Inference for Matching Estimators: A Consistent Variance Estimator under Heteroskedasticity

Xiang Meng* Aaron Smith† Luke Miratrix‡

September 27, 2025

Abstract

Matching estimators are widely used in causal inference, but valid inference based on them remains challenging. Classical results show that matching estimators have non-negligible bias and nonstandard asymptotics, while resampling approaches such as the bootstrap fail to capture their distributional properties. Recent work, most notably Otsu and Rai (2017), developed wild bootstrap procedures that are asymptotically valid, but in practice these methods can exhibit severe undercoverage for a wide range of sample sizes in realistic designs, where the asymptotics do not kick in until the sample sizes are enormous.

In this paper, we refine existing solutions along three fronts. First, we establish a central limit theorem that extends the class of valid matching procedures, including nearest neighbor, radius, caliper, and synthetic-control-based matching, under general heteroskedastic error structures. Our martingale-based proof weakens the regularity conditions required for asymptotic normality. Second, we propose a simple and computationally efficient variance estimator that only requires treated-to-control matching, making it practical in applications with large control pools, with full theoretical justification provided. Third, we demonstrate through extensive simulations that these refinements translate into major improvements in finite-sample

*Corresponding author: xmeng@g.harvard.edu. Postdoctoral Fellow at Dana-Farber Cancer Institute. Work done while completing Ph.D. in the Department of Statistics, Harvard University.

†Associate Professor, Department of Mathematics and Statistics, University of Ottawa.

‡Professor, Harvard Graduate School of Education.

coverage: our method achieves near-nominal coverage rates (94–99%) in designs where state-of-the-art bootstrap methods can under-cover by 20 percentage points, even with thousands of observations.

By combining theoretical justification with strong non-asymptotic performance, our framework provides a practical and reliable solution for inference with matching estimators. An R package implementing our method is available at `https://github.com/jche/scmatch2`.

1 Introduction

Matching and weighting estimators are fundamental tools in causal inference for estimating treatment effects from observational data. These methods enable researchers to draw population-level inferences about treatment effects by comparing treated units with similar control units based on observed covariates (Rosenbaum and Rubin, 1983; Rubin, 1973) or by reweighting observations to achieve covariate balance (Hirano et al., 2003; Imbens, 2004). Valid population inference—the ability to generalize findings beyond the specific sample to the broader population—is crucial for policy decisions and scientific understanding across diverse fields including economics (Dehejia and Wahba, 1999; Heckman et al., 1997), epidemiology (Stuart, 2010), and policy evaluation (Smith and Todd, 2005).

The foundational asymptotic theory for matching was established by Abadie and Imbens (2006), who showed that matching estimators exhibit nonstandard behavior and slower bias decay than other nonparametric methods. This prompted further developments in bias correction (Abadie and Imbens, 2011) and martingale representations for inference (Abadie and Imbens, 2012). At the same time, Abadie and Imbens (2008) demonstrated that the standard bootstrap fails for matching estimators, motivating alternatives such as the wild bootstrap proposed by Otsu and Rai (2017). That procedure is asymptotically valid and represents the current state-of-the-art. However, as we show through simulations, it can produce unreliable inference in finite samples—sometimes missing nominal coverage by 20 percentage points even in moderately large datasets.

There is also a long tradition of practice-oriented guidance. Hill and Reiter (2006) compared interval estimators for one-to-one matching with replacement, noting instability in standard "matched-pairs" formulas. Austin and Cafri (2020) developed sandwich estimators for survival out-

comes under matching with replacement. Bodory et al. (2020) provided a systematic finite-sample comparison, highlighting cases where bootstrap methods perform well and others where they underperform. Closest to our work, Abadie and Spiess (2022) study regression after matching and emphasize the importance of accounting for induced dependence. Their analysis and ours share the insight that valid inference requires constructing variance estimates within matched clusters. The key difference lies in how the error process is proxied: their robust standard errors rely on residuals from a post-matching regression, which delivers consistency only under correct specification of the regression model. By contrast, our estimator is fully model-free, using within-cluster dispersion of control outcomes as error proxies, and therefore remains consistent without requiring correct outcome model specification.

This paper revisits the inference problem for matching with refinements that yield both new theoretical insights and substantial empirical improvements. Our contributions are threefold. First, we establish a central limit theorem for a broad class of matching procedures—including nearest neighbor, radius, caliper, and synthetic-control-based matching—under weak dependence induced by matching. Our martingale-based proof generalizes and strengthens earlier results by introducing novel regularity conditions that expand the class of procedures known to be valid. Second, we decompose the asymptotic variance into two interpretable components: sampling variability from residual outcome noise and population variability from treatment effect heterogeneity. This decomposition clarifies the distinct roles of different uncertainty sources in driving estimator variability. Third, we propose a computationally simple and theoretically justified variance estimator that discovers a key covariance correction term capturing the interaction between control weights and error variance heterogeneity. Unlike the classical estimator of Abadie and Imbens (2006), our approach requires only treated-to-control matching, and unlike bootstrap methods, it maintains validity across challenging scenarios. Our simulation evidence demonstrates that these refinements translate to dramatically better finite-sample coverage performance: in the nonlinear setting of Otsu and Rai (2017) with sample sizes up to 5,000, our method maintains 96.3% coverage while the wild bootstrap achieves only 75.2%.

The remainder of this paper is organized as follows. Section 2 establishes the problem setup, introducing our notation, assumptions, and the matching estimator for the average treatment effect on the treated. Section 3 develops the inference framework, analyzes the bias-variance decomposition

of the matching estimator, establishes conditions for asymptotic normality, and presents our central limit theorem. Section 4 introduces our variance estimator, begins with the derivative control condition that generalizes previous assumptions, develops the estimator under both homogeneous and heterogeneous error structures, and proves its consistency.

The remainder of this paper is organized as follows. Section 2 establishes the problem setup, introducing our notation, assumptions, and the matching estimator for the average treatment effect on the treated. Section 3 develops the inference framework, analyzing the bias-variance decomposition of the matching estimator, establishing conditions for asymptotic normality, and presenting our central limit theorem. Section 4 introduces our variance estimator, begins with the derivative control condition that generalizes previous assumptions, develops the estimator under both homogeneous and heterogeneous error structures, and proves its consistency. We also compare our approach with existing methods in the literature. Section 5 presents simulation evidence showing that our method maintains proper coverage while the wild bootstrap fails under control unit reuse, using both the challenging nonlinear setting of Otsu and Rai (2017) and the multi-dimensional design of Che et al. (2024). Section 6 applies our method to evaluate an education program in Brazil, illustrating how proper variance estimation affects substantive conclusions in practice. Section 7 concludes with a discussion of extensions and future research directions. Technical proofs and additional results appear in the appendix.

2 Problem Setup and Overview of Main Results

We consider a setting with n observations, each representing a unit in our study population. The sample consists of n_T treated units and n_C control units, with $n = n_T + n_C$.

For each unit i, we observe a tuple $\{Z_i, Y_i, \mathbf{X}_i\}$ where:

- $Z_i \in \{0, 1\}$ denotes its binary treatment status.
- $Y_i \in \mathbb{R}$ denotes its observed real-valued outcome.
- $\mathbf{X}_i \equiv \{X_{1i}, \ldots, X_{ki}\}^T \in \mathbb{R}^k$ denotes its k-dimensional real-valued covariate vector.

We adopt the potential outcomes framework where each unit has two potential outcomes: $Y_i(1)$ and $Y_i(0)$. Here, $Y_i(1)$ represents the outcome if unit i receives treatment, and $Y_i(0)$ represents the

outcome if unit i does not receive treatment. The fundamental problem of causal inference is that we only observe one of these potential outcomes for each unit. Specifically, the observed outcome for unit i is $Y_i \equiv (1-Z_i)Y_i(0)+Z_iY_i(1)$ under the stable unit treatment value assumption (SUTVA).

We assume the data consist of i.i.d. draws of tuples $(Y_i(0), Y_i(1), Z_i, \mathbf{X}_i)$ from a common distribution that does not depend on the sample size n. For each unit i, the generic random variables $(Y(0), Y(1), Z, \mathbf{X})$ represent the population distribution from which the observed data are drawn. Throughout the paper, indexed variables (e.g., $\mathbf{X}_i$) refer to specific observations, while non-indexed variables (e.g., $\mathbf{X}$) refer to the generic random variables representing the population distribution.

We further assume a model where potential outcomes are generated as:

$$\begin{aligned} Y_i(0) &= f_0(\mathbf{X}_i) + \epsilon_{0,i} \\ Y_i(1) &= f_1(\mathbf{X}_i) + \epsilon_{1,i}, \end{aligned}$$

where $f_z(\mathbf{x}) = E[Y(z) \mid \mathbf{X} = \mathbf{x}]$ for $z \in \{0, 1\}$ denote the response surfaces (Hahn et al., 2020; Hill, 2011) under treatment and control. The error terms $\epsilon_{0,i}$ and $\epsilon_{1,i}$ represent the deviations of the individual potential outcomes from their respective conditional expectations, with conditional variances $\sigma^2_{0,i}$ and $\sigma^2_{1,i}$ respectively. Further distributional assumptions about these error terms are detailed in Section 3.2.

Our estimand of interest is the average treatment effect on the treated (ATT):

$$\tau \;=\; E\big[f_1(X_i) - f_0(X_i) \mid Z_i = 1\big].$$

We can further decompose the individual treatment effect as

$$Y_i(1) - Y_i(0) = \tau(\mathbf{X}_i) + \big(\epsilon_{1,i} - \epsilon_{0,i}\big),$$

where $\tau(\mathbf{X}_i) = f_1(\mathbf{X}_i) - f_0(\mathbf{X}_i)$ captures the systematic component of treatment effect variation explained by covariates, while the residual term $(\epsilon_{1,i} - \epsilon_{0,i})$ represents idiosyncratic noise. This distinction between systematic and idiosyncratic variation will later play a central role in our variance decomposition.

2.1 Matching Estimator

We write the set of all treated units' indices as $\mathcal{T} = \{i : Z_i = 1\}$, the set of all control units' indices as $\mathcal{C} = \{i : Z_i = 0\}$, and $t \in \mathcal{T}$, $j \in \mathcal{C}$ as individual treated and control units respectively. For each treated unit $t \in \mathcal{T}$, let $\mathcal{C}_t \subseteq \mathcal{C}$ denote an arbitrary set of control units assigned as its matches; the collection $\{\mathcal{C}_t : t \in \mathcal{T}\}$ is then called a matching. Finally, we denote the size of a set $\mathcal{S}$ as $|\mathcal{S}|$. The matching estimator of the ATT takes the form

$$\hat{\tau}(w) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \big(Y_t - \sum_{j \in \mathcal{C}_t} w_{jt} Y_j\big), \tag{1}$$

where $w_{jt} \in [0, 1]$ is the weight assigned to the matched control unit j for treated unit t, with $\sum_{j \in \mathcal{C}_t} w_{jt} = 1$ for each $t \in \mathcal{T}$. This formulation encompasses many common procedures: for instance, in M-nearest neighbor matching (Rubin, 1973; Abadie and Imbens, 2006; Stuart, 2010), each $\mathcal{C}_t$ consists of the M nearest controls to t with equal weights $w_{jt} = 1/M$, while in synthetic-control-style matching Che et al. (2024), w_{jt} is chosen by solving an optimization problem to approximate $\mathbf{X}_t$ by a convex combination of $\{\mathbf{X}_j : j \in \mathcal{C}_t\}$.

2.2 Variance Estimator (Proposed)

A central difficulty in inference for $\hat{\tau}(w)$ is variance estimation. Classical bootstrap methods fail because they do not capture the complex reuse of controls, while analytic estimators such as Abadie and Imbens (2006) require matching within both treatment groups, which is computationally heavy and rarely used in practice.

We propose a variance estimator that is both simple and practical. Our construction focuses first on the measurement error component of the variance, which arises from the noise terms $\epsilon_{1,i}$ and $\epsilon_{0,i}$. Formally, this component takes the form

$$V_E \;=\; \frac{1}{n_T^2} \left(\sum_{t \in \mathcal{T}} \sigma_{1,t}^2 \;+\; \sum_{j \in \mathcal{C}} (w_j)^2 \sigma_{0,j}^2 \right).$$

Here $\sigma_{z,t}^2 = Var(\epsilon_{z,t} \mid \mathbf{X}_t)$ denotes the conditional variance of the treated and control potential outcome for treated unit t, $\sigma_{0,j}^2 = Var(\epsilon_{0,j} \mid \mathbf{X}_j)$ is the corresponding conditional variance of the

control potential outcome for control unit j, and $w_j = \sum_{t \in \mathcal{T}} w_{jt}$ is the total weight assigned to control unit j across all treated units in the matching.

This measurement error component represents the fundamental randomness in outcomes even after systematic covariate adjustment. Ignoring it leads to severe underestimation of uncertainty. Later (Section 3), we show how this component combines with an additional population heterogeneity component to yield the total variance of $\hat{\tau}(w)$.

To estimate the measurement error component, we use within-cluster variation of control outcomes as model-free proxies for error variance. For each treated unit t, define

$$s_t^2 = \frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} (Y_j - \bar{Y}_t)^2, \qquad \bar{Y}_t = \frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} Y_j.$$

Our plug-in estimator is then

$$\hat{V}_E = \frac{1}{n_T^2} \left(\sum_{t \in \mathcal{T}} s_t^2 \; + \; \sum_{j \in \mathcal{C}} (w_j)^2 s_j^2 \right). \tag{2}$$

Here the first term aggregates within-cluster variances across treated units, while the second term adjusts for the reuse of controls: heavily reused controls with large w_j contribute disproportionately to the variance of $\hat{\tau}(w)$.

This estimator has two attractive features: (i) it is computationally efficient, requiring only treated-to-control matching; (ii) it is model-free, relying only on empirical dispersion rather than regression residuals

In Section 4, we show that $\hat{V}_E$ consistently estimates the measurement error component, and extend it to the full variance estimator $\hat{V}$ that also incorporates treatment effect heterogeneity. We now turn from variance estimation to the broader inference problem: what conditions are required for $\hat{\tau}(w)$ to be asymptotically normal, and how the different variance components together determine its limiting distribution.

3 The Inference Problem

We now turn to inference for $\hat{\tau}(w)$. The goals of this section are threefold: (i) introduce the assumptions needed for asymptotic analysis, (ii) establish a bias–variance decomposition that clarifies the roles of systematic bias, sampling error, and population heterogeneity, and (iii) present our central limit theorem that delivers a variance formula V. In the following section, we will then show how to consistently estimate V in practice.

To construct valid confidence intervals for our matching estimator $\hat{\tau}$, we require asymptotic normality of the form:

$$\frac{\sqrt{n_T}\,(\hat{\tau} - \tau)}{V^{-1/2}} \xrightarrow{d} N(0, 1).$$

The difference between the matching estimator $\hat{\tau}(w)$ (defined in Equation (1)) and the estimand τ can be decomposed into three components:

$$\hat{\tau}(w) - \tau = \hat{\tau}(w) - \tau_{\text{SATT}} + \tau_{\text{SATT}} - \tau = B_n + E_n + P_n \tag{3}$$

where τ_{SATT} is the sample average treatment effect on the treated (SATT):

$$\tau_{\text{SATT}} = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \big(f_1(X_t) - f_0(X_t)\big).$$

$$B_n = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{C}_t} w_{jt} \big(f_0(X_t) - f_0(X_j)\big)$$

represents bias from imperfect covariate matching.

$$E_n = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \Big(\epsilon_t - \sum_{j \in \mathcal{C}_t} w_{jt} \epsilon_j\Big)$$
$$= \frac{1}{n_T} \sum_{t \in \mathcal{T}} \epsilon_t - \frac{1}{n_T} \sum_{j \in \mathcal{C}} w_j \epsilon_j$$

captures measurement error from random variation in unobserved factors.

$$P_n = \tau_{\text{SATT}} - \tau$$

measures representation error between sample and population treatment effects.

where $w_j = \sum_{t \in \mathcal{T}} w_{jt}$ is the total weight assigned to control unit j across all matched treated units.

3.1 Assumptions

To proceed, we require a set of conditions on the covariates, treatment assignment, and matching procedure.

Assumption 1 (Compact support)**.** *The covariate vector* $\mathbf{X}$ *is a* k*-dimensional random vector with a density with respect to Lebesgue measure on* $\mathbb{R}^k$ *with compact support* $\mathbb{X}$*. The density of* X *is bounded and bounded away from zero on its support.*

The compact support assumption helps ensure that the covariate space is well-behaved, which facilitates consistent estimation and rules out pathological cases where the distribution of covariates becomes too sparse or unbounded.

Assumption 2 (Unconfoundedness and overlap (Rubin, 1974))**.** *For almost every* $x \in \mathbb{X}$ *there exists* $\eta > 0$ *such that*

1. $(Y(1), Y(0)) \perp\!\!\!\perp Z \mid \mathbf{X}$,
2. $\eta < \Pr(Z = 1 \mid \mathbf{X} = x) < 1 - \eta$.

This assumption states that, conditional on the observed covariates, treatment assignment is independent of the potential outcomes, and that both treated and control units are sufficiently represented across the covariate space. By the law of large numbers, it follows that $n_T/n \to \Pr(Z = 1)$ and $n_C/n \to 1 - \Pr(Z = 1)$ almost surely, and hence $n_T/n_C \to \theta$ for some $\theta \in \left(\frac{\eta}{1-\eta}, \frac{1-\eta}{\eta}\right)$.

Define the matching radius for a treated unit t with covariate value $\mathbf{X}_t$ as:

$$r\left(\mathcal{C}_t\right) = \sup_{j \in \mathcal{C}_t} \left\|\mathbf{X}_t - \mathbf{X}_j\right\|.$$

This radius represents the maximum distance between a treated unit and any of its matched controls. The probabilistic properties of this radius will be crucial for establishing our theoretical results.

