu and Barter, 2024).

Conformal methods produce prediction intervals with exact finite-sample marginal coverage under exchangeability of the observed and future data points (Vovk et al., 2005; Angelopoulos et al., 2024). These methods typically split the data into training and calibration subsets, construct a preliminary predictor on the training set, and adjust it on the calibration set to guarantee coverage. A prominent example is Conformalized Quantile Regression (CQR), which uses estimated conditional quantiles to build tighter prediction intervals (Romano et al., 2019).

Estimation procedure for CQR. The key idea of CQR is to combine quantile regression with conformal calibration:

1. **Split the data.** Randomly divide the dataset into a training set $\mathcal{D}_{\text{train}}$ and a calibration set $\mathcal{D}_{\text{calib}}$. The split fraction is typically 80/20.

2. **Fit quantile regression models.** On $\mathcal{D}_{\text{train}}$, estimate the conditional lower and upper quantile functions $\hat{q}_{\alpha/2}(x)$ and $\hat{q}_{1-\alpha/2}(x)$, often quantile random forest (Meinshausen and Ridgeway, 2006), qGAM (Fasiolo et al., 2017) or neural networks to approximate conditional quantiles for levels $\alpha/2$ and $1-\alpha/2$.

3. **Compute conformity scores.** For each $(X_i, Y_i) \in \mathcal{D}_{\text{calib}}$, compute the nonconformity score:

$$s_i = \max\{\hat{q}_{\alpha/2}(X_i) - Y_i,\ Y_i - \hat{q}_{1-\alpha/2}(X_i),\ 0\}.$$

   This measures how far Y_i lies outside the estimated conditional quantile interval.

4. **Calibrate using empirical quantiles.** Let $Q_{1-\alpha}(s_1, \ldots, s_m)$ be the $(1-\alpha)$-empirical quantile of the scores from the calibration set ($m = |\mathcal{D}_{\text{calib}}|$).

5. **Construct prediction intervals.** For a new point x, the CQR prediction set is:

$$\tilde{C}(x) = \left[\hat{q}_{\alpha/2}(x) - Q_{1-\alpha},\ \hat{q}_{1-\alpha/2}(x) + Q_{1-\alpha}\right].$$

This adjustment ensures that the final interval achieves marginal coverage at level $1-\alpha$ in finite samples under exchangeability, while leveraging conditional quantile estimates for tighter intervals.

However, exchangeability (slightly weaker assumption than i.i.d.) can fail in the presence of covariate shift, e.g., in observational studies comparing treated and untreated units. In such settings, even defining marginal coverage requires specifying the *target covariate distribution*: should coverage be with respect to $P_{X|T=1}$ (treated), $P_{X|T=0}$ (untreated), or a mixture P_X? This point is emphasized in Lei and Candès (2021). If one could attain conditional coverage, covariate shift would not pose a problem (recall that conditional coverage implies marginal coverage under any P_X) but such guarantees remain scarce (Gibbs et al., 2025).

To address distributional shift, weighted conformal prediction adjusts calibration via importance weights derived from the likelihood ratio between covariate distributions; when this ratio is known, one can guarantee exact marginal coverage for the chosen target population (Tibshirani et al., 2019).

When the ratio (or propensity score $\pi(x)$) is estimated, asymptotically valid marginal coverage is still achievable, with strong empirical performance (Lei and Candès, 2021). Recent approaches refine this idea by incorporating likelihood-ratio regularization for high-dimensional covariates (Joshi et al., 2025) or leveraging unlabeled test data to adapt coverage under label scarcity (Kasa et al., 2025). For settings with both covariate shift and posterior drift, weighted conformal classifiers have been proposed (Wang and Qiao, 2025).

Appendix B. Additional Experiments: misspecified ρ and non-Gaussianity

B.1. How vital is the assumption of Gaussianity?

We evaluate the sensitivity of our counterfactual estimation method to violations of the Gaussianity assumption in the joint distribution of potential outcomes. Specifically, we use the Synthetic dataset described in Appendix C.2, but replace the Gaussian error terms with non-Gaussian marginals coupled through different copulas. Formally, for each unit i, we generate

$$(\varepsilon_i^0, \varepsilon_i^1) \overset{i.i.d}{\sim} \mathrm{Copula}_\rho(F_0, F_1)\,,$$

where F_t denotes the marginal distribution of ε_i^t (e.g., $t = 0, 1$ could follow Student-t, Laplace, or Chi-square distributions), and Copula_ρ is a copula with correlation ρ. By Sklar's theorem, this ensures that the joint distribution of $(\varepsilon_i^0, \varepsilon_i^1)$ has the specified marginals while preserving the desired correlation structure through Copula_ρ. We experiment with Gaussian and Gumbel copulas to capture symmetric as well as asymmetric dependence patterns.

We vary the following factors:

- Marginal distributions: Gaussian, Student-t ($df = 3$), Laplace, and Chi-square ($df = 3$),
- Copula families: Gaussian and Gumbel,
- Cross-world correlation: $\rho \in \{0, 0.5, 1\}$,
- Sample size: $n \in \{100, 300, 500, 2000\}$ with covariate dimension fixed at $d = 1$.

For each configuration, we generate 50 replications and compare our estimate $\hat{\mu}_\rho$ against the oracle estimator

$$\hat{Y}^{cf}_{\text{oracle}} := \mathbb{E}[Y^{\text{cf}} \mid X, Y^{\text{obs}}, T],$$

which leverages the true joint distribution. We report the performance gap

$$\text{Gap} = \text{MSE}_{\text{RCP}} - \text{MSE}_{\text{oracle}}, \qquad \text{MSE}_{\text{RCP}} = \tfrac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i^{cf} - Y_i^{cf})^2, \qquad \hat{Y}_i^{cf} = \hat{\mu}_\rho.$$

Figure 4 summarizes the results. In all cases, the gap decreases with n, demonstrating that our $RCP(\rho)$ estimator converges to the oracle regardless of the marginal distribution or copula. The effect of non-Gaussianity is therefore limited to finite samples: convergence is noticeably slower under heavy-tailed or skewed marginals, particularly when $\rho = 1$, but the asymptotic behavior remains unchanged. By contrast, under independence ($\rho = 0$), our estimator is nearly indistinguishable from the oracle even in small samples.

In conclusion, violations of Gaussianity do not seem to threaten the validity of our method, but they can slow finite-sample convergence; especially under large cross-world dependence.

B.2. Details about Figure 3 and misspecified ρ

To study the effect of misspecifying the cross-world correlation ρ, we carried out a grid experiment on synthetic data. For each design point, we distinguish between the **true** value ρ_{true} used in the data-generating process (DGP), and the **assumed** value ρ_{est} used in our estimator $\hat{\mu}_\rho$.

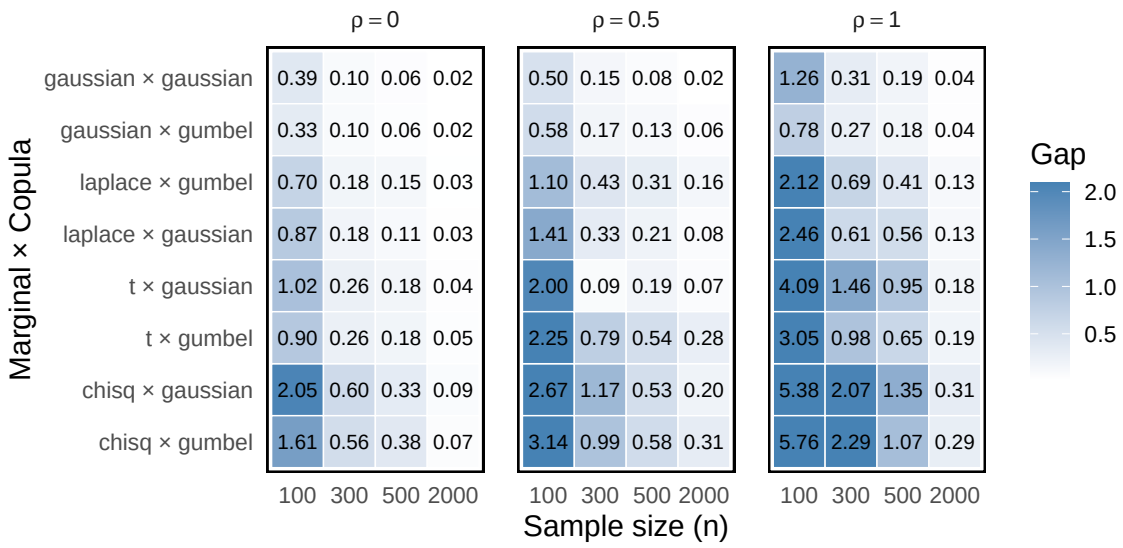

Figure 4: Gap $= \text{MSE}_{\text{our}} - \text{MSE}_{\text{oracle}}$ calculated across different marginal–copula distributions of potential outcomes $\big(Y(0), Y(1)\big)$. Here, we only considered correctly specified ρ in the estimation.

We consider the *synthetic* dataset (see Section C.1), a univariate covariate setting ($d = 1$), two sample sizes ($n = 200$ and $n = 2000$), and repeated each experiment 50 times to reduce Monte Carlo variability. The true correlation ρ_{dgp} was varied over a grid $\{0, 0.1, \ldots, 1\}$, and for each value we estimated counterfactuals under a grid of assumed correlations $\rho_{\text{est}} \in \{0, 0.1, \ldots, 1\}$.