Assumption 3 (Exponential Tail Condition)**.** *Let $r(\mathcal{C}_t)$ denote the (random) radius used for treated unit t, i.e., the maximum ℓ_2-distance from X_t to the controls in $\mathcal{C}_t$. There exist constants $C_1 \geq 1$ and $C_2 > 0$, not depending on n, such that for all $u \geq 0$ and all treated t,*

$$\Pr\big(n_C^{1/k}\, r(\mathcal{C}_t) > u\big) \;\leq\; C_1 \exp\big(-C_2\, u^k\big).$$

This assumption requires that the probability of a large scaled radius decays at a Weibull-k rate, $\exp(-cu^k)$. The shape parameter k reflects the covariate dimension, so higher k implies faster decay. Intuitively, this ensures increasingly accurate matches as n grows. Equivalently, $P\big(n_C r(\mathcal{C}_t)^k > t\big) \leq C_1 e^{-C_2 t}$, where $r(\mathcal{C}_t)^k$ approximates the volume of the matched region. Abadie and Imbens (2006) show that the number of times a control is reused, $K(j)$, is of order $n_C r(\mathcal{C}_t)^k$, so bounding this volume stabilizes reuse and underpins the CLT.

Many matching methods satisfy this condition. For fixed M-nearest neighbor matching, Abadie and Imbens (2006, proof of Lemma 3, p. 262) show it holds when covariates have bounded overlapping density, since the matching radius shrinks predictably with n. Radius matching with a data-adaptive caliper, such as the M-th nearest neighbor distance, also yields the required Weibull-k bound. In practice, researchers often choose the caliper by inspecting nearest-neighbor distance histograms (Che et al., 2024), which balances coverage and radius size.

Remark on bias. A crucial challenge in matching is that B_n shrinks at the slow rate $O_p(n_T^{-1/k})$ (Abadie and Imbens, 2006), slower than the $n_T^{-1/2}$ rate of conventional CLTs. Because our focus is on variance estimation, we take this fact as given and refer readers to Abadie and Imbens (2011) for explicit bias-corrected estimators.

3.2 Error Variance Assumptions

To analyze the large-sample behavior of $\hat{V}$, we also require structure on the conditional error variances. Importantly, we impose only moment bounds; Gaussianity of the errors is not required.

Let us denote the conditional variances of the potential outcomes as:

$$\begin{aligned}
\sigma^2_{0,i} \;&=\; \sigma^2_0(X_i) = E\big[\big(Y_i(0) \;-\; f_0(\mathbf{X}_i)\big)^2 \,\big|\, \mathbf{X}_i\big] = E[\epsilon^2_{0,i} \,|\, \mathbf{X}_i], \\
\sigma^2_{1,i} \;&=\; \sigma^2_1(X_i) = E\big[\big(Y_i(1) \;-\; f_1(\mathbf{X}_i)\big)^2 \,\big|\, \mathbf{X}_i\big] = E[\epsilon^2_{1,i} \,|\, \mathbf{X}_i].
\end{aligned} \tag{4}$$

We now define a class of variance functions with properties that enable consistent estimation in the matched setting.

Definition 3.1 (Regular variance function)**.** *A function $\sigma^2 : \mathcal{X} \to \mathbb{R}_+$ is said to be a regular variance function if it satisfies the following:*

- ***Uniform continuity.*** *$\sigma^2(\cdot)$ is uniformly continuous (or Lipschitz) on the support $\mathcal{X} \subset \mathbb{R}^d$ of X.*
- ***Boundedness.*** *There exist constants $0 < \sigma^2_{\min} < \sigma^2_{\max} < \infty$ such that*

$$\sigma^2_{\min} \leq \sigma^2(x) \leq \sigma^2_{\max} \quad \textit{for all } x \in \mathcal{X}.$$

- ***Higher-order moment bound.*** *There exists a constant $C < \infty$ and an exponent $\delta > 0$ such that*

$$\sup_{x \in \mathcal{X}} \mathbb{E}\big[\big|\epsilon_i\big|^{2+\delta} \;\big|\; X_i = x\big] \;\leq\; C.$$

  Here ϵ_i generically denotes either $\epsilon_{0,i}$ or $\epsilon_{1,i}$.

The first condition ensures that matched units have similar variances. Specifically, for any matching scheme with $\|X_{tj} - X_t\| \to 0$ (as guaranteed by Assumption 3), we have $\sigma^2(X_{tj}) \to \sigma^2(X_t)$. Hence, $\sigma^2_j \approx \sigma^2_t$ for $j \in \mathcal{C}_t$ whenever $\mathcal{C}_t$ is constructed by matching on X. In particular,

$$\max_{j \in \mathcal{C}_t} \big|\sigma^2(X_{tj}) \;-\; \sigma^2(X_t)\big| \;\to 0,$$

provided that $\max_{j \in \mathcal{C}_t} \|X_{tj} - X_t\| \to 0$. Definition 3.1 generalizes Assumption 4.1 in Abadie and Imbens (2006), which assumes Lipschitz continuity.

The boundedness condition ensures that the conditional variance is bounded away from both zero and infinity, preventing degeneracy and controlling the influence of outliers. The third condition imposes a uniform bound on a higher-order conditional moment of the errors. This assumption is standard in high-dimensional estimation and facilitates the use of maximal inequalities and uniform convergence tools.

We now formally state the assumption we make on the conditional variances of the potential outcomes:

Assumption 4 (Regular error variances)**.** *Both $\sigma_0^2(x)$ and $\sigma_1^2(x)$ are regular variance functions.*

3.3 Decomposition of Asymptotic Variance Components

From Equation (3), $\hat{\tau}(w) - \tau$ decomposes into bias B_n, measurement error E_n, and heterogeneity P_n. The latter two terms drive the asymptotic variance once bias is subtracted.

Measurement Error Component V_E. We first consider the component due to residual outcome noise. Conditional on the covariates $\mathbf{X}$ and treatment assignment vector $\mathbf{Z}$, the variance of E_n is given by:

$$\begin{aligned} V_E :=& \mathbb{E}[E_n^2 \mid \mathbf{X}, \mathbf{Z}] \\ =& \frac{1}{n_T^2} \left(\sum_{t \in \mathcal{T}} \sigma_{1,t}^2 + \sum_{j \in \mathcal{C}} (w_j)^2 \sigma_{0,j}^2 \right), \end{aligned} \tag{5}$$

where $w_j = \sum_{t \in \mathcal{T}} w_{jt}$ is the total weight assigned to control unit j across all matched treated units.

The first term reflects the direct contribution of treated units through their outcome variances, while the second term captures how control units contribute via squared weight accumulation. Notably, reused controls (with large w_j) disproportionately affect the overall variance, creating a fundamental bias-variance tradeoff in matching. Prior work including Kallus (2020) and Che et al. (2024) leverage this variance structure to study how tighter matches (which reduce bias) can increase variance due to heavy reuse of control units.

Population Heterogeneity Component V_P. The second term $P_n = \tau_{\text{SATT}} - \tau$ captures how the realized sample of treated units may differ from the target population of treated units. That is, even if outcomes were observed without error, the sample ATT may deviate from the population ATT due to treatment effect heterogeneity.

We define the population heterogeneity component as the asymptotic variance of P_n:

$$V_P := \frac{1}{n_T} \mathbb{E}\left[(\tau(X_i) - \tau)^2 \mid Z_i = 1 \right]. \tag{6}$$

We can verify that P_n has asymptotic variance V_P as defined above. This depends only on the dispersion of treatment effects among treated units, and vanishes under homogeneous treatment

effects.

3.4 The Central Limit Theorem

We now present our main asymptotic normality result, which forms the basis for valid inference.

Theorem 3.2 (Central Limit Theorem)**.** *Under Assumptions 1, 2, 3 and 4, as $n_T \to \infty$:*

$$\frac{\sqrt{n_T}\left(\hat{\tau} - B_n - \tau\right)}{V^{-1/2}} \quad \xrightarrow{d} \quad N(0,1),$$

where

$$V = n_T \cdot (V_E + V_P).$$

When $k \leq 2$, the bias term B_n shrinks faster and can be ignored, yielding the same CLT without bias correction.

Proof: See Appendix A.

This theorem extends the seminal results of Abadie and Imbens (2006) by covering a broader class of matching estimators. In particular, it applies to procedures beyond fixed M-nearest neighbor with uniform weights, including radius matching, caliper matching, and synthetic-control–style weights. Our framework accommodates a wide range of weighting schemes and clarifies the variance decomposition in terms of measurement error and treatment effect heterogeneity.

Our proof approach builds on the martingale representation of Abadie and Imbens (2012) but refines it by incorporating the drift term required for a valid martingale CLT. This refinement allows us to handle the dependence created by control reuse more directly and to establish asymptotic normality under weaker regularity conditions. Together, these results provide a more general theoretical foundation for inference with modern matching methods.

For practical inference, we therefore need a consistent estimator $\hat{V}$ of V, which is the focus of Section 4.

4 The Standard Error Estimator

In Section 3.4, we established a CLT for the matching estimator $\hat{\tau}(w)$ with asymptotic variance $V = n_T(V_E + V_P)$. To use this result in practice, we need a consistent estimator of V. Because V decomposes into the measurement error component V_E and the heterogeneity component V_P, our strategy is to begin with V_E, which presents the main technical challenge, and then extend the estimator to cover V_P.

4.1 Consistent Estimation of V_E

We now turn to the construction of an estimator for V_E. We first state the assumptions that make estimation feasible, then introduce two estimators. The first is a straightforward plug-in estimator based on cluster residuals. The second rewrites the estimator into a pooled-variance form motivated by homoskedasticity (a t-test–like variance) with an additional covariance adjustment that captures heteroskedasticity. We show that both estimators are asymptotically equivalent and consistent.

Assumption 5 (Smoothness of outcome regression)**.** *The regression function f is continuously differentiable on the compact support $\mathbb{X}$ of $\mathbf{X}$.*

This mild smoothness assumption ensures that f' is bounded on $\mathbb{X}$, which will be used in bounding approximation errors within matched clusters.

Assumption 6 (Estimable Treatment Variance)**.** *We assume that the conditional variances of the potential outcomes are related in a way that allows estimation from control unit residuals:*

$$\sigma^2_{0,i} = \sigma^2_{1,i} = \sigma^2_i.$$

Furthermore, we assume σ^2_i is regular in the sense of Definition 3.1.

This assumption underpins the entire estimation strategy in this section. It allows us to estimate the unobservable treated variances $\sigma^2_{1,i}$ using the corresponding control variances, which can be recovered from matched control outcomes.

4.1.1 Estimator 1: Direct Plug-in

Recall from Equation 5 that the measurement error variance is given by:

$$\begin{aligned} V_E &= \mathbb{E}[E_n^2 \mid \mathbf{X}, \mathbf{Z}] \\ &= \frac{1}{n_T^2}\left(\sum_{t\in\mathcal{T}} \sigma_{1,t}^2 + \sum_{j\in\mathcal{C}} (w_j)^2 \sigma_{0,j}^2\right) \\ &= \frac{1}{n_T^2}\left(\sum_{t\in\mathcal{T}} \sigma_t^2 + \sum_{j\in\mathcal{C}} (w_j)^2 \sigma_j^2\right), \end{aligned}$$

where the last equality is due to Assumption 6, and $w_j = \sum_{t\in\mathcal{T}} w_{jt}$ is the total weight assigned to control unit j across all matched treated units.

A natural approach is to estimate the individual variances σ_t^2 and σ_j^2 using cluster-based residual variance estimates. For each matched cluster consisting of treated unit t and its matched controls $\mathcal{C}_t$, we define a cluster as the set $\{t\}\cup\mathcal{C}_t$. For control units j that belong to multiple clusters (i.e., are reused across different treated units), we allow such overlap and assign j to one cluster arbitrarily for the purpose of defining s_j^2.[1] The residual variance for cluster t is defined to be:

$$s_t^2 = \frac{1}{|\mathcal{C}_t| - 1} \sum_{j\in\mathcal{C}_t} \left(Y_j - \bar{Y}_t\right)^2, \quad \text{where} \quad \bar{Y}_t = \frac{1}{|\mathcal{C}_t|} \sum_{j\in\mathcal{C}_t} Y_j. \tag{7}$$

This approach uses only control outcomes because we use the difference between individual control outcomes and their cluster mean as the error proxy. Importantly, this cluster mean $\bar{Y}_t$ is not obtained from any parametric model but rather from the empirical average within the matched set, enabling model-free variance estimation. This is a key advantage of our approach: we do not need to specify or estimate outcome regression models to obtain variance estimates.

Based on this cluster-based variance estimation, our general estimator for V_E is:

$$\hat{V}_E = \frac{1}{n_T^2}\left(\sum_{t\in\mathcal{T}} s_t^2 + \sum_{j\in\mathcal{C}} (w_j)^2 s_j^2\right). \tag{8}$$

[1] Overlap does not affect the asymptotic theory, since the contribution of each j is accounted for via its total weight $w_j = \sum_{t\in\mathcal{T}} w_{jt}$.

We show it is consistent with the theorem below.

Theorem 4.1 (Consistency of the General Variance Estimator)**.** *Under Assumptions 3, 4, and 5, the estimator in Equation* (8) *satisfies*

$$n_T \left| \hat{V}_E - V_E \right| \xrightarrow{p} 0 \qquad \text{as } n_T \to \infty.$$

Proof: see Appendix B

The key intuition is the "power of averaging": each individual cluster variance estimates are noisy, but the aggregation across many clusters smooths out individual noises, in the same spirit as White's heteroskedasticity-consistent estimator (White, 1980).

4.1.2 Estimator 2: Pooled-variance Form with Covariance Adjustment

We next present an alternative estimator that highlights the structure of V_E. It combines a pooled-variance component, motivated by the homoskedastic benchmark, with an adjustment term that corrects for heterogeneity via a covariance form. Specifically, we define

$$\begin{aligned} \hat{V}_E^{alt} = &\left(S^2 - \frac{1}{\sum_{t\in\mathcal{T}} |\mathcal{C}_t|/n_T} \operatorname{Cov}_v\big(|\mathcal{C}_t|, \sigma_t^2\big) \right) \left(\frac{1}{n_T} + \frac{1}{\operatorname{ESS}(\mathcal{C})} \right) \\ &+ \frac{1}{n_T} \operatorname{Cov}_p\big(w_j, s_j^2\big) , \end{aligned} \tag{9}$$

where S^2 is the pooled variance in Equation (13), Cov_v denotes covariance under the uniform distribution on treated units, and Cov_p denotes covariance under the normalized weights $p_j = w_j/n_T$ on controls (a random measure).

This formulation makes clear that $\hat{V}_E^{alt}$ consists of a t-test–like variance component plus a heteroskedasticity correction. We now explain how it arises.

Motivation from the homoskedastic benchmark. If $\sigma^2(x) \equiv \sigma^2$, then

$$\begin{aligned} V_E &= \frac{1}{n_T^2} \left(\sum_{t\in\mathcal{T}} \sigma^2 + \sum_{j\in\mathcal{C}} (w_j)^2 \sigma^2 \right) \\ &= \sigma^2 \left(\frac{1}{n_T} + \frac{1}{\operatorname{ESS}(\mathcal{C})} \right), \end{aligned} \tag{10}$$

where $\mathrm{ESS}(\mathcal{C})$ is the effective sample size of the weighted control sample:

$$\mathrm{ESS}(\mathcal{C}) = \frac{(\sum_{j\in\mathcal{C}} w_j)^2}{\sum_{j\in\mathcal{C}} w_j^2}. \tag{11}$$

This metric quantifies the number of independent observations that would provide equivalent precision under equal weighting (Potthoff et al., 2024), and reflects efficiency loss from reusing controls with varying weights.

This motivates the plug-in form:

$$\hat{V}_E^{homo} = S^2 \left(\frac{1}{n_T} + \frac{1}{\mathrm{ESS}(\mathcal{C})} \right), \tag{12}$$

where S^2 is a pooled variance estimator for σ^2 defined across matched clusters where each treated unit is matched to more than one control (excluding singleton control matches). Specifically:

$$S^2 = \frac{1}{N_C} \sum_{t\in\mathcal{T}_+} |\mathcal{C}_t| s_t^2 \quad \text{with} \quad N_C = \sum_{t\in\mathcal{T}_+} |\mathcal{C}_t|, \tag{13}$$

where $\mathcal{T}_+ = \{t \in \mathcal{T} : |\mathcal{C}_t| > 1\}$ excludes singleton clusters, since variance cannot be estimated from clusters with only one control unit.

Lemma 4.2 shows that S^2 is consistent for the average treated variance up to a correction term that depends on the covariance between cluster size $|\mathcal{C}_t|$ and σ_t^2.

Lemma 4.2 (Consistency of the Pooled Variance Estimator)**.** *Let $\{\mathcal{C}_t, t \in \mathcal{T}\}$ be a collection of matched control sets. Under Assumptions 3, 4, and 5, as $n_T \to \infty$:*

$$\left| S^2 - \frac{1}{\sum_{t\in\mathcal{T}} |\mathcal{C}_t|/n_T} \mathrm{Cov}_v \left(|\mathcal{C}_t|, \sigma_t^2 \right) - \frac{1}{n_T} \sum_{t\in\mathcal{T}} \sigma_t^2 \right| \xrightarrow{a.s.} 0. \tag{14}$$

where v denotes the uniform distribution on the treated set $\mathcal{T}$.

Proof: See Appendix C.

This lemma shows that our pooled variance estimator S^2, although originally motivated under homoskedasticity, remains consistent on average even when error variances σ_t^2 are heterogeneous, up to a correction term. The key intuition is the "power of averaging": each local variance estimate s_t^2

converges to its target σ_t^2 within shrinking clusters, and aggregation across many clusters smooths out local noise, in the same spirit as White's heteroskedasticity-consistent estimator (White, 1980).