For each pair $(\rho_{\text{true}}, \rho_{\text{est}})$, we generated synthetic data, computed counterfactual estimates with our method using ρ_{est}, and compared performance against the oracle estimator $\mathbb{E}[Y^{cf} \mid X, Y^{obs}, T]$. We measured performance using the mean squared error (MSE) of counterfactual predictions, and summarized results via the Gap $= \text{MSE}_{\text{our}} - \text{MSE}_{\text{oracle}}$. Results (Figure 3) show that the gap increases systematically with the degree of misspecification $|\rho_{\text{est}} - \rho_{\text{true}}|$. When the assumed correlation is close to the truth, the gap shrinks as n grows, and bias vanishes asymptotically. In contrast, for larger misspecifications, the bias persists even at large n, indicating that asymptotic consistency requires $\rho_{\text{est}} \approx \rho_{\text{true}}$. These results show the importance of approximate domain knowledge of ρ: even approximate information about its value can yield large gains over methods that implicitly assume $\rho = 0$ or $\rho = 1$.

Appendix C. Appendix: Numerical Experiments

We provide full details about our experiments below.

C.1. Datasets

We investigate three types of data-generating mechanisms:

- **Synthetic** (taken from (Bodik et al., 2025)): For the univariate case ($d = 1$), we draw $X \sim \mathrm{Unif}(-1, 1)$. When $d > 1$, we follow the setup in Wager and Athey (2018); Alaa et al. (2023); Lei and Candès (2021); Jonkers et al. (2024) and generate covariates $\mathbf{X} = (X_1, \ldots, X_d)$, where each $X_j = \Phi(\tilde{X}_j)$ and Φ is the standard normal CDF. The latent vector $(\tilde{X}_1, \ldots, \tilde{X}_d)$ is sampled from a multivariate Gaussian distribution with zero mean and constant pairwise correlation $\mathrm{Cov}(\tilde{X}_j, \tilde{X}_{j'}) = 0.25$ for $j \neq j'$. Treatment assignments are drawn from a propensity score function

  $$\pi(\mathbf{X}) = \frac{1 + |X_1|}{4} \in [0.25, 0.5],$$

  ensuring adequate overlap. The potential outcomes are defined as

  $$Y_i(0) = f_0(\mathbf{X}_i) + \varepsilon_i^0,$$
  $$Y_i(1) = f_0(\mathbf{X}_i) + \tau(\mathbf{X}_i) + \varepsilon_i^1,$$

  with noise terms jointly distributed as

  $$(\varepsilon_i^0, \varepsilon_i^1) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 2\rho \\ 2\rho & 4 \end{bmatrix} \right).$$

  The treatment effect function $\tau(\mathbf{x}) = \tau(x_1, x_2)$ is a smooth random polynomial depending on the first two covariates (or only on x_1 when $d = 1$), generated using a Perlin noise generator (Perlin, 1985) following Bodik and Chavez-Demoulin (2025). The baseline function is $f_0(x) = \beta^\top x$ with β drawn from a standard normal distribution.

- **IHDP (semi-synthetic):** Originally introduced in Hill (2011), this dataset contains 25 pretreatment covariates (e.g., birth weight, maternal age, education level) denoted by $\mathbf{X}$. The binary treatment T indicates whether the infant participated in the intervention program. Potential outcomes represent cognitive test scores, were simulated in Hill (2011) as

  $$Y_i(0) = f_0(X_i) + \varepsilon_i^0, \tag{8}$$
  $$Y_i(1) = f_1(X_i) + \varepsilon_i^1, \tag{9}$$

  where $\varepsilon_i^0, \varepsilon_i^1 \overset{\text{i.i.d.}}{\sim} N(0, 1)$. The functions f_0 and f_1 are either random linear (case "A") or nonlinear (case "B"). We only consider case "B".

  While the original setup fixes $\rho = 0$, we also consider a correlated noise version:

  $$\begin{pmatrix} \varepsilon_i^0 \\ \varepsilon_i^1 \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

  which better reflects empirical situations in which the two potential outcomes are not independent but share substantial underlying information.

- **Twins (real-world):** We use the U.S. twin birth records (1989–1991) described in Louizos et al. (2017), restricted to same-sex twins with both birth weights below $2\,\mathrm{kg}$. Each pair comes with detailed perinatal covariates, including maternal risk factors, prenatal care indicators, and demographic information. In this context, twins are viewed as natural counterfactuals for one another, so the potential outcomes can be conceptually "observed" by comparing mortality for the heavier twin ($T = 1$) and the lighter twin ($T = 0$) within the same pair. The outcome variable is one-year mortality. In our analysis, we work with a balanced sample containing a moderate number of individuals and a small set of covariates, obtained after standard preprocessing.

C.2. Interval Scores Results: Recommending C_ρ for $\rho \leq 0.5$ and C_ρ^{+CI} for $\rho > 0.5$

Figures 5 and 6 report the Interval Scores (IS) of the competing methods across all datasets considered in our experiments. The Interval Score jointly evaluates interval width and coverage, with lower values indicating more efficient and reliable prediction intervals. While GANITE is excluded from these comparisons because it does not provide prediction intervals out of the box, one could imagine extending it with Bayesian or conformalized post-processing layers to quantify uncertainty. For instance, sampling-based approaches could be added to its adversarial generator, or conformal calibration could be applied on top of GANITE outputs. However, such adaptations are not standard, and we therefore omit GANITE from the interval score plots.

Results. When using the bias-corrected C_ρ^{+CI} variant, our method achieves consistently strong results, typically outperforming all baselines across datasets. The only exception is when $\rho = 0$, in which case Direct Outcome (DO) estimators attain nearly identical performance. The main drawback of C_ρ^{+CI} lies in its computational cost, since constructing bootstrap confidence intervals is substantially more demanding than computing C_ρ. Moreover, when ρ is large, estimation error in $\hat{\mu}_\rho$ can induce bias, leading to undercoverage and consequently poor Interval Scores. In practice, we therefore recommend using the uncorrected C_ρ intervals when $\rho \leq 0.5$, while for $\rho > 0.5$ the bias-corrected C_ρ^{+CI} intervals are preferable, as they yield the greatest empirical gains. **Recommendation: C_ρ is satisfactory if $\rho \leq 0.5$, and ideally use C_ρ^{+CI} if $\rho > 0.5$.**

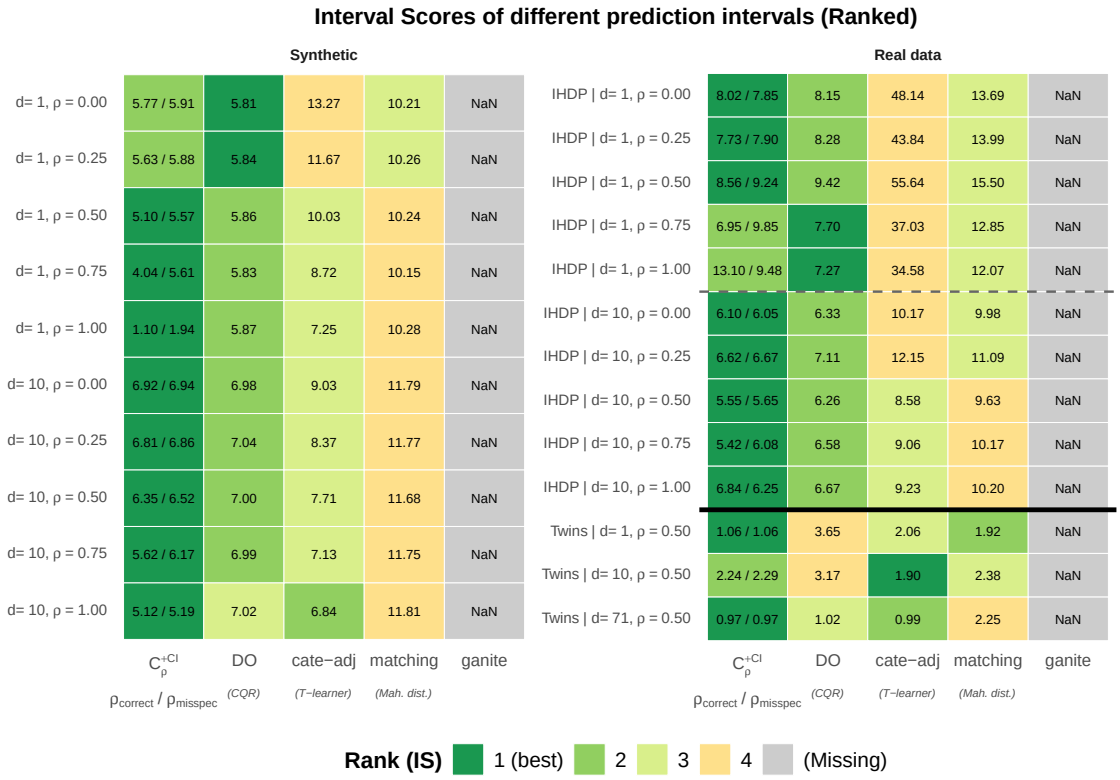

Figure 5: Interval Scores of different prediction interval methods across all datasets. Here, C_ρ^{+CI}, the bias-corrected version of C_ρ introduced in Section 3.4, is used. GANITE is excluded since it does not provide a natural way of constructing prediction intervals.