The correction term $\frac{1}{\sum_{t\in\mathcal{T}}|\mathcal{C}_t|/n_T}\operatorname{Cov}_v\left(|\mathcal{C}_t|,\sigma_t^2\right)$ arises because in S^2, we weight each residual s_j^2 by $\frac{|\mathcal{C}_t|}{N_C}$. This term disappears when $|\mathcal{C}_t|$ is the same for all t (for example, in fixed M-NN matching) or under homoskedasticity where σ_t^2 is constant. When this correction term is non-zero, it is positive when treated units with larger variance σ_t^2 tend to have larger clusters $|\mathcal{C}_t|$ (e.g., in radius/caliper schemes where noisier cases attract larger neighborhoods). The term is negative in the opposite pattern.

Beyond homoskedasticity. Lemma 4.2 leads us to consider the properties of the following quantity:

$$V_{E,\lim}^* := \left(\frac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2\right)\left(\frac{1}{n_T}+\frac{1}{\operatorname{ESS}(\mathcal{C})}\right).$$

It turns out that $V_{E,\lim}^*$ converges in probability to the true variance V_E up to a correction term:

Lemma 4.3 (Asymptotic Equivalence to Error Variance)**.** *Under the same assumptions as Lemma 4.2,*

$$n_T\left|V_{E,\lim}^* + \frac{1}{n_T}\operatorname{Cov}_p\left(w_j,\sigma_j^2\right) - V_E\right| \xrightarrow{p} 0$$

where $\operatorname{Cov}_p\left(w_j,\sigma_j^2\right) = \frac{1}{n_T}\sum_{j\in\mathcal{C}}\left(w_j - \frac{\sum_{j'\in\mathcal{C}}w_{j'}^2}{n_T}\right)w_j\sigma_j^2$, and p is the random probability measure on $\mathcal{C}$ that assigns mass $p_j = w_j/n_T$ to each control unit j.

Proof: See Appendix D.

This lemma shows that our variance formula (Equation 12) resembles the pooled variance structure in a two-sample t-test assuming equal variances, up to an adjustment term, $\frac{1}{n_T}\operatorname{Cov}_p\left(w_j,\sigma_j^2\right)$. The adjustment term appears because $V_{E,\lim}^*$ depends only on the treated unit variances, while the true variance V_E includes the weighted average of control unit variances through the term $\frac{1}{n_T^2}\sum_{j\in\mathcal{C}}w_j^2\sigma_j^2$. The correction term bridges this difference.

The adjustment term is zero when either of two conditions holds: (1) When weights are uniform (as in M-NN matching with no overlapping controls), each control unit j receives weight $w_j = 1/M$ if matched to exactly one treated unit, and the correction term vanishes. (2) When errors are

homogeneous ($\sigma_j^2 = \sigma^2$ for all j), the correction term equals zero. When neither condition holds, the term is positive if higher-variance controls receive more reuse (i.e., $\mathrm{Cov}_p(w, \sigma^2) > 0$), and is negative if more stable controls receive more reuse ($\mathrm{Cov}_p(w, \sigma^2) < 0$).

Putting it together. We now establish consistency.

Theorem 4.4 (Consistency of the Alternative Estimator)**.** *Under Assumptions 3, 4, and 5, the alternative estimator in Equation* (9) *is consistent for* V_E*:*

$$n_T \left| \hat{V}_E^{alt} - V_E \right| \xrightarrow{p} 0 \qquad \text{as } n_T \to \infty.$$

Proof: See Appendix E.

Together, Theorems 4.1 and 4.4 show that both $\hat{V}_E$ and $\hat{V}_E^{alt}$ are consistent, and in fact asymptotically equivalent. The first estimator is the natural direct plug-in; the second highlights the structure of V_E as a t-test–like pooled variance plus a covariance adjustment for heteroskedasticity.

4.2 Consistent Estimation of V

Building on our analysis of the measurement error variance component V_E, we now develop a consistent estimator for the total variance V. While V_E captures the variance due to residual outcome noise, the complete variance V must also account for treatment effect heterogeneity among the treated units.

We start by exploring the relationship between the squared deviations of individual treatment effects and the components of the total variance:

$$\begin{aligned}
&E\left[\left(Y_t(1) - \hat{Y}_t(0) - \tau\right)^2\right] \\
\approx &E\left[(\tau(x) - \tau)^2\right] + E\left[\varepsilon_t^2 + \sum_{j \in C_t} w_{jt}^2 \varepsilon_j^2\right] \\
\approx &n_T V_P + \frac{1}{n_T}\left[\sum_{t \in \mathcal{T}} \sigma_t^2 + \sum_{j \in \mathcal{C}} \left(\sum_{t' \in \mathcal{T}} w_{jt'}^2\right) \sigma_j^2\right].
\end{aligned}$$

This expectation can be approximated empirically as:

$$E\left[\left(Y_t(1) - \hat{Y}_t(0) - \tau\right)^2\right] \approx \frac{1}{n_T}\sum_{t\in\mathcal{T}}\left(Y_t - \hat{Y}_t(0) - \hat{\tau}\right)^2.$$

By comparing the population expression above with the definition of V_P in Equation (6), and replacing population variances by sample analogues, we obtain the following estimator:

$$\begin{aligned}\hat{V}_P :=& \frac{1}{n_T^2}\sum_{t\in\mathcal{T}}\left(Y_t - \hat{Y}_t(0) - \hat{\tau}\right)^2 \\ & - \frac{1}{n_T^2}\left[\sum_{t\in\mathcal{T}}\hat{\sigma}_t^2 + \sum_{j\in\mathcal{C}}\left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\hat{\sigma}_j^2\right]\end{aligned}$$

We now heuristically combine the two components. While this expression includes variance estimates $\hat{\sigma}_t^2$ and $\hat{\sigma}_j^2$, the former terms will cancel out and the latter will ultimately be approximated by S^2 as in our estimator for V_E. One does not need to worry about the precise form of these terms at this stage—they serve to motivate the algebraic derivation below.

Combining this with our estimator for V_E, we obtain:

$$\begin{aligned}\hat{V} =& n_T \cdot (\hat{V}_E + \hat{V}_P) \\ =& \frac{1}{n_T}\left[\sum_{t\in\mathcal{T}}\hat{\sigma}_t^2 + \sum_{j\in\mathcal{C}}\left(\sum_{t'\in\mathcal{T}} w_{jt'}\right)^2\hat{\sigma}_j^2\right] \\ & + \frac{1}{n_T}\sum_{t\in\mathcal{T}}\left(Y_t - \hat{Y}_t(0) - \hat{\tau}\right)^2 \\ & - \frac{1}{n_T}\left[\sum_{t\in\mathcal{T}}\hat{\sigma}_t^2 + \sum_{j\in\mathcal{C}}\left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\hat{\sigma}_j^2\right]\end{aligned}$$

Through algebraic simplification, this expression reduces to:

$$\hat{V} = \frac{1}{n_T}\sum_{t\in\mathcal{T}}\left(Y_t - \hat{Y}_t(0) - \hat{\tau}\right)^2 + \frac{1}{n_T}\left[\sum_{j\in\mathcal{C}} s_j^2\left[\left(\sum_{t'\in\mathcal{T}} w_{jt'}\right)^2 - \left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\right]\right] \tag{15}$$

This estimator effectively combines the empirical squared deviations with a correction term that accounts for the matching structure.

Theorem 4.5 (Consistency of the Total Variance Estimator)**.** *Under Assumptions 3, 4, and 5, the*

proposed estimator $\hat{V}$ is consistent:

$$\left|\hat{V} - V\right| \xrightarrow{p} 0 \quad as\, n_T \to \infty.$$

Proof: See Appendix F

4.3 Comparison with Abadie and Imbens (2006) Estimator

To position our work within the existing literature and highlight its advantages, we now compare our variance estimator with that proposed by Abadie and Imbens (2006). This comparison is particularly relevant as their work established the foundational theory for matching estimators, and our analysis builds upon and extends their approach for practical applications in modern causal inference settings.

Adapting their estimator to our notation:

$$\widehat{V}_{AI06} = \frac{1}{n_T^2} \sum_{t \in \mathcal{T}} \hat{\sigma}_t^2 + \frac{1}{n_T^2} \sum_{j \in \mathcal{C}} \left(\sum_{t \in \mathcal{T}} w_{jt} \right)^2 \hat{\sigma}_j^2, \tag{16}$$

where $w_{jt} = 1/M$ if unit j is among the M closest controls to unit t, and $w_{jt} = 0$ otherwise, and $\hat{\sigma}_i^2$ is an estimate of the conditional outcome variance for unit i, defined as:

$$\hat{\sigma}_i^2 = \frac{M}{M+1} \left(Y_i - \frac{1}{M} \sum_{m=1}^{M} Y_{m(i)} \right)^2.$$

Here, $Y_{m(i)}$ denotes the outcome of the m-th closest unit to unit i among units with the same treatment status, and M is a fixed small number (typically set to match the number of matches used in the estimator).

The fundamental methodological difference lies in variance estimation approaches. Abadie and Imbens (2006) estimates variance by comparing each unit to its nearest same-treatment neighbors individually: $\hat{\sigma}_i^2 = \frac{M}{M+1} \left(Y_i - \frac{1}{M} \sum_{m=1}^{M} Y_{m(i)} \right)^2$. In contrast, our estimator calculates variance within matched clusters: $s_t^2 = \frac{1}{|\mathcal{C}_t|-1} \sum_{j \in \mathcal{C}_t} \left(Y_j - \bar{Y}_t \right)^2$, pooling information across all controls matched to each treated unit.

This difference in approach leads to several important practical advantages and trade-offs. First, our estimator requires only matching controls to treated units, whereas Abadie and Imbens (2006) requires matching for both treatment and control groups—significantly reducing computational burden when the control group is large. However, this computational advantage comes at the cost of requiring our homoskedasticity assumption ($\sigma_t^2 = \sigma_c^2$ for matched pairs), while Abadie and Imbens (2006) can accommodate arbitrary heteroskedasticity across units.

Second, Abadie and Imbens (2006)'s approach necessitates matching treated units with other treated units to estimate $\hat{\sigma}_t^2$. This becomes problematic when the treated group is small or highly heterogeneous in covariates, as finding good same-treatment matches becomes difficult or impossible. Our approach avoids this issue entirely by focusing on control-to-treated matching, making it particularly suitable for ATT estimation where treated samples are typically small.

Third, our framework naturally accommodates flexible weighting schemes, including kernel weights, caliper matching weights, and optimal transportation weights, whereas Abadie and Imbens (2006)'s approach was primarily designed for fixed-number nearest neighbor matching with equal weights.

The main limitation of our approach is that we do not utilize within-treated-group variation for variance estimation—we do not use the observed outcomes Y_t of treated units when estimating σ^2, potentially discarding valuable information. This efficiency loss is the price of our computational simplicity and homoskedasticity assumption. However, this limitation is typically minor in ATT applications where the treated group is small relative to the control group, and within-treated-group variation becomes unreliable when the number of treated units is small. Many influential ATT applications feature relatively small treated samples, including job training program evaluations (LaLonde, 1986), educational interventions (Abadie et al., 2002), and health policy assessments (Keele et al., 2023), where Imbens (2004) notes that ATT estimation is often preferred precisely because treatment is relatively rare or targeted.

5 Simulation

In this section, we conduct simulation studies to validate the two main theoretical results established in earlier sections: Theorem 3.2 (Central Limit Theorem) and the consistency of our variance

estimator. The primary focus is threefold: first, to verify the asymptotic normality of our estimator, second, to assess whether confidence intervals constructed using our variance estimator achieve near-nominal coverage, and third, to compare the performance of our variance estimator to that of existing methods, demonstrating how our approach substantially outperforms the state-of-the-art bootstrap variance estimator proposed by Otsu and Rai (2017). These simulations provide empirical insights into the reliability and robustness of our methods under different data-generating scenarios and matching conditions.

5.1 Otsu–Rai DGP: Challenging Nonlinear Setting

We begin with the simulation design of Otsu and Rai (2017), which features nonlinear response surfaces known to be challenging for bootstrap inference. We fix the treatment effect at $\tau = 0$, set $(\gamma_1, \gamma_2) = (0.15, 0.7)$ for the propensity score, and vary the covariate dimension $K \in \{2, 4, 8\}$. The error term is drawn as $\epsilon_i \sim N(0, 0.2^2)$. Outcome functions $m(\cdot)$ are taken from the six nonlinear curves reported in Otsu and Rai (2017) and reproduced in Table 1.

Formally, the data generating process is:

$$
\begin{aligned}
&\{Y_i, Z_i, \mathbf{X}_i\}_{i=1}^n, \\
&Y_i(1) = \tau + m(\|\mathbf{X}_i\|) + \epsilon_i, \quad Y_i(0) = m(\|\mathbf{X}_i\|) + \epsilon_i, \\
&Z_i = \mathbb{I}\{P(\mathbf{X}_i) \geq v_i\}, \quad v_i \sim U[0,1], \\
&P(\mathbf{X}_i) = \gamma_1 + \gamma_2 \|\mathbf{X}_i\|, \quad \mathbf{X}_i = (X_{1i}, \ldots, X_{Ki})', \\
&X_{ji} = \xi_i |\zeta_{ji}| / \|\boldsymbol{\zeta}_i\|, \quad j = 1, \ldots, K, \\
&\xi_i \sim U[0,1], \quad \boldsymbol{\zeta}_i \sim N(\mathbf{0}, I_K).
\end{aligned}
$$

We implement 5-nearest neighbor matching with uniform weighting ($w_{jt} = 1/5$) across 500 replications. The sample sizes n_T and n_C are determined by the propensity score design, resulting in approximately balanced treatment and control groups with a 1:1 ratio. We also vary the total sample size n from 250 to 5000.

Figure 1 presents our main empirical findings. It shows a substantial performance gap between inference methods: our pooled variance estimator consistently achieves coverage rates much closer

Table 1: Nonlinear outcome functions $m(z)$ used in simulations

Curves	$m(z)$
1	$0.15 + 0.7z$
2	$0.1 + z/2 + \exp\left(-200(z-0.7)^2\right)/2$
3	$0.8 - 2(z-0.9)^2 - 5(z-0.7)^3 - 10(z-0.6)^{10}$
4	$0.2 + \sqrt{1-z} - 0.6(0.9-z)^2$
5	$0.2 + \sqrt{1-z} - 0.6(0.9-z)^2 - 0.1z\cos(30z)$
6	$0.4 + 0.25\sin(8z-5) + 0.4\exp\left(-16(4z-2.5)^2\right)$

to the nominal 95% level compared to the wild bootstrap method proposed by Otsu and Rai (2017). Across all covariate dimensions, our method maintains coverage rates between 93.8% and 99.0%, with an overall average of 96.7%, while the bootstrap method exhibits severe undercoverage ranging from 74.6% to 96.8%, averaging only 81.7%.

The performance differential becomes more obvious as sample size increases and covariate dimensionality decreases. Most notably, at the largest sample size ($n = 5000$) with low-dimensional covariates ($K = 2$), the bootstrap method achieves only 75.2% coverage across all six nonlinear curves. In contrast, our method maintains 96.3% coverage at this sample size, demonstrating robustness even in challenging settings.

The superior coverage performance of our method comes with appropriately wider confidence intervals. Our method produces confidence intervals with an average width of 0.092 compared to 0.057 for the bootstrap method. On average, the confidence interval length under our method is about 1.64 times larger than that under the bootstrap method across all sample sizes, covariate dimensions, and curve IDs. The bootstrap method's narrower intervals are artificially optimistic due to its failure to account for the true sampling variability induced by control unit dependencies. Detailed figures of confidence interval length can be found at Figure 3 in the Appendix.

One limitation of our estimator is the tendency toward slight overcoverage, particularly evident in high-dimensional settings where coverage rates occasionally reach 100%. This conservative behavior can be attributed to the fact that confidence interval lengths remain relatively stable across dimensions (averaging 0.092–0.093), while the underlying sampling variability may decrease in some settings. The challenging nature of the Otsu-Rai data generating process, where complex nonlinear outcome functions create additional estimation complexity, contributes to this conservative performance. We leave the investigation of refined interval calibration in high-dimensional settings as an

important direction for future research.