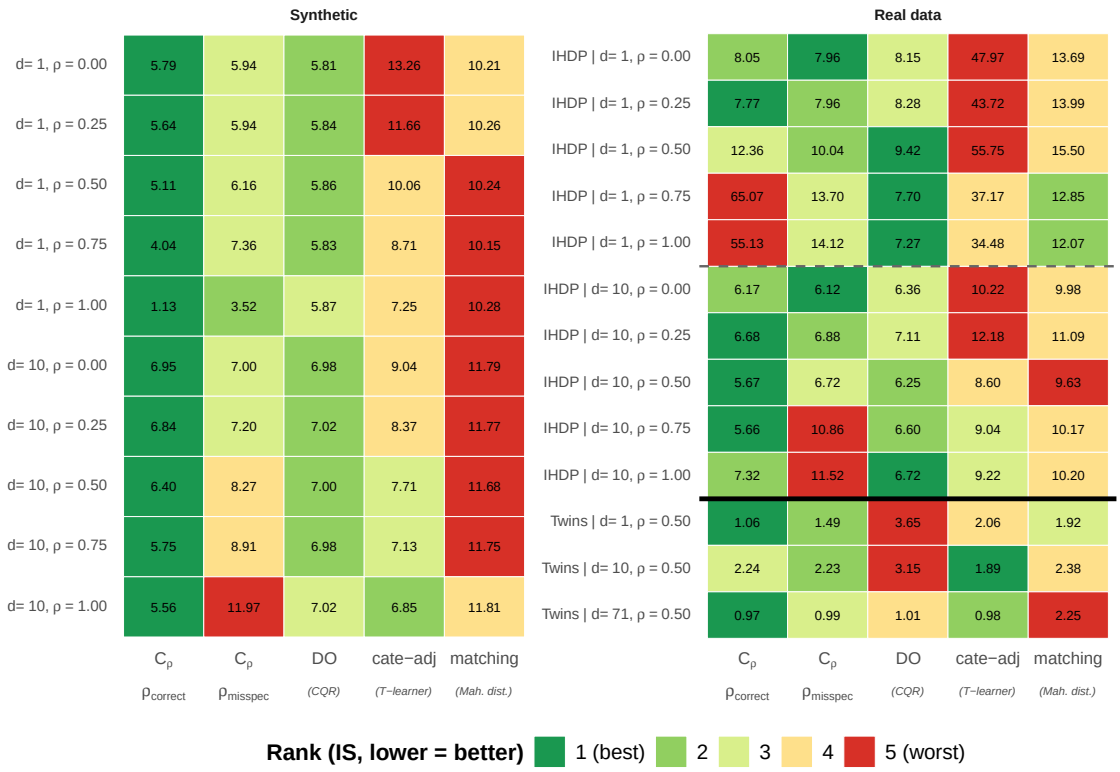

Figure 6: Interval Scores of different prediction interval methods across all datasets. Here, the uncorrected C_ρ intervals, as defined in Section 3, are used.

MSE of Counterfactual Estimators (Ranked)

Synthetic

	RCP(ρ) $\rho_{correct}$	RCP(ρ) $\rho_{misspec}$	DO *(CQR)*	cate−adj *(T−learner)*	matching *(Mah. dist.)*	ganite
d= 1, ρ = 0.00	2.14±0.02	2.24±0.03	2.85±0.03	3.77±0.03	2.54±0.02	9.95±0.78
d= 1, ρ = 0.25	2.02±0.02	2.16±0.03	2.85±0.03	3.28±0.03	2.56±0.02	10.66±1.05
d= 1, ρ = 0.50	1.63±0.01	1.80±0.03	2.84±0.03	2.77±0.02	2.54±0.02	8.22±0.50
d= 1, ρ = 0.75	0.99±0.01	1.12±0.02	2.84±0.03	2.31±0.02	2.54±0.02	11.88±3.00
d= 1, ρ = 1.00	0.08±0.00	0.14±0.02	2.90±0.03	1.85±0.02	2.58±0.03	9.91±1.01
d= 10, ρ = 0.00	3.04±0.04	3.04±0.05	2.94±0.04	3.61±0.05	4.19±0.08	12.57±1.65
d= 10, ρ = 0.25	2.90±0.04	2.98±0.04	2.98±0.04	3.23±0.04	4.20±0.07	12.12±1.48
d= 10, ρ = 0.50	2.56±0.04	2.64±0.05	2.96±0.04	2.79±0.05	4.18±0.08	11.10±1.11
d= 10, ρ = 0.75	2.03±0.04	2.10±0.04	2.93±0.04	2.32±0.04	4.21±0.07	11.57±1.13
d= 10, ρ = 1.00	1.39±0.05	1.49±0.05	2.98±0.05	1.97±0.06	4.29±0.08	11.56±1.60

Real data

	RCP(ρ) $\rho_{correct}$	RCP(ρ) $\rho_{misspec}$	DO *(CQR)*	cate−adj *(T−learner)*	matching *(Mah. dist.)*	ganite
IHDP | d= 1, ρ = 0.00	11.87±3.75	11.16±3.69	15.76±5.01	49.43±17.45	14.62±4.64	215.67±52.35
IHDP | d= 1, ρ = 0.25	9.17±2.44	9.49±2.39	13.19±3.41	40.30±11.71	12.53±3.39	224.68±33.16
IHDP | d= 1, ρ = 0.50	19.13±8.25	20.36±8.70	29.42±12.48	104.75±49.13	28.37±12.33	263.58±74.62
IHDP | d= 1, ρ = 0.75	6.37±1.06	6.66±1.11	9.53±1.51	26.54±4.94	9.03±1.46	183.09±26.72
IHDP | d= 1, ρ = 1.00	6.28±1.64	5.96±1.61	9.03±2.07	25.32±7.18	8.53±1.98	150.65±26.08
IHDP | d= 10, ρ = 0.00	3.29±0.54	3.26±0.54	3.20±0.51	5.43±1.06	4.37±0.77	211.20±41.44
IHDP | d= 10, ρ = 0.25	5.41±1.27	5.58±1.29	5.74±1.34	10.37±2.71	7.86±1.93	330.91±62.15
IHDP | d= 10, ρ = 0.50	2.59±0.34	2.65±0.34	2.98±0.39	4.10±0.68	3.53±0.44	192.14±30.46
IHDP | d= 10, ρ = 0.75	2.89±0.40	2.93±0.40	3.43±0.40	4.77±0.77	4.63±0.64	234.34±35.05
IHDP | d= 10, ρ = 1.00	3.24±0.64	3.23±0.64	4.02±0.68	5.77±1.26	4.96±0.86	233.05±42.52
Twins | d= 1, ρ = 0.50	0.10±NA	0.10±NA	0.18±NA	0.10±NA	0.16±NA	0.21±NA
Twins | d= 10, ρ = 0.50	0.12±NA	0.13±NA	0.17±NA	0.10±NA	0.18±NA	0.17±NA
Twins | d= 71, ρ = 0.50	0.10±NA	0.10±NA	0.14±NA	0.10±NA	0.17±NA	0.10±NA

Rank (MSE) 1 (best) 2 3 4 5 (worst)

Figure 7: Extended version of Figure 2, additionally displaying the standard deviations of the MSE estimates within each cell.

Appendix D. Proofs

Lemma 7 *Let $x \in \mathcal{X}$ and suppose $\big(Y(1), Y(0)\big) \mid X = x$ is Gaussian with $\rho = \mathrm{cor}\big(Y(1), Y(0) \mid X = x\big) \in [-1, 1]$.*

Consider the asymptotic scenario: $\hat{\mu}_t(x) = \mu_t(x)$ and suppose that we found conditionally valid prediction intervals:

$$\mathbb{P}\big(Y(t) \leq \hat{\mu}_t(x) + u_t(x) \mid X = x\big) = 0.95, \quad \mathbb{P}\big(Y(t) \geq \hat{\mu}_t(x) - l_t(x) \mid X = x\big) = 0.95, \quad t = 0, 1.$$

Then, C_ρ prediction intervals from Definition 3 are optimal in a sense that it is the smallest set satisfying:

$$\mathbb{P}\big(Y(1) \in C_\rho(X, Y(0)) \mid X = x, Y(0) = y\big) \geq 0.9,$$

for any $y \in \mathbb{R}$. Moreover, $\hat{\mu}_\rho(x, y)$ is the optimal point predictor in the sense that it minimizes the mean squared error:

$$\hat{\mu}_\rho(x, y) = \operatorname*{argmin}_{c \in \mathbb{R}} \mathbb{E}\big[(Y(1) - c)^2 \mid X = x, Y(0) = y\big].$$

Proof We use the following fact:

For a bivariate Gaussian random variables (Z_1, Z_0):

$$\begin{pmatrix} Z_0 \\ Z_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix} \right),$$

it is well known that:

$$Z_1 \mid Z_0 = z \sim \mathcal{N}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_0}(z - \mu_0),\ \sigma_1^2(1 - \rho^2) \right).$$

Moreover, the shortest prediction interval with a given coverage is symmetric around the mean.

First, we introduce some notation:

- Let $c := \Phi^{-1}(0.95) \approx 1.6449$ denote the 0.95 quantile of a standard Gaussian random variable.
- Let $\sigma_t^2(x) := \mathrm{Var}(Y(t) \mid X = x)$ denote the conditional variance.
- $\mu_t(x) + u_t(x) = \mathrm{Quantile}_{0.95}(Y(t) \mid X = x)$.
- Since $Y(t) \mid X = x$ is symmetrical around the mean, we have $l_t(x) = u_t(x)$. Therefore, $u_t(x) = c \cdot \sigma_t(x)$, by the standard form of the quantile function for a Gaussian distribution. Therefore, $\lambda(x) = \frac{\sigma_1(x)}{\sigma_0(x)}$.

Due to Gaussianity assumption, it holds that:

$$Y(1) \mid Y(0) = y, X = x \sim \mathcal{N}\left(\mu_1(x) + \rho\frac{\sigma_1(x)}{\sigma_0(x)}(y - \mu_0(x)),\ (1 - \rho^2)\sigma_1^2(x) \right)$$

which directly gives us

$$\mathbb{P}(Y(1) \leq \mu_1(x) + \rho\frac{\sigma_1(x)}{\sigma_0(x)}(y_0 - \mu_0(x)) + \sqrt{1-\rho^2} \cdot c \cdot \sigma_1(x) \mid X = x, Y(0) = y_0) = 0.95.$$

Using our notation and previously established results, we get

$$\mathbb{P}(Y(1) \leq \hat{\mu}_\rho(x, y_0) + \sqrt{1-\rho^2} \cdot u_1(x) \mid X = x, Y(0) = y_0) = 0.95,$$

and analogously

$$\mathbb{P}(Y(1) \geq \hat{\mu}_\rho(x, y_0) - \sqrt{1-\rho^2} \cdot l_1(x) \mid X = x, Y(0) = y_0) = 0.95.$$