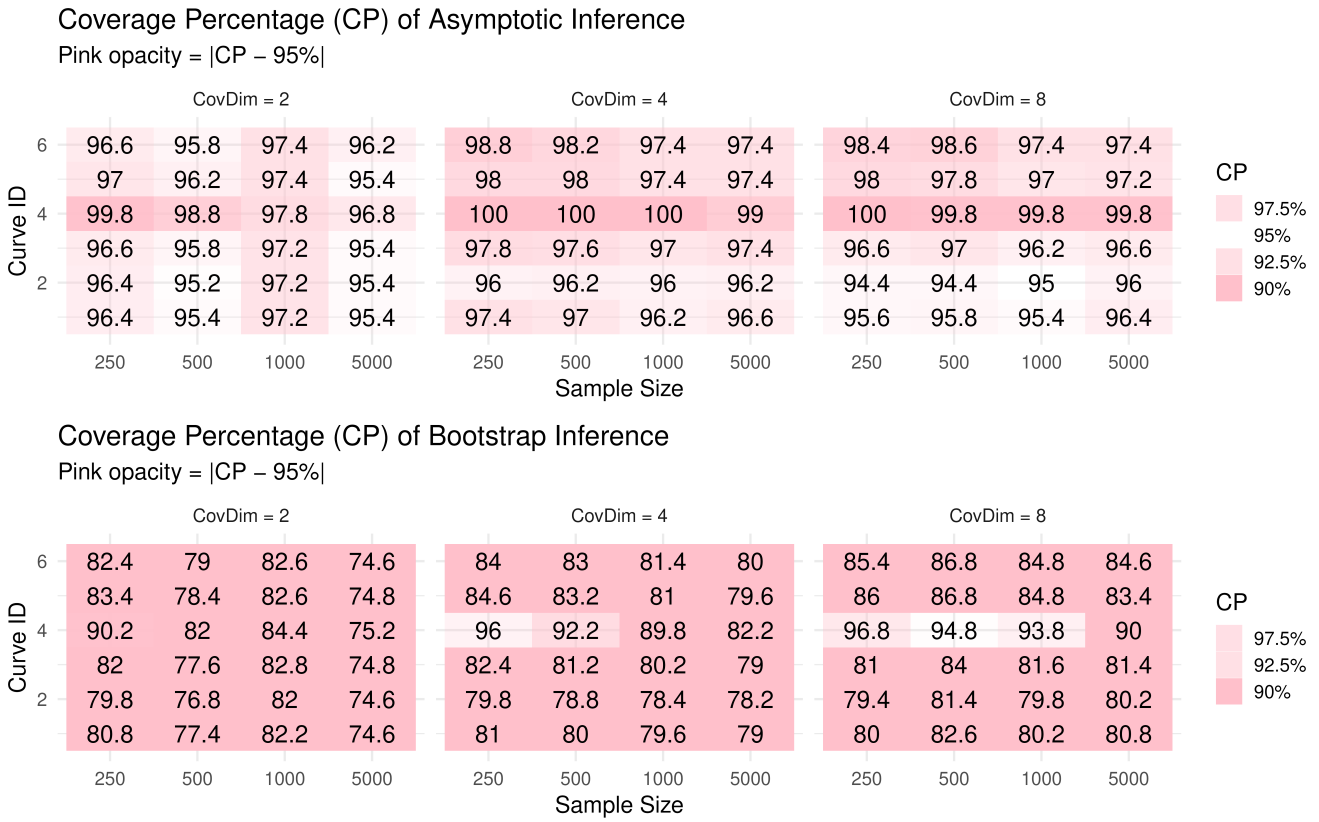

Figure 1: Simulation results for the Otsu-Rai data generating process across varying covariate dimensions ($K = 2, 4, 8$), sample sizes ($n = 250, 500, 1000, 5000$), and nonlinear outcome functions (curves 1–6). Coverage percentages for asymptotic inference (our method) versus bootstrap inference. Pink opacity indicates deviation from the nominal 95% rate. Our pooled variance estimator maintains coverage close to the nominal rate while the bootstrap method exhibits severe undercoverage, particularly at large sample sizes and low dimensions.

5.2 Che et al. DGP: Multi-Dimensional Validation

To provide comprehensive validation of our theoretical framework, we conduct additional simulations following the design from Che et al. (2024). This four-dimensional setting with varying degrees of population overlap provides secondary evidence of our method's robustness across different scenarios.

We maintain a 1:5 control–treated ratio and vary the total sample size across $n \in \{120, 240, 600, 1200, 2400\}$, corresponding to treated sample sizes of $n_T \in \{20, 40, 100, 200, 400\}$ respectively. Co-

variates are drawn from a 4-dimensional multivariate normal distribution $N((0.5, 0.5, 0.5, 0.5)^T, \Sigma)$, where the covariance matrix Σ has diagonal elements $\Sigma_{ii} = 1$ for all i and off-diagonal elements $\Sigma_{ij} = 0.8$ for $i \neq j$. For each unit with covariates (x_1, x_2, x_3, x_4), we generate outcomes via $Y = f_0(x_1, x_2, x_3, x_4) + Z \cdot \tau(x_1, x_2, x_3, x_4) + \epsilon$, where f_0 is the density function for the same multivariate normal distribution and $\tau(x_1, x_2, x_3, x_4) = 3\sum_{i=1}^{4} x_i$ is the heterogeneous treatment effect function. We vary the degree of overlap by adjusting the distribution parameters and use 5-nearest neighbor matching with uniform weighting ($w_{jt} = 1/5$) across 500 replications.

We consider two error variance structures to test the robustness of our method:

$$\text{Homoskedastic:} \quad \epsilon_i \sim \mathcal{N}(0, 0.5^2) \tag{17}$$

$$\text{Covariate-dependent:} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2(\mathbf{X}_i)) \quad \text{where} \quad \sigma^2(\mathbf{X}_i) = 0.25 + 0.5 \cdot \|\mathbf{X}_i - \bar{\mathbf{X}}\| \tag{18}$$

Figure 2 shows that while the performance gap between methods is smaller than in the Otsu-Rai setting, our pooled variance estimator consistently outperforms the bootstrap method across both variance structures. In the homoskedastic case, our method achieves coverage rates very close to the nominal 95% level, while the bootstrap method consistently falls below 95%, showing systematic undercoverage across all overlap scenarios.

The difference in performance is clearer in the covariate-dependent variance setting. While our method generally maintains coverage close to 95%, the bootstrap method performs reasonably well in most scenarios but fails dramatically when the degree of overlap is very high. This shows the bootstrap's inability to properly account for the complex dependency structure that emerges when high-quality control units are extensively reused across multiple treated units, particularly when variance heterogeneity compounds the estimation challenges.

The confidence interval analysis shows that our method consistently produces wider intervals than the bootstrap, with the difference being more pronounced in the covariate-dependent variance setting. On average, the CI length under our method is about 1.06 times larger than the bootstrap CI length. Our method produces appropriately conservative intervals while the bootstrap method's narrower intervals are artificially optimistic because it fails to account for the true sampling variability induced by overlapped controls and variance heterogeneity. Detailed figures of confidence interval length can be found at Figure 3 in the Appendix.

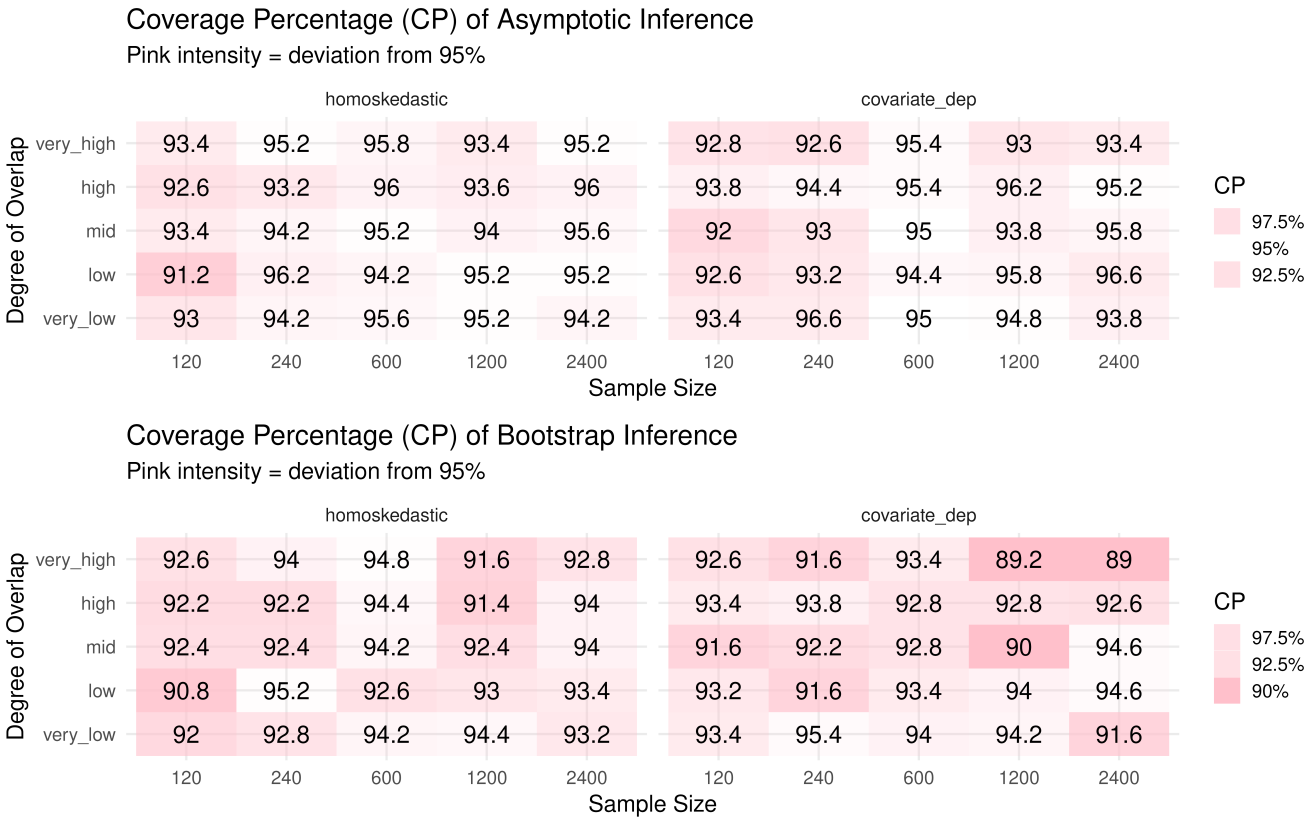

Figure 2: Simulation results for the Che et al. data generating process across varying degrees of population overlap and two error variance structures. Coverage percentages comparing our pooled variance estimator (asymptotic inference) with the wild bootstrap method across different overlap scenarios and variance structures. Results show our method maintains coverage closer to the nominal 95% rate, particularly in challenging covariate-dependent variance settings with high overlap.

5.3 Summary of Simulation Evidence

The simulation evidence provides strong support for our theoretical framework and demonstrates the practical importance of our contributions. In challenging settings with substantial control unit reuse—a common occurrence in real-world matching applications—our method maintains proper coverage while the current state-of-the-art bootstrap approach can fail dramatically. The robustness of our approach across different data generating processes, dimensions, and overlap patterns makes it a reliable tool for practitioners seeking valid population inference in matching-based causal studies.

Additional results examining the performance of individual variance components and bias correction effects are presented in Appendix J.2.

6 Application: Education Program Evaluation in Brazil

To illustrate the practical importance of our variance estimation framework, we analyze data from Brazil's "Jovem de Futuro" (Young of the Future) education program, following the experimental design of Barros et al. (2012) and Ferman (2021). This application demonstrates how our robust inference methods affect substantive conclusions in a setting with extensive control unit reuse—precisely the scenario where existing methods can fail.

6.1 Data and Matching Design

The Jovem de Futuro program offered management strategies and conditional grants to schools in Rio de Janeiro and São Paulo from 2010-2012. Following Ferman (2021)'s approach, we employ a within-study comparison design where experimental control schools (those randomized to receive no intervention) serve as our "treatment" group. We match these $n_T = 54$ experimental control schools to $n_C = 4{,}447$ non-participating schools to estimate what should be a null effect if matching successfully removes selection bias.

Pre-treatment covariates consist of standardized test scores from 2007-2009 and a state indicator. Table 2 shows substantial pre-treatment differences between experimental and non-participating schools, with experimental schools having consistently lower baseline scores across all years, motivating the use of matching methods.

Table 2: Summary Statistics for Brazilian School Data

Variable	Non-participating Schools (Control)	Experimental Schools (Treated)	Standardized Difference
Score 2007	0.047	−0.028	−0.075
Score 2008	0.008	−0.010	−0.018
Score 2009	0.023	−0.025	−0.048
São Paulo (%)	78.1	72.2	−5.9
Sample size	4,447	54	

Note: Test scores are standardized with mean 0 and standard deviation 1 in the full sample. São Paulo percentage indicates the proportion of schools from São Paulo state.

We implement radius matching following Che et al. (2024), using the L^∞ distance metric with a distance caliper of $c = 0.35$. We impose covariate-specific calipers of 0.2 standard deviations for each pre-treatment test score (2007-2009) and require near-exact matching on state with a caliper of 0.001. This configuration ensures high-quality matches while maintaining adequate sample size—49 of 54 treated units (91%) find at least one match within the specified radius, with the remaining 5 units matched adaptively to their nearest neighbor. For matched units, we apply synthetic control weights to minimize covariate imbalance within each matched set. The matching procedure achieves excellent covariate balance, with post-matching standardized differences below 0.1 for all covariates.

6.2 Control Unit Reuse and Effective Sample Size

A key feature of this application is the minimal reuse of control schools in our matched sample. Table 3 presents diagnostics that characterize the dependency structure created by matching. The mean control reuse of 1.02 indicates that control schools are rarely matched to multiple treated units—on average, each control school is matched to just one treated unit, with only a small fraction matched to two treated units.

This minimal control reuse arises from the combination of a small treated sample ($n_T = 54$) relative to the large control reservoir ($n_C = 4{,}447$), along with our radius matching design that allows variable numbers of matches per treated unit. The maximum reuse of only 2 indicates that even the best control schools are matched to at most two treated units.

Despite the limited control reuse, the effective sample size (ESS) of 95 is notably lower than the 155 unique controls used, with an ESS ratio of 61.3%. This reduction in effective sample size

Table 3: Control Unit Reuse and Effective Sample Size in the Matched Sample

Statistic	Value
Mean control reuse	1.02
Median control reuse	1
Maximum control reuse	2
Proportion of controls matched to multiple treated units	1.9%
Effective sample size (ESS)	95
Number of unique controls	155
ESS/Number of unique controls	61.3%

Note: Mean control reuse measures the average number of times each control unit is matched to treated units. A value of 1 indicates no reuse; higher values indicate greater dependency in the matched sample. The effective sample size accounts for both control reuse and the heterogeneous weights from synthetic control optimization.

is primarily driven by the synthetic control weighting scheme rather than control reuse per se. The synthetic control optimization assigns heterogeneous weights to minimize covariate imbalance within each matched set, with some controls receiving substantially higher weights than others. This unequal weighting, while improving covariate balance, reduces the effective independent information in the sample. Our variance estimator properly accounts for this reduction through the ESS calculation, ensuring valid inference despite the loss of effective sample size from the weighting scheme.

6.3 Treatment Effect Estimates and Inference

Table 4 presents estimates of the average treatment effect on the treated using different inference methods. The point estimate of 0.035 is close to zero, as expected in this within-study comparison. This null effect is by design: we are comparing experimental control schools (randomized to receive no treatment) to observationally similar non-participating schools. If our matching procedure successfully removes selection bias, we should find no systematic difference between these groups, validating the matching method's ability to create appropriate counterfactuals.

Both inference methods produce similar standard errors (0.030 vs 0.029), with 95% confidence intervals that include zero. This similarity is expected given the minimal control reuse in our matched sample (mean reuse of 1.02). With limited dependency structure, the wild bootstrap

Table 4: ATT Estimates and Variance Components for 2010 Test Scores

Method	Point Estimate	SE	95% CI	$n_T\hat{V}_E$	$n_T\hat{V}_P$
Our pooled variance estimator	0.035	0.030	$(-0.024, 0.094)$	0.055	-0.400
Wild bootstrap (Otsu-Rai)	0.035	0.029	$(-0.022, 0.092)$	0.045	–

Note: Both methods use radius matching with synthetic control weights. The variance components show $n_T\hat{V}_E$ (sampling variance due to residual noise) and $n_T\hat{V}_P$ (variance due to treatment effect heterogeneity). Wild bootstrap based on 1,000 replications.

performs adequately, and both methods lead to the same substantive conclusion: we cannot reject the null hypothesis of no selection bias after matching.

An important feature of our variance estimator is the decomposition into measurement error variance ($\hat{V}_E$) and population heterogeneity variance ($\hat{V}_P$). The scaled variance components show $n_T\hat{V}_E = 0.055$, indicating modest sampling variability due to residual outcome noise. The estimate $n_T\hat{V}_P = -0.400$ is negative, which occurs because we obtain $n_T\hat{V}_P$ through subtraction: we subtract $n_T\hat{V}_E$ in Equation (9) from $\hat{V}$ in Equation (15). While theoretically this decomposition is accurate in probability limit, in finite samples it is possible for $\hat{V} < n_T\hat{V}_E$ due to sampling variability of both estimators. The negative value suggests minimal treatment effect heterogeneity in this sample. Developing improved estimators for $n_T\hat{V}_P$ that ensure non-negativity while maintaining consistency remains an interesting direction for future work.

7 Conclusion

We have presented a new framework for inference with matching estimators that strengthens both theoretical and practical foundations. Our analysis establishes a central limit theorem for a broad class of matching procedures under heteroskedastic errors, extends variance decompositions to include a previously unrecognized covariance term, and introduces a variance estimator that is computationally simple and consistent.

Simulations demonstrate that these refinements have substantial practical value. While the wild bootstrap of Otsu and Rai (2017) often undercovers, our estimator consistently delivers confidence intervals with coverage close to the nominal rate, even in moderately sized samples. The improvements are not subtle: coverage gaps can reach 20 percentage points, underscoring how minor-seeming theoretical adjustments can yield dramatic empirical benefits.

Methodologically, our work parallels the role of heteroskedasticity-robust variance estimation in regression: it provides a theoretically justified and broadly applicable correction that improves inference reliability. Practically, our results equip applied researchers with a variance estimator that is easy to compute, requires only treated-to-control matching, and produces trustworthy confidence intervals in settings where existing approaches falter.

We view these contributions as refinements to a well-studied problem rather than a wholesale rethinking of matching inference. Yet the payoff of these refinements is large: by carefully addressing overlooked variance components and grounding inference in rigorous asymptotics, we achieve both theoretical clarity and dramatic gains in empirical performance. Future work may extend these tools to weighting methods and other causal estimators, further unifying the inferential foundations of design-based approaches in causal inference.

References

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. Econometrica, 74(1):235–267, 2006.

Alberto Abadie and Guido W Imbens. On the failure of the bootstrap for matching estimators. Econometrica, 76(6):1537–1557, 2008.

Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics, 29(1):1–11, 2011.