Hence, we proved that

$$\mathbb{P}\big(Y(1) \in C_\rho(X, Y(0)) \mid X = x, Y(0) = y_0\big) = 0.9.$$

The fact that C_ρ prediction interval is the smallest possible interval achieving the desired coverage follows directly from symmetry+continuity of Gaussian variable. The fact that $\hat{\mu}_\rho(x, y)$ is the optimal point predictor follows directly since

$$\hat{\mu}_\rho(x, y) = \mathbb{E}\big[Y(1) \mid X = x, Y(0) = y\big].$$

■

Theorem 4 *Let $x \in \mathcal{X}$ and suppose $\big(Y(1), Y(0)\big) \mid X = x$ is Gaussian with $\rho = \mathrm{cor}\big(Y(1), Y(0) \mid X = x\big) \in [-1, 1]$.*

Let $\hat{\mu}_t(x)$ be consistent estimators of $\mu_t(x)$, and assume the prediction interval widths $l_t(x), u_t(x)$ are asymptotically conditionally valid, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\big(Y(t) \leq \hat{\mu}_t(x) + u_t(x) \mid X = x\big) = 0.95, \qquad \lim_{n\to\infty} \mathbb{P}\big(Y(t) \geq \hat{\mu}_t(x) - l_t(x) \mid X = x\big) = 0.95,$$

for $t = 0, 1$. Then, for any fixed $y \in \mathbb{R}$:

1. *$\hat{\mu}_\rho(x, y)$ is a consistent estimator of the conditional mean,*

$$\hat{\mu}_\rho(x, y) \xrightarrow{p} \mathbb{E}\big[Y(1) \mid X = x, Y(0) = y\big], \quad \text{as } n \to \infty.$$

2. *The C_ρ prediction intervals achieve asymptotic conditional coverage,*

$$\lim_{n\to\infty} \mathbb{P}\big(Y(1) \in C_\rho(X, Y(0)) \mid X = x, Y(0) = y\big) = 0.9.$$

3. *The C_ρ prediction intervals are **asymptotically optimal** in a sense that, in the limit, C_ρ is the smallest set satisfying (4).*

Proof Under the Gaussian assumption, Lemma 7 implies

$$\mathbb{E}\big[Y(1) \mid X = x, Y(0) = y\big] = \mu_1(x) + \rho\,\frac{\sigma_1(x)}{\sigma_0(x)}\big(y - \mu_0(x)\big). \tag{10}$$

By consistency, $\hat{\mu}_t(x) \xrightarrow{p} \mu_t(x)$ for $t = 0, 1$. Moreover, since the upper and lower bounds converge to the 0.95 and 0.05 conditional quantiles of $Y(t) \mid X = x$, their total width satisfies

$$l_t(x) + u_t(x) \xrightarrow{p} \text{Quantile}_{0.95}(Y(t) \mid X = x) - \text{Quantile}_{0.05}(Y(t) \mid X = x) = 2z_{0.95}\sigma_t(x).$$

Thus,

$$\lambda(x) = \frac{l_1(x) + u_1(x)}{l_0(x) + u_0(x)} \xrightarrow{p} \frac{\sigma_1(x)}{\sigma_0(x)}.$$

Substituting into $\hat{\mu}_\rho(x, y)$,

$$\hat{\mu}_\rho(x, y) \xrightarrow{p} \mu_1(x) + \rho\,\frac{\sigma_1(x)}{\sigma_0(x)}\big(y - \mu_0(x)\big),$$

which coincides with (10), proving consistency of the point estimator.

For the prediction interval C_ρ, Lemma 7 further states that, under Gaussianity, $C_\rho(X, Y(0))$ is the minimal set achieving 90% conditional coverage for $Y(1) \mid X = x, Y(0) = y$. Since $l_t(x)$ and $u_t(x)$ converge to their true quantiles, the constructed interval converges to this optimal set. Hence,

$$\lim_{n\to\infty} \mathbb{P}\big(Y(1) \in C_\rho(X, Y(0)) \mid X = x, Y(0) = y\big) = 0.9,$$

and the optimality then follows trivially from Lemma 7. ■

Lemma 2 *Consider two distributions P and P' satisfying strong ignorability and overlap, that share the same marginal laws of $Y(0) \mid X$ and $Y(1) \mid X$ but differ in their cross-world dependence structure; that is, $\rho(x) \neq \rho'(x)$.*

Then P and P' induce the same observable distribution of (X, T, Y), yet their cross-world regressions can differ:

$$\mathbb{E}_P[Y(1) \mid X = x, Y(0) = y] \neq \mathbb{E}_{P'}[Y(1) \mid X = x, Y(0) = y]$$

for some (x, y). Therefore $\mu(x, y)$ is not identified from the observed data and its value depends on the unobserved cross-world correlation $\rho(x)$.