Alberto Abadie and Guido W Imbens. A martingale representation for matching estimators. Journal of the American Statistical Association, 107(498):833–843, 2012.

Alberto Abadie and Jörg Spiess. Robust post-matching inference. Journal of the American Statistical Association, 117(540):1811–1827, 2022.

Alberto Abadie, Joshua Angrist, and Guido Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica, 70(1):91–117, 2002.

Peter C Austin and Guy Cafri. Variance estimation when using propensity-score matching with

replacement with survival or time-to-event outcomes. Statistics in Medicine, 39(11):1623–1640, 2020.

Ricardo Barros, Mirela de Carvalho, Samuel Franco, and Andrezza Rosalém. Impacto do projeto jovem de futuro. Est. Aval. Educ, pages 214–226, 2012.

Hugo Bodory, Lorenzo Camponovo, Martin Huber, and Michael Lechner. The finite sample performance of inference methods for propensity score matching and weighting estimators. Journal of Business & Economic Statistics, 38(1):183–200, 2020.

Jonathan Che, Xiang Meng, and Luke Miratrix. Caliper synthetic matching: Generalized radius matching with local synthetic controls. arXiv preprint arXiv:2411.05246, 2024.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94(448): 1053–1062, 1999.

Bruno Ferman. Matching estimators with few treated and many control observations. Journal of Econometrics, 225(2):295–307, 2021.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). Bayesian Analysis, 15(3):965–1056, 2020.

Peter Hall and Christopher C Heyde. Martingale limit theory and its application. Academic press, 2014.

James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The Review of Economic Studies, 64(4):605–654, 1997.

Jennifer Hill and Jerome P. Reiter. Interval estimation for treatment effects using propensity score matching. Statistics in Medicine, 25(14):2230–2256, 2006.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1):217–240, 2011.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4):1161–1189, 2003.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and Statistics, 86(1):4–29, 2004.

Nathan Kallus. Generalized optimal matching methods for causal inference. J. Mach. Learn. Res., 21:62–1, 2020.

Luke J. Keele, Eli Ben-Michael, Avi Feller, Rachel Kelz, and Luke Miratrix. Hospital quality risk standardization via approximate balancing weights. The Annals of Applied Statistics, 17(2), June 2023. ISSN 1932-6157. doi: 10.1214/22-AOAS1629. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-17/issue-2/Hospital-quality-risk-standardization-via-approximate-balancing-weights/10.1214/22-AOAS1629.full.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. The American economic review, pages 604–620, 1986.

Taisuke Otsu and Yoshiyasu Rai. Bootstrap inference of matching estimators for average treatment effects. Journal of the American Statistical Association, 112(520):1720–1732, 2017.

Richard F Potthoff, Max A Woodbury, and Kenneth G Manton. "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. 2024.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Donald B Rubin. Matching to remove bias in observational studies. Biometrics, pages 159–183, 1973.

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

Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? Journal of Econometrics, 125(1-2):305–353, 2005.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. Statistical Science, 25(1):1–21, 2010.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: journal of the Econometric Society, pages 817–838, 1980.

Appendix

A Proof of Theorem 3.2

We prove that

$$\frac{\sqrt{n_T}(\hat{\tau} - B_n - \tau)}{\sqrt{V_E + V_P}} \xrightarrow{d} N(0,1).$$

Step 1: Decomposition. Recall the decomposition

$$\hat{\tau} - \tau = P_n + E_n + B_n,$$

where

$$P_n = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \{\tau(X_t) - \tau\},$$
$$E_n = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \epsilon_t - \frac{1}{n_T} \sum_{j \in \mathcal{C}} w_j \epsilon_j.$$

The bias term B_n is handled separately. The main task is to establish a joint CLT for (P_n, E_n).

Step 2: Martingale representation. Let $\mathcal{F}_i = \sigma\{(Z_\ell, X_\ell, Y_\ell) : \ell \leq i\}$. Define

$$M_n^{(P)} = \frac{1}{\pi} \sum_{i=1}^{n} Z_i(\tau(X_i) - \tau),$$
$$M_n^{(E)} = \frac{1}{\pi} \sum_{i=1}^{n} \{Z_i \epsilon_i - (1 - Z_i) w_i \epsilon_i\},$$

where w_i is the total weight assigned to control i across all matches (and equals 1 if i is treated). Then $M^{(P)}$ and $M^{(E)}$ are martingales with respect to $\{\mathcal{F}_i\}$.

For $M^{(P)}$, note that

$$E[M_n^{(P)} | \mathcal{F}_{n-1}] = M_{n-1}^{(P)} + \tfrac{1}{\pi} E[Z_n(\tau(X_n) - \tau) \mid \mathcal{F}_{n-1}] = M_{n-1}^{(P)},$$

since $E[Z_n | X_n] = \pi$ and $E[\tau(X_n) - \tau] = 0$. For $M^{(E)}$, a similar calculation using $E[\epsilon_n | X_n, Z_n] = 0$

yields

$$E[M_n^{(E)}|\mathcal{F}_{n-1}] = M_{n-1}^{(E)}.$$

Thus both are martingales. Moreover,

$$P_n = \frac{\pi}{n_T} M_n^{(P)}, \qquad E_n = \frac{\pi}{n_T} M_n^{(E)}.$$

Step 3: Martingale CLT setup. Define the martingale difference array

$$X_{n,k} = \frac{1}{\sqrt{n}} \begin{pmatrix} \Delta M_k^{(P)} \\ \Delta M_k^{(E)} \end{pmatrix}, \qquad \Delta M_k^{(P)} = \tfrac{1}{\pi} Z_k(\tau(X_k) - \tau), \;\; \Delta M_k^{(E)} = \tfrac{1}{\pi}(Z_k - (1 - Z_k)w_k)\epsilon_k.$$

Then $S_n = \sum_{k=1}^n X_{n,k} = (M_n^{(P)}, M_n^{(E)})^\top / \sqrt{n}$.

Step 4: Quadratic variations. The conditional variance for $\Delta M_k^{(P)}$ is

$$\begin{aligned} E[(\Delta M_k^{(P)})^2|\mathcal{F}_{k-1}] &= \frac{1}{\pi^2} E[Z_k(\tau(X_k) - \tau)^2 \mid \mathcal{F}_{k-1}] \\ &= \frac{1}{\pi} Var(\tau(X) \mid Z = 1) =: \tfrac{n_T V_P}{\pi}. \end{aligned}$$

Hence

$$\sum_{k=1}^n E[(\Delta M_k^{(P)})^2|\mathcal{F}_{k-1}] = n \cdot \tfrac{n_T V_P}{\pi}.$$

For $\Delta M_k^{(E)}$,

$$E[(\Delta M_k^{(E)})^2|\mathcal{F}_{k-1}] = \frac{1}{\pi^2} E[(Z_k + (1 - Z_k)w_k^2)\sigma_{k,Z_k}^2(X_k) \mid \mathcal{F}_{k-1}],$$

where $\sigma_{k,z}^2(X_k) = E[\epsilon_k^2|X_k, Z_k = z]$. Taking expectations,

$$E[n_T^2 V_E] = \pi^2 E\left[\sum_{k=1}^n E[(\Delta M_k^{(E)})^2|\mathcal{F}_{k-1}]\right].$$

Proposition A.1. *We have*

$$\frac{1}{n}\left[\sum_{k=1}^{n} E\big[(\Delta M_k^{(E)})^2 \mid \mathcal{F}_{k-1}\big] - \frac{1}{\pi^2} E[n_T^2 V_E]\right] \xrightarrow{p} 0.$$

Proof. Note that the summands $E[(\Delta M_k^{(E)})^2|\mathcal{F}_{k-1}]$ are uniformly integrable and bounded in expectation by Assumption 4 and finite moments of the weights. By the predictable law of large numbers for martingales (Hall and Heyde, 2014, Theorem 2.18), the empirical averages converge to their expectations, yielding the result. □

Cross terms vanish: $E[\Delta M_k^{(P)} \Delta M_k^{(E)}|\mathcal{F}_{k-1}] = 0$, so the martingales are orthogonal.

Step 5: Lindeberg condition. For $M^{(P)}$, bounded or sub-Gaussian treatment effects imply $|\Delta M_k^{(P)}|$ is uniformly bounded, so the Lindeberg condition holds trivially.

For $M^{(E)}$, we must show

$$\frac{1}{n}\sum_{k=1}^{n} E\Big[(\Delta M_k^{(E)})^2\, 1\{|\Delta M_k^{(E)}| > \varepsilon\sqrt{n}\} \mid \mathcal{F}_{k-1}\Big] \xrightarrow{p} 0.$$

Fix $p = \frac{2+\delta}{2} > 1$. For any integrable Y and $t > 0$,

$$E[Y^2 1\{|Y| > t\} \mid \mathcal{F}_{k-1}] \le \frac{\big(E[|Y|^{2+\delta} \mid \mathcal{F}_{k-1}]\big)^{2/(2+\delta)} \big(E[Y^2 \mid \mathcal{F}_{k-1}]\big)^{\delta/(2+\delta)}}{t^{2\delta/(2+\delta)}}.$$

Apply this to $Y = \Delta M_k^{(E)}$ and $t = \varepsilon\sqrt{n}$. By Assumption 4, $\sup_x E[|\epsilon|^{2+\delta}|X = x, Z = z] < \infty$ and $\sigma_z^2(x)$ is bounded. Finite moments of w_k then imply

$$E[(\Delta M_k^{(E)})^{2+\delta} \mid \mathcal{F}_{k-1}] \le C, \qquad E[(\Delta M_k^{(E)})^2 \mid \mathcal{F}_{k-1}] \le C,$$

for some constant C independent of n, k. Thus

$$E[(\Delta M_k^{(E)})^2 1\{|\Delta M_k^{(E)}| > \varepsilon\sqrt{n}\} \mid \mathcal{F}_{k-1}] \le C' n^{-\delta/(2+\delta)},$$

and summing over k gives

$$\frac{1}{n}\sum_{k=1}^{n} E[(\Delta M_k^{(E)})^2 1\{|\Delta M_k^{(E)}| > \varepsilon\sqrt{n}\} \mid \mathcal{F}_{k-1}] \;\le\; C' n^{-\delta/(2+\delta)} \to 0.$$

Hence the Lindeberg condition holds.

Step 6: Limit distribution. By the two-dimensional martingale CLT,

$$\frac{1}{\sqrt{n}}\begin{pmatrix} M_n^{(P)} \\ M_n^{(E)} \end{pmatrix} \;\Rightarrow\; N\left(0,\; \begin{pmatrix} \frac{n_T V_P}{\pi} & 0 \\ 0 & \pi^{-1} E[n_T V_E] \end{pmatrix}\right).$$

Since $P_n = \pi M_n^{(P)}/n_T$ and $E_n = \pi M_n^{(E)}/n_T$, with $n_T/n \to \pi$, we obtain

$$\sqrt{n_T}\,(P_n + E_n) \;\Rightarrow\; N(0,\, V_P + V_E)\,.$$

Therefore

$$\frac{\sqrt{n_T}(\hat{\tau} - B_n - \tau)}{\sqrt{V_E + V_P}} \;\Rightarrow\; N(0,1).$$

This establishes Theorem 3.2.

B Proof of Theorem 4.1

Proof. Expand the difference:

$$n_T\left(\hat{V}_E - V_E\right) = \frac{1}{n_T}\sum_{t\in\mathcal{T}}(s_t^2 - \sigma_t^2) + \frac{1}{n_T}\sum_{j\in\mathcal{C}} w_j^2\,(s_j^2 - \sigma_j^2).$$

The first term has exactly the same form as Term A in the proof of Lemma 4.2, except scaled by $1/n_T$. Using the same decomposition (sampling error, cross-product, interaction, and systematic differences, cf. Equations (20)–(20e)), and applying the same moment bounds and shrinking–cluster arguments, we obtain

$$\frac{1}{n_T}\sum_{t\in\mathcal{T}}(s_t^2 - \sigma_t^2) \;\xrightarrow{p}\; 0.$$

For the second term, note that $w_j^2 \le K(j)^2$ where $K(j)$ is the reuse count of control j. By

Lemma C.2, $K(j)$ has all finite moments under the exponential tail condition, and hence $\mathbb{E}[w_j^2] < \infty$. The same argument as above (treating $s_j^2 - \sigma_j^2$ as an error term with bounded moments, independent across controls given covariates) shows

$$\frac{1}{n_T} \sum_{j \in \mathcal{C}} w_j^2 \left(s_j^2 - \sigma_j^2\right) \xrightarrow{p} 0.$$

Combining both parts yields the stated result. □

C Proof of Lemma 4.2

Proof. Let us decompose the difference between our variance estimator and the true average variance:

$$\begin{aligned}
S^2 - \frac{1}{n_T} \sum_{t \in \mathcal{T}} \sigma_t^2 &= \frac{1}{N_C} \sum_{t \in \mathcal{T}} |\mathcal{C}_t| s_t^2 - \frac{1}{n_T} \sum_{t \in \mathcal{T}} \sigma_t^2 \\
&= \sum_{t \in \mathcal{T}} u_t s_t^2 - \frac{1}{n_T} \sum_{t \in \mathcal{T}} \sigma_t^2 \\
&= \sum_{t \in \mathcal{T}} (u_t s_t^2 - \frac{1}{n_T} \sigma_t^2) \\
&= \sum_{t \in \mathcal{T}} \underbrace{(u_t s_t^2 - u_t \sigma_t^2)}_{\text{Term A}} + \sum_{t \in \mathcal{T}} \underbrace{(u_t \sigma_t^2 - \frac{1}{n_T} \sigma_t^2)}_{\text{Term B}}
\end{aligned}$$

where $u_t = \frac{|\mathcal{C}_t|}{N_C}$ represents the weight of cluster t in the pooled estimator. Note that

$$N_C = \sum_{t \in \mathcal{T}} |\mathcal{C}_t| \tag{19}$$

is the total number of matches[2].

The lemma holds if two conditions are established:

1. **Term A vanishes:** $\sum_{t \in \mathcal{T}} (u_t s_t^2 - u_t \sigma_t^2) \to 0$ in probability as $n_T \to \infty$;

[2]If a control unit is matched to multiple treated units, it contributes to N_C multiple times. For example, if a control unit is matched to three treated units, it adds 3 to N_C rather than 1.

2. **Term B equals the covariance adjustment:**

$$\sum_{t\in\mathcal{T}}\Big(u_t\sigma_t^2 - \tfrac{1}{n_T}\sigma_t^2\Big) = \frac{1}{\sum_{t\in\mathcal{T}}|\mathcal{C}_t|/n_T}\ \mathrm{Cov}_v\big(|\mathcal{C}_t|, \sigma_t^2\big)\,.$$

Subtracting Term B from both sides then yields exactly the form in Equation (14). We handle Term A in Section C.1 and Term B in Section C.2.

C.1 Proof that Term A goes to zero

We first analyze Term A, which measures the difference between the estimated and true variance within each cluster. For a fixed treatment t, for each individual matched control j in $\mathcal{C}_t$, we focus on the summand in $s_t^2 = \frac{1}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}(Y_j - \bar{Y}_t)^2$ (introduced in Equation 7) and expand the squared deviation:

$$\begin{aligned}
(Y_j - \bar{Y}_t)^2 &= \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k) + \epsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\epsilon_k\right)^2 \\
&= \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k)\right)^2 \\
&\quad + 2\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k)\right)\left(\epsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\epsilon_k\right) \\
&\quad + \left(\epsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\epsilon_k\right)^2
\end{aligned}$$

Therefore, the difference between the sample variance and the true variance can be written as:

$$
\begin{aligned}
s_t^2 - \sigma_t^2 &= \frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} (Y_j - \bar{Y}_t)^2 - \sigma_t^2 \\
&= \underbrace{\left(\frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \epsilon_j^2 - \sigma_t^2 \right)}_{\text{Sampling error}} \\
&+ \underbrace{\frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[-2\epsilon_j \left(\frac{1}{|\mathcal{C}_t|} \sum_{\substack{k \in \mathcal{C}_t \\ k \neq j}} \epsilon_k \right) \right]}_{\text{Cross-product of errors}} \\
&+ \underbrace{\frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[2 \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right) \left(\epsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \epsilon_k \right) \right]}_{\text{Interaction between function and errors}} \\
&+ \underbrace{\frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right)^2 \right]}_{\text{Systematic differences within cluster}}
\end{aligned}
$$

□

Now, Term A becomes the following decomposition:

$$\text{Term A} = \sum_{t\in\mathcal{T}}(u_t s_t^2 - u_t\sigma_t^2) \tag{20a}$$

$$= \underbrace{\sum_{t\in\mathcal{T}}\frac{u_t}{|\mathcal{C}_t|}\sum_{j\in\mathcal{C}_t}(\varepsilon_j^2 - \sigma_t^2)}_{\text{Sampling error}} \tag{20b}$$

$$+ \underbrace{\sum_{t\in\mathcal{T}}\frac{u_t}{|\mathcal{C}_t|}\sum_{j\in\mathcal{C}_t}\left[-2\varepsilon_j\left(\frac{1}{|\mathcal{C}_t|}\sum_{\substack{k\in\mathcal{C}_t\\k\neq j}}\varepsilon_k\right)\right]}_{\text{Cross-product of errors}} \tag{20c}$$

$$+ \underbrace{\sum_{t\in\mathcal{T}}\frac{u_t}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}\left[-2\Big(f_0(X_j) - \overline{f}_{0,t}\Big)\Big(\varepsilon_j - \overline{\varepsilon}_t\Big)\right]}_{\text{Interaction between function and errors}} \tag{20d}$$

$$+ \underbrace{\sum_{t\in\mathcal{T}}\frac{u_t}{|\mathcal{C}_t|-1}\sum_{j\in\mathcal{C}_t}\left[\Big(f_0(X_j) - \overline{f}_{0,t}\Big)^2\right]}_{\text{Systematic differences within cluster}} \tag{20e}$$

Let's focus on the first component of Term A, the sampling error:

$$\begin{aligned}
(20b) &= \sum_{t\in\mathcal{T}}\frac{u_t}{|\mathcal{C}_t|}\sum_{j\in\mathcal{C}_t}(\varepsilon_j^2 - \sigma_t^2)\\
&= \sum_{c\in\mathcal{C}}\sum_{t\in\mathcal{T}_c}\frac{1}{\sum_{c\in\mathcal{C}}K(c)}(\varepsilon_c^2 - \sigma_t^2)\\
&= \sum_{c\in\mathcal{C}}\sum_{t\in\mathcal{T}_c}\frac{1}{\sum_{c\in\mathcal{C}}K(c)}(\varepsilon_c^2 - \sigma_c^2 + \sigma_c^2 - \sigma_t^2)\\
&= \underbrace{\frac{1}{\sum_{c\in\mathcal{C}}K(c)}\sum_{c\in\mathcal{C}}K(c)\left(\varepsilon_c^2 - \sigma_c^2\right)}_{\text{first term of first component}} \qquad (21a)\\
&\quad + \underbrace{\frac{1}{\sum_{c\in\mathcal{C}}K(c)}\sum_{c\in\mathcal{C}}\sum_{t\in\mathcal{T}_c}\left(\sigma_c^2 - \sigma_t^2\right)}_{\text{second term of first component}} \qquad (21b)
\end{aligned}$$

where $\mathcal{T}_c$ is the set of treated units matched to control unit c. $K(c) = |\mathcal{T}_c|$ represents the number of times control unit c is used across all matches. Note that $\sum_{c\in\mathcal{C}}K(c) = \sum_{t\in\mathcal{T}}|\mathcal{C}_t| = N_C$ is the total number of matches (Equation 19).