Proof To see that the cross-world regression can differ, consider the following example: Fix x and assume that, under P, the pair $(Y(0), Y(1)) \mid X = x$ is bivariate normal with mean vector $(\mu_0(x), \mu_1(x))$ and covariance matrix

$$\Sigma_{\rho(x)}(x) = \begin{pmatrix} \sigma_0^2(x) & \rho(x)\sigma_0(x)\sigma_1(x) \\ \rho(x)\sigma_0(x)\sigma_1(x) & \sigma_1^2(x) \end{pmatrix},$$

where $\sigma_0(x), \sigma_1(x) > 0$ and $\rho(x) \in (-1, 1)$. Define P' analogously, but with correlation $\rho'(x) \neq \rho(x)$, keeping the same means and variances. Then for both P and P' we have

$$Y(0) \mid X = x \sim \mathcal{N}(\mu_0(x), \sigma_0^2(x)), \qquad Y(1) \mid X = x \sim \mathcal{N}(\mu_1(x), \sigma_1^2(x)),$$

so the marginals $Y(0) \mid X$ and $Y(1) \mid X$ coincide under P and P', and hence the observable distribution of (X, T, Y) is the same.

However, for a bivariate normal vector, the conditional expectation is

$$\mathbb{E}_P[Y(1) \mid X = x, Y(0) = y] = \mu_1(x) + \rho(x)\,\frac{\sigma_1(x)}{\sigma_0(x)}\big(y - \mu_0(x)\big),$$

and similarly

$$\mathbb{E}_{P'}[Y(1) \mid X = x, Y(0) = y] = \mu_1(x) + \rho'(x)\,\frac{\sigma_1(x)}{\sigma_0(x)}\big(y - \mu_0(x)\big).$$

If $\rho(x) \neq \rho'(x)$ and $\sigma_0(x), \sigma_1(x) > 0$, these two quantities differ for all $y \neq \mu_0(x)$. Since $Y(0) \mid X = x$ is non-degenerate and has a continuous density, the event $\{Y(0) \neq \mu_0(x)\}$ has positive probability, so there exists a set of (x, y) with positive probability such that

$$\mathbb{E}_P[Y(1) \mid X = x, Y(0) = y] \neq \mathbb{E}_{P'}[Y(1) \mid X = x, Y(0) = y]\,.$$

Thus P and P' generate the same observed data but different cross-world regressions, which proves that $\mu(x, y)$ is not identified from the distribution of (X, T, Y) and depends on the unobserved cross-world correlation $\rho(x)$. ■

Lemma 8 (Special cases of ρ)

- *If $\rho = 0$ and $\tilde{C}_1(X)$ is marginally valid, then $C_\rho(X, Y(0))$ is also marginally valid:*

$$\mathbb{P}(Y(1) \in \tilde{C}_1(X)) \geq 0.9 \implies \mathbb{P}(Y(1) \in C_\rho(X, Y(0))) \geq 0.9.$$

  If additionally $Y(0) \perp\!\!\!\perp Y(1) \mid X = x$ and $\tilde{C}_1(X)$ is conditionally valid, then $C_\rho(X, Y(0))$ is also conditionally valid; that is,

$$\mathbb{P}(Y(1) \in \tilde{C}_1(X) \mid X = x) \geq 0.9 \implies \mathbb{P}(Y(1) \in C_\rho(X, Y(0)) \mid X = x, Y(0) = y) \geq 0.9,$$

  for any $x \in \mathcal{X}, y \in \mathcal{Y}$.

- *If $\rho = \pm 1$ and $\mu(x, y_0) = \hat{\mu}(x, y_0)$, then*

$$\mathbb{P}(Y(1) \in C_\rho(X, Y(0)) \mid X = x, Y(0) = y) = 1.$$

  If we have confidence intervals satisfying $\mathbb{P}(\mu(x, y_0) \in \hat{\mu}(x, y_0) \pm r(x, y_0)) = 1 - \beta$, then

$$\mathbb{P}(Y(1) \in C_\rho^{+CI}(X, Y(0)) \mid X = x, Y(0) = y) = 1 - \beta.$$

Proof **Case** $\rho = 0$: By definition, $C_\rho(X, Y(0)) = \tilde{C}_1(X)$, so marginal validity is preserved. If $Y(0) \perp\!\!\!\perp Y(1) \mid X$, then conditioning on $Y(0)$ does not affect the validity, hence conditional validity also holds.

Case $\rho = \pm 1$: Perfect (anti-)correlation implies a deterministic linear relationship: for fixed $X = x$, we have

$$Y(1) = a_x + b_x Y(0) \quad \text{for some } a_x, b_x \in \mathbb{R}.$$

Thus,

$$\text{Var}(Y(1) \mid X = x, Y(0) = y) = 0 \quad \Rightarrow \quad \mathbb{P}(Y(1) = \mu(x, y) \mid X = x, Y(0) = y) = 1.$$

If $\mu(x, y) = \hat{\mu}(x, y)$, then $C_\rho(x, y) = \{\mu(x, y)\}$, implying perfect coverage. If instead $\mu(x, y)$ lies in a confidence interval with coverage $1 - \beta$, then

$$\mathbb{P}(Y(1) \in C_\rho^{+CI}(x, y) \mid X = x, Y(0) = y) \geq 1 - \beta.$$

■