C.1.1 First term, 21a goes to zero under a $(2+\delta)/2$-moment condition.

Write

$$S_n := \frac{1}{\sum_{c\in\mathcal{C}} K(c)} \sum_{c\in\mathcal{C}} K(c)\left(\varepsilon_c^2 - \sigma_c^2\right) = \sum_{c\in\mathcal{C}} a_c\,\xi_c, \qquad a_c := \frac{K(c)}{\sum_{c'\in\mathcal{C}} K(c')},\ \ \xi_c := \varepsilon_c^2 - \sigma_c^2.$$

We separately discuss cases when $q = (2+\delta)/2 \in (1,2]$ and $q > 2$ because different techniques are used.

Case $q = (2+\delta)/2 \in (1,2]$ Conditional on the matching covariates $\mathcal{X} := \{X_i, Z_i\}_{i=1}^n$ (hence on $\{K(c)\}_{c\in\mathcal{C}}$), the $\{\xi_c\}_{c\in\mathcal{C}}$ are independent with $\mathbb{E}[\xi_c \mid \mathcal{X}] = 0$ and have uniformly bounded q-th moments by Lemma C.1 below. By the von Bahr–Esseen inequality for $1 \le q \le 2$,

$$\mathbb{E}\big[\big|S_n\big|^q \,\big|\, \mathcal{X}\big] \;\le\; 2\sum_{c\in\mathcal{C}} |a_c|^q\, \mathbb{E}\big[|\xi_c|^q \mid \mathcal{X}\big] \;\le\; 2C \sum_{c\in\mathcal{C}} a_c^q \;=\; 2C \cdot \frac{\sum_{c\in\mathcal{C}} K(c)^q}{\big(\sum_{c\in\mathcal{C}} K(c)\big)^q}.$$

Taking expectations and then Markov's inequality yields, for any $\varepsilon > 0$,

$$\mathbb{P}\big(|S_n| > \varepsilon\big) \;\le\; \frac{2C}{\varepsilon^q}\, \mathbb{E}\left[\frac{\sum_{c\in\mathcal{C}} K(c)^q}{\big(\sum_{c\in\mathcal{C}} K(c)\big)^q}\right].$$

Thus, if

$$\frac{1}{\big(\sum_{c\in\mathcal{C}} K(c)\big)^q} \sum_{c\in\mathcal{C}} K(c)^q \;\xrightarrow{p}\; 0, \qquad q = (2+\delta)/2, \tag{22}$$

we have $S_n \xrightarrow{p} 0$ by bounded convergence.

(22) is true: for control reuse counts $\{K(c)\}_{c\in\mathcal{C}}$, we have $\mathbb{E}[K(c)^q \mid Z = 0] < \infty$ and $\mathbb{E}[K(c) \mid Z = 0] > 0$ (see Lemma C.2) Hence, by the law of large numbers

$$\frac{1}{n_C}\sum_{c\in\mathcal{C}} K(c) \xrightarrow{p} \mathbb{E}[K(c) \mid Z = 0], \qquad \frac{1}{n_C}\sum_{c\in\mathcal{C}} K(c)^q \xrightarrow{p} \mathbb{E}[K(c)^q \mid Z = 0],$$

so that

$$\frac{\sum_c K(c)^q}{\big(\sum_c K(c)\big)^q} = \frac{\frac{1}{n_C}\sum_c K(c)^q}{\big(\frac{1}{n_C}\sum_c K(c)\big)^q} \cdot n_C^{1-q} \;\xrightarrow{p}\; 0 \quad \text{since } q > 1.$$

Case $q > 2$ By Rosenthal's inequality (for independent mean-zero summands and $q \geq 2$), there exists $C_q < \infty$ such that

$$\mathbb{E}[|S_n|^q \,|\, \mathcal{X}] \;\leq\; C_q \left\{ \Big(\sum_c a_c^2 \, \mathbb{E}[\xi_c^2 \mid \mathcal{X}] \Big)^{q/2} \;+\; \sum_c |a_c|^q \, \mathbb{E}[|\xi_c|^q \mid \mathcal{X}] \right\}.$$

Using $\sup_c \mathbb{E}[\xi_c^2] \leq M_2 < \infty$,

$$\mathbb{E}[|S_n|^q \,|\, \mathcal{X}] \;\leq\; C_q \left\{ M_2^{q/2} \Big(\sum_c a_c^2 \Big)^{q/2} \;+\; M_q \sum_c a_c^q \right\}.$$

Since $a_c = K(c) / \sum_{c'} K(c')$,

$$\sum_c a_c^2 = \frac{\sum_c K(c)^2}{\big(\sum_c K(c)\big)^2}, \qquad \sum_c a_c^q = \frac{\sum_c K(c)^q}{\big(\sum_c K(c)\big)^q}.$$

Hence

$$\mathbb{E}[|S_n|^q] \;\leq\; C_q \left\{ M_2^{q/2} \, \mathbb{E}\left[\Big(\tfrac{\sum_c K(c)^2}{(\sum_c K(c))^2} \Big)^{q/2} \right] + M_q \, \mathbb{E}\left[\tfrac{\sum_c K(c)^q}{(\sum_c K(c))^q} \right] \right\}. \tag{23}$$

Therefore, if

$$\frac{\sum_c K(c)^2}{\big(\sum_c K(c)\big)^2} \xrightarrow{p} 0 \quad \text{and} \quad \frac{\sum_c K(c)^q}{\big(\sum_c K(c)\big)^q} \xrightarrow{p} 0, \tag{24}$$

then $\mathbb{E}[|S_n|^q] \to 0$ and by Markov, $S_n \xrightarrow{p} 0$.

Again, (24) is true due to law of large numbers.

Lemma C.1 (Uniform q-moment for ξ_c for any $q \geq 1$)**.** *Let $q \geq 1$ and suppose*

$$\sup_x \; \mathbb{E}\left[\, |\varepsilon|^{2q} \,\middle|\, X = x \right] \;\leq\; C_\varepsilon \;<\; \infty.$$

Then

$$\sup_c \; \mathbb{E}\left[\, |\varepsilon_c^2 - \sigma_c^2|^q \right] \;\leq\; C \;<\; \infty,$$

for a constant C depending only on q and C_ε.

Proof. Use the inequality valid for all $q \geq 1$: $|u - v|^q \leq 2^{q-1}\big(|u|^q + |v|^q\big)$ with $u = \varepsilon_c^2$, $v = \sigma_c^2$:

$$\mathbb{E}\left[|\varepsilon_c^2 - \sigma_c^2|^{\,q}\right] \;\leq\; 2^{q-1}\Big\{\mathbb{E}\left[|\varepsilon_c|^{\,2q}\right] + \mathbb{E}\left[(\sigma_c^2)^q\right]\Big\}.$$

For the second term, apply conditional Jensen with the convex map $x \mapsto x^q$:

$$(\sigma_c^2)^q = \big(\mathbb{E}[\varepsilon_c^2 \mid X_c]\big)^q \;\leq\; \mathbb{E}\left[\,|\varepsilon_c|^{\,2q} \mid X_c\right].$$

Taking expectations and using the uniform bound on the conditional $2q$-th moment,

$$\mathbb{E}\left[(\sigma_c^2)^q\right] \leq \mathbb{E}\left[\,\mathbb{E}(|\varepsilon_c|^{\,2q} \mid X_c)\right] \leq C_\varepsilon, \quad \text{and} \quad \mathbb{E}\left[|\varepsilon_c|^{\,2q}\right] \leq C_\varepsilon.$$

Thus $\mathbb{E}[|\varepsilon_c^2 - \sigma_c^2|^{\,q}] \leq 2^{q-1}(C_\varepsilon + C_\varepsilon) = 2^q C_\varepsilon$, uniformly in c. □

Lemma C.2 (Finite Moments of Matching Weights)**.** *Let $K(i)$ be the number of times control unit i is matched to units in the treated group. For controls,*

$$w_i = \sum_{t \in \mathcal{T}} w_{it} \;\leq\; K(i),$$

since $w_{it} \leq 1$ for each pair (i, t). Under the Exponential Tail Condition (Assumption 3), all moments of $K(i)$ are finite. Consequently, $\mathbb{E}[w_i^r] < \infty$ for all integers $r > 0$.

Proof. The bound $w_i \leq K(i)$ follows directly from the definition of matching weights, since each pairwise weight $w_{it} \leq 1$. The finiteness of all moments of $K(i)$ under the Exponential Tail Condition is established in the proof of Lemma 3 of Abadie and Imbens (2006) (p. 262). Since $w_i \leq K(i)$, we have $w_i^r \leq K(i)^r$ for all $r \geq 1$, and therefore $\mathbb{E}[w_i^r] \leq \mathbb{E}[K(i)^r] < \infty$. □

C.1.2 Second term 21a goes to zero:

$$\frac{1}{\sum_{c \in \mathcal{C}} K(c)} \sum_{c \in \mathcal{C}} \sum_{t \in \mathcal{T}_c} (\sigma_c^2 - \sigma_t^2) = \frac{1}{\sum_{c \in \mathcal{C}} K(c)} \sum_{c \in \mathcal{C}} K(c)(\sigma_c^2 - \bar{\sigma}_c^2) \tag{25}$$

where $\bar{\sigma}_c^2 = \frac{1}{K(c)} \sum_{t \in \mathcal{T}_c} \sigma_t^2$ is the average variance of the treated units matched to control unit c,

and $K(c) = |\mathcal{T}_c|$ represents the number of treated units to which control unit c is matched.

We can bound this term as follows:

$$\left| \frac{1}{\sum_{c\in\mathcal{C}} K(c)} \sum_{c\in\mathcal{C}} K(c)(\sigma_c^2 - \bar{\sigma}_c^2) \right| \leq \frac{1}{\sum_{c\in\mathcal{C}} K(c)} \sum_{c\in\mathcal{C}} K(c) \cdot \max_{c=1,\ldots,n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \tag{26}$$

$$= \max_{c=1,\ldots,n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \xrightarrow{\text{a.s.}} 0 \text{ as } n_c, n_T \to \infty \tag{27}$$

where the last convergence follows from Lemma C.3, which establishes the uniform convergence of variance differences across all control units.

Lemma C.3 (Uniform convergence of variances)**.** *Under Assumptions 3 and 4 (through the continuity condition in Definition 3.1), we have*

$$\max_{c=1,\ldots,n_c} \left|\sigma_c^2 - \bar{\sigma}_c^2\right| \xrightarrow{p} 0 \quad \textit{as } n_c, n_T \to \infty,$$

where $\sigma_c^2 = \sigma^2(X_c)$ *and* $\bar{\sigma}_c^2 = \frac{1}{K(c)} \sum_{t\in\mathcal{T}_c} \sigma^2(X_t)$*, with* $\mathcal{T}_c$ *the set of treated units matched to control* c*.*

Proof. Recall the matching radius for treated unit t:

$$r(\mathcal{C}_t) = \sup_{j\in\mathcal{C}_t} \|X_t - X_j\|.$$

Define the maximal (sample-wide) matching radius

$$r_{\max} := \max_{t\in\mathcal{T}} r(\mathcal{C}_t).$$

By Assumption 3, for any $u \geq 0$ and each treated t, $\Pr\big(n_C^{1/k} r(\mathcal{C}_t) > u\big) \leq C_1 e^{-C_2 u^k}$. A union bound over $t \in \mathcal{T}$ gives

$$\Pr\big(n_C^{1/k} r_{\max} > u\big) \;\leq\; n_T C_1 e^{-C_2 u^k}.$$

Fix $\varepsilon > 0$ and set $u = \varepsilon n_C^{1/k}$. Then $\Pr(r_{\max} > \varepsilon) \leq n_T C_1 e^{-C_2 \varepsilon^k n_C} \to 0$ as $n_C, n_T \to \infty$, hence $r_{\max} \xrightarrow{p} 0$.

By Assumption 4, $\sigma^2(\cdot)$ is Lipschitz: there exists $L < \infty$ such that $|\sigma^2(x) - \sigma^2(y)| \leq L\|x - y\|$

for all x, y. For any control c,

$$\begin{aligned}\left|\sigma_c^2 - \bar{\sigma}_c^2\right| &= \left|\sigma^2(X_c) - \frac{1}{K(c)} \sum_{t \in \mathcal{T}_c} \sigma^2(X_t)\right| \\ &\leq \frac{1}{K(c)} \sum_{t \in \mathcal{T}_c} \left|\sigma^2(X_c) - \sigma^2(X_t)\right| \\ &\leq \frac{L}{K(c)} \sum_{t \in \mathcal{T}_c} \|X_c - X_t\|.\end{aligned}$$

Each $t \in \mathcal{T}_c$ is a treated unit for which c was matched, so $\|X_c - X_t\| \leq r(\mathcal{C}_t) \leq r_{\max}$. Hence

$$\left|\sigma_c^2 - \bar{\sigma}_c^2\right| \;\leq\; L\, r_{\max} \quad \text{and thus} \quad \max_c \left|\sigma_c^2 - \bar{\sigma}_c^2\right| \;\leq\; L\, r_{\max}.$$

Since $r_{\max} \xrightarrow{p} 0$, the desired conclusion follows. □

C.1.3 Second component of Term A, (20c) goes to zero

For the second component of Term A (cross-product of errors):

$$\begin{aligned}&\sum_{t \in \mathcal{T}} \frac{u_t}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \left[-2\varepsilon_j \left(\frac{1}{|\mathcal{C}_t|} \sum_{\substack{k \in \mathcal{C}_t \\ k \neq j}} \varepsilon_k \right) \right] \\ =&\sum_{t \in \mathcal{T}} \frac{u_t}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \left[-2\varepsilon_j \frac{1}{|\mathcal{C}_t|} \sum_{\substack{k \in \mathcal{C}_t \\ k \neq j}} \varepsilon_k \right] \\ =&\sum_{t \in \mathcal{T}} \frac{1}{\sum_{t \in \mathcal{T}} |\mathcal{C}_t|} \frac{1}{|\mathcal{C}_t|} \sum_{\substack{j,k \in \mathcal{C}_t \\ j \neq k}} (-4\varepsilon_j \varepsilon_k) \\ \leq&\frac{1}{\sum_{t \in \mathcal{T}} |\mathcal{C}_t|} \sum_{\substack{j,k \in \mathcal{C} \\ j \neq k}} -4 \cdot \frac{K(j,k)}{2} \varepsilon_j \varepsilon_k \\ =&\frac{1}{\sum_{c \in \mathcal{C}} K(c)} \sum_{\substack{j,k \in \mathcal{C} \\ j \neq k}} -2 \cdot K(j,k) \varepsilon_j \varepsilon_k\end{aligned}$$

where $K(j,k)$ represents the number of times control units j and k appear together in the same matched cluster. Since $|\mathcal{C}_t| \geq 2$ for all clusters (as we exclude singleton clusters), we have $\frac{1}{|\mathcal{C}_t|} \leq \frac{1}{2}$,

which gives us the inequality in the last step.

To establish that this term converges to zero in probability, we apply a similar two–step proof argument as in the previous subsection (Section C.1.1). First, note that each cross–product has mean zero since $E[\varepsilon_j \varepsilon_k] = 0$ by independence of errors across units. Second, observe that the pairwise reuse count $K(j,k)$ is automatically controlled by the individual reuse counts, because two units can be matched together at most as many times as the less frequently used unit appears; formally, $K(j,k) \leq \min\{K(j), K(k)\}$. This ensures that the aggregate weight on cross–products is bounded in the same way as in the first–term analysis. Therefore, by applying the same second–moment condition on the errors, inequalities and the law of large numbers, we conclude that the variance of this cross–product sum vanishes, and hence the term converges to zero in probability.

C.1.4 Third component of Term A, (20d) goes to zero

For the third component of Term A (interaction between function values and errors):

$$(A3) = \sum_{t \in \mathcal{T}} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[-2 \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right) \left(\varepsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right) \right].$$

By the Mean Value Theorem and Assumption 5, we can bound the first factor:

$$\begin{aligned} \left| f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right| &\leq \max_{k \in \mathcal{C}_t} |f_0(X_j) - f_0(X_k)| \\ &\leq \sup_{j \in \mathcal{C}_t} |f_0'(X_j')| \cdot \max_{j,k \in \mathcal{C}_t} \|X_j - X_k\| \\ &\leq \sup_{j \in \mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t), \end{aligned}$$

where X_j' lies on the line segment between X_j and X_k.

Therefore:

$$
\begin{aligned}
|(A3)| &\leq \sum_{t\in\mathcal{T}} \frac{u_t}{|\mathcal{C}_t|-1} \sum_{j\in\mathcal{C}_t} 2\cdot \sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t) \cdot \left|\varepsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\varepsilon_k\right| \\
&\leq 2\cdot \sup_{t\in\mathcal{T}} \left[\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right] \cdot \sum_{t\in\mathcal{T}} \frac{u_t}{|\mathcal{C}_t|-1} \sum_{j\in\mathcal{C}_t} \left|\varepsilon_j - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\varepsilon_k\right| \\
&= 2\cdot \sup_{t\in\mathcal{T}} \left[\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right] \cdot \frac{1}{\sum_{t\in\mathcal{T}}|\mathcal{C}_t|} \sum_{c\in\mathcal{C}} K(c) \left|\varepsilon_c - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\varepsilon_k\right| \\
&= 2\cdot \sup_{t\in\mathcal{T}} \left[\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right] \cdot \frac{1}{\sum_{c\in\mathcal{C}}K(c)} \sum_{c\in\mathcal{C}} K(c) \left|\varepsilon_c - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\varepsilon_k\right|.
\end{aligned}
$$

The term

$$
\sup_{t\in\mathcal{T}} \left[\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right]
$$

goes to zero by the following lemma.

Lemma C.4 (Slope–radius product vanishes)**.** *Under Assumption 1, Assumption 5, and the exponential tail condition on matching radii (Assumption 3),*

$$
M_n \;:=\; \sup_{t\in\mathcal{T}} \;\left[\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right] \;\xrightarrow{p}\; 0.
$$

We can then show that the weighted error differences satisfy

$$
\frac{1}{\sum_{t\in\mathcal{T}}|\mathcal{C}_t|} \sum_{c\in\mathcal{C}} K(c) \left|\varepsilon_c - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t}\varepsilon_k\right| \xrightarrow{p} 0 \quad \text{as } n_T \to \infty,
$$

using arguments similar to those in Section C.1.1. Therefore, $(A3) \xrightarrow{p} 0$ as $n_T \to \infty$.

C.1.5 Fourth component of Term A, (20e) goes to zero

For the fourth and final component of Term A (systematic differences within clusters):

$$
(A4) = \sum_{t\in\mathcal{T}} \frac{u_t}{|\mathcal{C}_t|-1} \sum_{j\in\mathcal{C}_t} \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k)\right)^2.
$$

As in the analysis of (A3), we apply the Mean Value Theorem to bound each squared difference:

$$\begin{aligned}\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|}\sum_{k\in\mathcal{C}_t} f_0(X_k)\right)^2 &\le \left(\max_{k\in\mathcal{C}_t} |f_0(X_j) - f_0(X_k)|\right)^2 \\ &\le \left(\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot \max_{j,k\in\mathcal{C}_t} \|X_j - X_k\|\right)^2 \\ &\le \left(\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right)^2,\end{aligned}$$

where X_j' lies on the line segment between X_j and X_k.

Thus:

$$\begin{aligned}|(A4)| &\le \sum_{t\in\mathcal{T}} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j\in\mathcal{C}_t} \left(\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right)^2 \\ &= \sum_{t\in\mathcal{T}} \frac{u_t \cdot |\mathcal{C}_t|}{|\mathcal{C}_t| - 1} \left(\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right)^2 \\ &\le 2 \cdot \sum_{t\in\mathcal{T}} u_t \left(\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right)^2 \\ &\le 2 \cdot \left(\sup_{t\in\mathcal{T}} \left[\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right]\right)^2.\end{aligned}$$

By Lemma C.4, we have

$$\sup_{t\in\mathcal{T}} \left[\sup_{j\in\mathcal{C}_t} |f_0'(X_j')| \cdot r(\mathcal{C}_t)\right] = o_p(1).$$

Therefore, $(A4) \xrightarrow{p} 0$ as $n_T \to \infty$.

C.2 Term B

For Term B,

$$\text{Term B} = \sum_{t\in\mathcal{T}} \left(\frac{|\mathcal{C}_t|}{\sum_{t'\in\mathcal{T}} |\mathcal{C}_{t'}|} - \frac{1}{n_T}\right) \sigma_t^2 = \frac{\sum_{t\in\mathcal{T}} |\mathcal{C}_t|\, \sigma_t^2}{\sum_{t\in\mathcal{T}} |\mathcal{C}_t|} \; - \; \frac{1}{n_T} \sum_{t\in\mathcal{T}} \sigma_t^2.$$

Let v be the uniform distribution on $\mathcal{T}$, so for any sequence a_t, $\mathbb{E}_v[a_t] = \frac{1}{n_T}\sum_{t\in\mathcal{T}} a_t$. Then

$$\mathrm{Cov}_v\big(|\mathcal{C}_t|, \sigma_t^2\big) = \mathbb{E}_v\big[|\mathcal{C}_t|\,\sigma_t^2\big] - \mathbb{E}_v[|\mathcal{C}_t|]\;\mathbb{E}_v\big[\sigma_t^2\big] = \frac{1}{n_T}\sum_t |\mathcal{C}_t|\,\sigma_t^2 - \left(\frac{1}{n_T}\sum_t |\mathcal{C}_t|\right)\left(\frac{1}{n_T}\sum_t \sigma_t^2\right).$$

Dividing both sides by $\frac{1}{n_T}\sum_t |\mathcal{C}_t|$ gives

$$\frac{\mathrm{Cov}_v(|\mathcal{C}_t|, \sigma_t^2)}{\frac{1}{n_T}\sum_t |\mathcal{C}_t|} = \frac{\sum_t |\mathcal{C}_t|\,\sigma_t^2}{\sum_t |\mathcal{C}_t|} - \frac{1}{n_T}\sum_t \sigma_t^2 = \text{Term B}.$$

Hence,

$$\boxed{\text{Term B} = \frac{1}{\sum_{t\in\mathcal{T}} |\mathcal{C}_t|/n_T}\;\mathrm{Cov}_v\big(|\mathcal{C}_t|, \sigma_t^2\big)\,.}$$

D Proof of Lemma 4.3

Proof. Recall

$$V^*_{E,\lim} \;=\; \Big(\tfrac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2\Big)\Big(\tfrac{1}{n_T} + \tfrac{1}{\mathrm{ESS}(\mathcal{C})}\Big) = \frac{1}{n_T^2}\sum_{t\in\mathcal{T}}\sigma_t^2 + \frac{1}{n_T^2}\sum_{j\in\mathcal{C}} w_j^2\cdot\Big(\tfrac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2\Big),$$

and

$$V_E \;=\; \frac{1}{n_T^2}\sum_{t\in\mathcal{T}}\sigma_t^2 + \frac{1}{n_T^2}\sum_{j\in\mathcal{C}} w_j^2\sigma_j^2.$$

Hence

$$V^*_{E,\lim} - V_E = \frac{1}{n_T^2}\sum_{j\in\mathcal{C}} w_j^2\Big[\tfrac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2 - \sigma_j^2\Big].$$

Introduce the matched averages $\overline{\sigma_t^2} = \sum_{j\in\mathcal{C}_t} w_{jt}\sigma_j^2$. Decompose:

$$V^*_{E,\lim} - V_E = \underbrace{\tfrac{1}{n_T}\tfrac{1}{\mathrm{ESS}(\mathcal{C})}\sum_{t\in\mathcal{T}}(\sigma_t^2 - \overline{\sigma_t^2})}_{(I)} \;+\; \underbrace{\Big(\tfrac{1}{n_T}\tfrac{1}{\mathrm{ESS}(\mathcal{C})}\sum_{t\in\mathcal{T}}\overline{\sigma_t^2} - \tfrac{1}{n_T^2}\sum_{j\in\mathcal{C}} w_j^2\sigma_j^2\Big)}_{(II)}.$$

Term (I). By Regular Variance and Shrinking Clusters, $\frac{1}{n_T}\sum_t(\sigma_t^2 - \overline{\sigma_t^2}) \to 0$. By Lemma D.1,

$n_T/\mathrm{ESS}(\mathcal{C}) = O_p(1)$. Therefore

$$n_T \cdot (I) = \frac{n_T}{\mathrm{ESS}(\mathcal{C})} \cdot \frac{1}{n_T} \sum_t (\sigma_t^2 - \overline{\sigma_t^2}) \xrightarrow{p} 0.$$

Term (II). Compute

$$\frac{1}{n_T} \frac{1}{\mathrm{ESS}(\mathcal{C})} \sum_{t \in \mathcal{T}} \overline{\sigma_t^2} = \frac{1}{n_T^2} \cdot \frac{\sum_{j'} w_{j'}^2}{n_T} \sum_{j \in \mathcal{C}} w_j \sigma_j^2,$$

so

$$(II) = \frac{1}{n_T^2} \sum_{j \in \mathcal{C}} \Big(\frac{\sum_{j'} w_{j'}^2}{n_T} w_j - w_j^2 \Big) \sigma_j^2 = -\frac{1}{n_T} \mathrm{Cov}_p(w_j, \sigma_j^2),$$

using Lemma D.2.

Putting the pieces together,

$$n_T \Big(V_{E,\mathrm{lim}}^* - V_E + \tfrac{1}{n_T} \mathrm{Cov}_p(w_j, \sigma_j^2) \Big) = n_T \cdot (I) \xrightarrow{p} 0,$$

which yields the stated result. □

D.1 Bounded Ratio Lemma

Assumption 7 (Bounded maximum reuse)**.** *The maximum reuse count is bounded in probability:*

$$K_n := \max_{j \in \mathcal{C}} K(j) = O_p(1),$$

where $K(j) := \#\{t \in \mathcal{T} : w_{jt} > 0\}$.

Lemma D.1 (Bounded ratio $n_T/\mathrm{ESS}(\mathcal{C})$)**.** *Under Assumption 7,*

$$\frac{n_T}{\mathrm{ESS}(\mathcal{C})} \;=\; \frac{\sum_{j \in \mathcal{C}} w_j^2}{n_T} \;=\; O_p(1).$$

Proof. By definition,

$$\mathrm{ESS}(\mathcal{C}) = \frac{\Big(\sum_{j \in \mathcal{C}} w_j \Big)^2}{\sum_{j \in \mathcal{C}} w_j^2} = \frac{n_T^2}{\sum_{j \in \mathcal{C}} w_j^2},$$

so

$$\frac{n_T}{\mathrm{ESS}(\mathcal{C})} = \frac{\sum_{j\in\mathcal{C}} w_j^2}{n_T}.$$

Thus it suffices to show that $\sum_{j\in\mathcal{C}} w_j^2 = O_p(n_T)$.

For a given control $j \in \mathcal{C}$, write

$$w_j \;=\; \sum_{t\in\mathcal{T}} w_{jt}, \qquad K(j) := \#\{t \in \mathcal{T} : w_{jt} > 0\}.$$

By Cauchy–Schwarz,

$$w_j^2 = \Big(\sum_{t\in\mathcal{T}} w_{jt}\Big)^2 \;\le\; K(j) \sum_{t\in\mathcal{T}} w_{jt}^2.$$

Summing over all controls $j \in \mathcal{C}$,

$$\sum_{j\in\mathcal{C}} w_j^2 \;\le\; \sum_{j\in\mathcal{C}} K(j) \sum_{t\in\mathcal{T}} w_{jt}^2.$$

Now swap the order of summation:

$$\sum_{j\in\mathcal{C}} K(j) \sum_{t\in\mathcal{T}} w_{jt}^2 = \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{C}} K(j)\, w_{jt}^2.$$

Define $K_n := \max_{j\in\mathcal{C}} K(j)$. Then

$$\sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{C}} K(j)\, w_{jt}^2 \;\le\; K_n \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{C}_t} w_{jt}^2.$$

For each treated unit $t \in \mathcal{T}$, the weights satisfy $\sum_{j\in\mathcal{C}_t} w_{jt} = 1$. Hence

$$\sum_{j\in\mathcal{C}_t} w_{jt}^2 \le \Big(\sum_{j\in\mathcal{C}_t} w_{jt}\Big)^2 = 1.$$

Therefore,

$$\sum_{j\in\mathcal{C}} w_j^2 \;\le\; K_n \sum_{t\in\mathcal{T}} 1 = K_n\, n_T.$$

By Assumption 7, $K_n = O_p(1)$, so $\sum_{j\in\mathcal{C}} w_j^2 = O_p(n_T)$. It follows that

$$\frac{n_T}{\mathrm{ESS}(\mathcal{C})} = \frac{\sum_{j\in\mathcal{C}} w_j^2}{n_T} = O_p(1).$$

□

D.2 Covariance form of the heteroskedastic correction

Lemma D.2 (Covariance form of the heteroskedastic correction)**.** *With $p_j := w_j/n_T$ and $\mathrm{ESS}(\mathcal{C}) = n_T^2/\sum_j w_j^2$, the term*

$$T = \frac{1}{n_T^2}\sum_{j\in\mathcal{C}}\left(\frac{\sum_{j'} w_{j'}^2}{n_T} - w_j\right) w_j s_j^2 = -\frac{1}{n_T}\,\mathrm{Cov}_p\left(w_j, s_j^2\right).$$

Proof. Recall two facts:

- $\sum_{j\in\mathcal{C}} w_j = n_T$ (each treated contributes total weight 1 across its matched controls),
- The effective sample size $\mathrm{ESS}(\mathcal{C}) = \frac{\left(\sum_j w_j\right)^2}{\sum_j w_j^2} = \frac{n_T^2}{\sum_j w_j^2}$. Equivalently, $\sum_j w_j^2/n_T^2 = 1/\mathrm{ESS}(\mathcal{C})$.

Now set

$$p_j = \frac{w_j}{n_T} \quad \left(\text{so}\sum_j p_j = 1\right), \quad q_j = \frac{w_j^2}{\sum_\ell w_\ell^2} \quad \left(\text{so}\sum_j q_j = 1\right).$$

Then a few lines of algebra give

$$T = \frac{1}{\mathrm{ESS}}\left(\sum_j p_j s_j^2 - \sum_j q_j s_j^2\right) = \frac{1}{\mathrm{ESS}}\left(\mathbb{E}_p\left[s^2\right] - \mathbb{E}_q\left[s^2\right]\right)$$

Next relate $\mathbb{E}_q$ to $\mathbb{E}_p$. Because $q_j \propto w_j p_j$,

$$\mathbb{E}_q\left[s^2\right] = \frac{\mathrm{ESS}}{n_T}\mathbb{E}_p\left[ws^2\right] = \frac{\mathrm{ESS}}{n_T}\left(\mathrm{Cov}_p\left(w, s^2\right) + \mathbb{E}_p[w]\mathbb{E}_p\left[s^2\right]\right)$$

and since $\mathbb{E}_p[w] = \sum_j p_j w_j = \sum_j w_j^2/n_T = n_T/\mathrm{ESS}$,

$$\mathbb{E}_q\left[s^2\right] = \mathbb{E}_p\left[s^2\right] + \frac{\mathrm{ESS}}{n_T}\,\mathrm{Cov}_p\left(w, s^2\right)$$

Plugging back,

$$T = -\frac{1}{n_T}\,\mathrm{Cov}_p\left(w_j, s_j^2\right) \qquad \text{with } p_j = \frac{w_j}{n_T}$$

□

E Proof of Theorem 4.4

Proof. From Equation (9), write

$$\begin{aligned} n_T\big(\hat{V}_E^{alt} - V_E\big) &= n_T\Big(\tfrac{1}{n_T} + \tfrac{1}{\mathrm{ESS}(\mathcal{C})}\Big)\Big(S^2 - \tfrac{1}{\sum_{t\in\mathcal{T}}|\mathcal{C}_t|/n_T}\,\mathrm{Cov}_v(|\mathcal{C}_t|, \sigma_t^2) - \tfrac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2\Big) \\ &\quad + n_T\Big(\tfrac{1}{n_T} + \tfrac{1}{\mathrm{ESS}(\mathcal{C})}\Big)\Big(\tfrac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2 + \tfrac{1}{n_T}\mathrm{Cov}_p(w_j, \sigma_j^2)\Big) - n_T V_E \\ &\quad + \Big(\mathrm{Cov}_p(w_j, s_j^2) - \mathrm{Cov}_p(w_j, \sigma_j^2)\Big). \end{aligned} \tag{28}$$

Consider each line in (28):

1. First line. By Lemma 4.2, the inner parentheses converge to zero in probability. Moreover, $n_T\big(\frac{1}{n_T} + \frac{1}{\mathrm{ESS}(\mathcal{C})}\big) = O_p(1)$ by Lemma D.1. Hence the entire first line is $o_p(1)$.

2. Second line. By Lemma 4.3,

$$n_T\Big(\tfrac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2 + \tfrac{1}{n_T}\mathrm{Cov}_p(w_j, \sigma_j^2)\Big) - n_T V_E \;\xrightarrow{p}\; 0.$$

3. Third line. For the difference of covariances, expand

$$\mathrm{Cov}_p(w_j, s_j^2) - \mathrm{Cov}_p(w_j, \sigma_j^2) = \frac{1}{n_T}\sum_{j\in\mathcal{C}}(w_j - \bar{w})\, w_j\, (s_j^2 - \sigma_j^2),$$

where $\bar{w} = \frac{1}{n_T}\sum_{j\in\mathcal{C}} w_j^2$. Each s_j^2 is a consistent estimator of σ_j^2 within clusters (see the proof of Term A in Lemma 4.2), and the weights $\{w_j\}$ have bounded moments by Lemma C.2. Therefore

this term also vanishes in probability.

Combining all three parts shows that the right-hand side of (28) converges to zero in probability, proving the claim. □

F Proof of Theorem 4.5

Proof. We prove this by showing that each component of $\hat{V}$ converges in probability to the corresponding component of $V = n_T \cdot (V_E + V_P)$.

Step 1: Decomposition of the main term

First, we decompose the primary component of our estimator:

$$\frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{Y}_t(0) - \hat{\tau} \right)^2 = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{Y}_t(0) \right)^2 - \hat{\tau}^2$$

Step 2: Expansion using the structural model

Next, we expand $\frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{Y}_t(0) \right)^2$ using our structural assumptions. Recall that:

- $Y_t = f_1(X_t) + \epsilon_{1,t}$
- $\hat{Y}_t(0) = \sum_{j \in \mathcal{C}_t} w_{jt} Y_j = \sum_{j \in \mathcal{C}_t} w_{jt} (f_0(X_j) + \epsilon_{0,j})$

Therefore:

$$Y_t - \hat{Y}_t(0) = f_1(X_t) - \sum_{j \in \mathcal{C}_t} w_{jt} f_0(X_j) + \epsilon_{1,t} - \sum_{j \in \mathcal{C}_t} w_{jt} \epsilon_{0,j}$$

Expanding the squared term:

$$\begin{aligned}
&\frac{1}{n_T}\sum_{t\in\mathcal{T}}\left(Y_t-\hat{Y}_t(0)\right)^2 \\
&=\frac{1}{n_T}\sum_{t\in\mathcal{T}}\left[f_1(X_t)-\sum_{j\in\mathcal{C}_t}w_{jt}f_0(X_j)\right]^2 \quad \text{(Term I)} \\
&\quad+\frac{1}{n_T}\sum_{t\in\mathcal{T}}\left[\epsilon_{1,t}-\sum_{j\in\mathcal{C}_t}w_{jt}\epsilon_{0,j}\right]^2 \quad \text{(Term II)} \\
&\quad+\frac{2}{n_T}\sum_{t\in\mathcal{T}}\left[f_1(X_t)-\sum_{j\in\mathcal{C}_t}w_{jt}f_0(X_j)\right]\left[\epsilon_{1,t}-\sum_{j\in\mathcal{C}_t}w_{jt}\epsilon_{0,j}\right] \quad \text{(Term III)}
\end{aligned}$$

Step 2a: Analysis of Term I

By Assumptions 3 and **??**, we have that $\sum_{j\in\mathcal{C}_t}w_{jt}f_0(X_j)\to f_0(X_t)$ uniformly in t. Therefore:

$$\text{Term I}=\frac{1}{n_T}\sum_{t\in\mathcal{T}}\left[f_1(X_t)-f_0(X_t)\right]^2+o_p(1)=\frac{1}{n_T}\sum_{t\in\mathcal{T}}\tau(X_t)^2+o_p(1)$$

Step 2b: Analysis of Term II

Expanding Term II:

$$\begin{aligned}
\text{Term II}&=\frac{1}{n_T}\sum_{t\in\mathcal{T}}\left[\epsilon_{1,t}^2+\left(\sum_{j\in\mathcal{C}_t}w_{jt}\epsilon_{0,j}\right)^2-2\epsilon_{1,t}\sum_{j\in\mathcal{C}_t}w_{jt}\epsilon_{0,j}\right] \\
&=\frac{1}{n_T}\sum_{t\in\mathcal{T}}\epsilon_{1,t}^2+\frac{1}{n_T}\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{C}_t}\sum_{j'\in\mathcal{C}_t}w_{jt}w_{j't}\epsilon_{0,j}\epsilon_{0,j'}-\frac{2}{n_T}\sum_{t\in\mathcal{T}}\epsilon_{1,t}\sum_{j\in\mathcal{C}_t}w_{jt}\epsilon_{0,j}
\end{aligned}$$

Under Assumption 4, we have $E[\epsilon_{1,t}^2|X_t]=\sigma_{1,t}^2$ and $E[\epsilon_{0,j}^2|X_j]=\sigma_{0,j}^2$. By the law of large numbers and independence of errors:

$$\begin{aligned}
\text{Term II}&\xrightarrow{p}\frac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_{1,t}^2+\frac{1}{n_T}\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{C}_t}w_{jt}^2\sigma_{0,j}^2 \\
&=\frac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_{1,t}^2+\frac{1}{n_T}\sum_{j\in\mathcal{C}}\left(\sum_{t'\in\mathcal{T}}w_{jt'}^2\right)\sigma_{0,j}^2
\end{aligned}$$

Step 2c: Analysis of Term III

Term III involves cross-products between systematic and error components. Since errors have conditional mean zero and are independent of covariates, by the law of large numbers:

$$\text{Term III} \xrightarrow{p} 0$$

Step 3: Combining Terms I and the $\hat{\tau}^2$ correction

From Step 2a, we have:

$$\text{Term I} - \hat{\tau}^2 = \frac{1}{n_T}\sum_{t\in\mathcal{T}} \tau(X_t)^2 - \hat{\tau}^2 + o_p(1)$$

Recall $\tau_{SATT} = \frac{1}{n_T}\sum_{t\in\mathcal{T}} \tau(X_t)$ be the sample average treatment effect on the treated. Then:

$$\frac{1}{n_T}\sum_{t\in\mathcal{T}} \tau(X_t)^2 - \hat{\tau}^2 = \frac{1}{n_T}\sum_{t\in\mathcal{T}} \tau(X_t)^2 - \tau_{SATT}^2 + \tau_{SATT}^2 - \hat{\tau}^2$$

Since $\hat{\tau} \xrightarrow{p} \tau_{SATT}$ (consistency of the matching estimator), we have $\tau_{SATT}^2 - \hat{\tau}^2 \xrightarrow{p} 0$.

Therefore:

$$\text{Term I} - \hat{\tau}^2 \xrightarrow{p} \frac{1}{n_T}\sum_{t\in\mathcal{T}} \tau(X_t)^2 - \tau_{SATT}^2 = \frac{1}{n_T}\sum_{t\in\mathcal{T}} (\tau(X_t) - \tau_{SATT})^2$$

By the law of large numbers, as $n_T \to \infty$:

$$\frac{1}{n_T}\sum_{t\in\mathcal{T}} (\tau(X_t) - \tau_{SATT})^2 \xrightarrow{p} E[(\tau(X) - \tau)^2 | Z = 1] = n_T V_P$$

Step 4: Analysis of the correction term

The correction term in $\hat{V}$ is:

$$S^2 \frac{1}{n_T}\left[\sum_{j\in\mathcal{C}}\left[\left(\sum_{t'\in\mathcal{T}} w_{jt'}\right)^2 - \left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\right]\right]$$

By Lemma 4.2, $S^2 \xrightarrow{p} \frac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2$. Under Assumption 6, $\sigma_{1,t}^2 = \sigma_{0,j}^2 = \sigma_t^2$.

The bracketed term converges to:

$$\frac{1}{n_T}\left[\sum_{j\in\mathcal{C}}\left[\left(\sum_{t'\in\mathcal{T}} w_{jt'}\right)^2 - \left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\right]\right] \xrightarrow{p} \frac{1}{n_T}\sum_{j\in\mathcal{C}}\left(w_j^2 - \sum_{t'\in\mathcal{T}} w_{jt'}^2\right)$$

where $w_j = \sum_{t'\in\mathcal{T}} w_{jt'}$.

Step 5: Final assembly

Combining all components:

$$\begin{aligned}
\hat{V} &= \frac{1}{n_T}\sum_{t\in\mathcal{T}}\left(Y_t - \hat{Y}_t(0) - \hat{\tau}\right)^2 + S^2\frac{1}{n_T}\left[\sum_{j\in\mathcal{C}}\left[\left(\sum_{t'\in\mathcal{T}} w_{jt'}\right)^2 - \left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\right]\right] \\
&\xrightarrow{p} n_T V_P + \frac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_{1,t}^2 + \frac{1}{n_T}\sum_{j\in\mathcal{C}}\left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\sigma_{0,j}^2 \\
&\quad + \frac{1}{n_T}\sum_{t\in\mathcal{T}}\sigma_t^2 \cdot \frac{1}{n_T}\sum_{j\in\mathcal{C}}\left(w_j^2 - \sum_{t'\in\mathcal{T}} w_{jt'}^2\right) \\
&= n_T V_P + n_T V_E \\
&= V
\end{aligned}$$

The last equality follows from the definition of V_E and algebraic manipulation using the homoskedasticity assumption.

Therefore, $|\hat{V} - V| \xrightarrow{p} 0$ as $n_T \to \infty$. □

G Compare the Lipschitz Condition to that in the Existing Litarature

In the existing literature, the function $f(x)$ is often assumed to be locally Lipschitz on any compact set $\mathcal{X} \subset \mathbb{R}$. This implies that for any compact set $\mathcal{X} = [a, b]$, there exists a constant $L_{\mathcal{X}} < \infty$ such that:

$$|f(x) - f(y)| \leq L_{\mathcal{X}}|x - y|, \quad \forall x, y \in \mathcal{X}.$$

For example, consider $f(x) = x^2$, where the derivative $f'(x) = 2x$. On $\mathcal{X} = [0, 100]$, the Lipschitz constant is:

$$L_{\mathcal{X}} = 2 \cdot \max_{x \in \mathcal{X}} |x| = 200.$$

This large constant makes the bound impractical in matching-based inference, where overly conservative bounds can restrict the formation of matched sets.

In contrast, our Derivative Control condition improves on the Lipschitz assumption by explicitly tying the slope of $f(x)$ to the size of the matched set. Specifically, it requires:

$$\sup_{x \in \mathcal{C}_t} \left| f'(x) \right| \cdot \text{radius}(\mathcal{C}_t) \leq M,$$

where:

- $\mathcal{C}_t$ is the matched set for a given t,
- $\text{radius}(\mathcal{C}_t)$ is the diameter of the matched set in x-space,
- M is a universal constant independent of the matched set size.

This condition offers several practical advantages:

1. Localized Control: Instead of requiring a single large Lipschitz constant $L_{\mathcal{X}}$ over a wide range, our condition focuses on smaller, localized matched sets.
2. Adaptive Bounds: When the derivative $f'(x)$ is large, our condition naturally enforces smaller matched set radii to maintain practical bounds. For instance:

$$\text{If } f'(x) = 100 \text{ (as for } x = 50), \text{ then } \text{radius}(\mathcal{C}_t) \leq \frac{M}{100}.$$

3. Real-World Applicability: In real-world matching problems, matched sets are typically small, and our condition aligns with this reality by providing sharper, more practical bounds than the overly conservative Lipschitz constant.

To summarize, while the Lipschitz assumption is valid on compact sets, the associated constants $L_{\mathcal{X}}$ can become impractically large for functions like $f(x) = x^2$ over wide intervals. By explicitly

accounting for both the derivative and the size of matched sets, our condition provides a more precise and practical framework for matching-based inference.

H Comparison with Theorem 1 of White (1980)

Our theorem, stated as Theorem 4.5, differs from Theorem 1 of White (1980) in several key aspects. While both results address consistency in variance estimation under heteroskedasticity, the differences lie in the frameworks, assumptions, and proof strategies.

H.1 Parametric vs. Nonparametric Framework

White's Theorem 1 is based on a regression model $Y_i = X_i\beta_0 + \varepsilon_i$, where ε_i represents independent but non-identically distributed (i.n.i.d.) errors. The parametric form $X_i\beta_0$ is central, and β_0 is estimated via ordinary least squares (OLS). Heteroskedasticity arises through $\mathrm{Var}(\varepsilon_i \mid X_i) = g(X_i)$, where $g(X_i)$ is a known (possibly parametric) function. In contrast, our theorem relies on a nonparametric matching estimator for treatment effects, without assuming a parametric form for $f(X_i)$. Matching is governed by hyperparameters like the number of neighbors or the maximum matching radius, but these are not estimated from the data in the regression sense. Heteroskedasticity arises through $\sigma^2(X_i)$, where $\sigma^2(\cdot)$ is a uniformly continuous function.

H.2 White's Setup: Estimating $\mathrm{Var}(\hat{\beta})$ vs. Cluster-Based Variance Estimation

White's Theorem 1 focuses on the heteroskedasticity-consistent (HC) covariance matrix estimator for $\hat{\beta}$. It defines the matrix

$$\hat{V}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 X_i' X_i, \quad \text{where} \quad \hat{\varepsilon}_i = Y_i - X_i\hat{\beta}.$$

White proves $\hat{V}_n \xrightarrow{\text{a.s.}} \bar{V}_n$, where $\bar{V}_n$ is the asymptotic covariance matrix of the regressors. Our theorem, on the other hand, defines cluster-level residual variance estimators s_t^2 for each treated

unit $t \in \mathcal{T}$, given its matched controls $\mathcal{C}_t$. The overall variance estimator is

$$S^2 = \frac{1}{n_T} \sum_{t \in \mathcal{T}} s_t^2, \quad \text{where} \quad s_t^2 = \frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} e_{tj}^2.$$

We prove $|S^2 - \frac{1}{n_T} \sum_{t \in \mathcal{T}} \sigma_t^2| \xrightarrow{\text{a.s.}} 0$, showing consistency for the average cluster variance.

H.3 Homoskedasticity in Matched Clusters vs. General Heteroskedasticity

White's Theorem 1 allows general heteroskedasticity: $\text{Var}(\varepsilon_i \mid X_i) = g(X_i)$, where $g(\cdot)$ can vary arbitrarily across observations. Errors are independent but not identically distributed (i.n.i.d.). Our theorem also allows heteroskedasticity: $\sigma^2(X_i)$ varies with X_i. However, within each matched cluster $\{t\} \cup \mathcal{C}_t$, we assume $\sigma_j^2 \approx \sigma_t^2$ for $j \in \mathcal{C}_t$, based on a uniform continuity (or Lipschitz) assumption on $\sigma^2(\cdot)$.

H.4 Proof Strategy and Key Assumptions

White's proof strategy relies on expanding $\hat{V}_n - \bar{V}_n$ and showing that

$$\hat{V}_n - \bar{V}_n = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\varepsilon}_i^2 X_i' X_i - E[\varepsilon_i^2 X_i' X_i] \right) \xrightarrow{\text{a.s.}} 0.$$

White uses assumptions on finite moments of ε_i and X_i (Assumptions 2–4 in White (1980)) and uniform integrability conditions. Our proof, in contrast, relies on showing that for matched clusters $\{t\} \cup \mathcal{C}_t$, the residual variance s_t^2 converges to the true variance σ_t^2. We leverage uniform continuity of $\sigma^2(\cdot)$ to argue that $\sigma_j^2 \to \sigma_t^2$ as $\|X_{tj} - X_t\| \to 0$. We then apply a version of the Law of Large Numbers (LLN) for matched clusters.

H.5 Summary of Differences

The key differences between White's theorem and our theorem can be summarized as follows. First, White's theorem is regression-based, while our theorem is matching-based. Second, White

assumes a parametric model $Y_i = X_i\beta_0 + \varepsilon_i$, whereas our model assumes a nonparametric $f_1(X)$, $f_0(X)$. Third, White's focus is on a robust covariance estimator for $\hat{\beta}$, while ours is on residual variance from matched clusters. Fourth, White allows fully general $g(X_i)$, whereas our clusters assume approximate homoskedasticity ($\sigma_j^2 \approx \sigma_t^2$). Finally, White's framework has no matching hyperparameters, while ours depends on predefined criteria for matching (e.g., number of neighbors or radius).

I Otsu and Rai Variance Estimator

I.0.1 Debiasing Method

A debiasing model estimates the conditional mean function $\mu(z, x) = E[Y \mid Z = z, X = x]$. It is used to offset the bias to achieve valid inference (see Section 3.4 for discussion of the issue). The debiased estimator is defined as:

$$\tilde{\tau}(w) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{\mu}(0, X_t) - \sum_{j \in \mathcal{C}_t} w_{jt}(Y_j - \hat{\mu}(0, X_j)) \right) \tag{29}$$

Additional implementation details include:

- **Model Choice**: Linear model
- **Training Data**: Control data only
- **Cross-fitting**: Implemented by dividing the control data into two halves

I.0.2 Variance Estimators

Bootstrap Variance Estimator.

- Step 1: Use data with $Z_i = 0$ to construct $\hat{\mu}(0, x) = \hat{E}[Y|Z = 0, X = x]$.
- Step 2: Construct debiased estimate for each treated unit $t \in \mathcal{T}$:

$$\tilde{\tau}_t = (Y_t - \hat{\mu}(0, X_t)) - \sum_{j \in \mathcal{C}_t} w_{jt}(Y_j - \hat{\mu}(0, X_j))$$

- Step 3: Construct the debiased estimator: $\tilde{\tau} = \frac{1}{n_t} \sum_{t \in \mathcal{T}} \tilde{\tau}_t$
- Step 4: Construct the debiased residuals $R_t = \tilde{\tau}_t - \tilde{\tau}$
- Step 5: Perform Wild bootstrap on $\{R_t\}$ with special sampling weights
- Step 6: Construct confidence interval from bootstrap distribution

J Other Simulation Results

J.1 Detailed Figures on CI Length

See Figure 3. For the Otsu-Rai DGP, our method produces confidence intervals with an average width of 0.092 compared to 0.057 for the bootstrap method. On average, the confidence interval length under our method is about 1.64 times larger than that under the bootstrap method across all sample sizes, covariate dimensions, and curve IDs. For the Che et al. DGP, the CI length under our method is about 1.06 times larger than the bootstrap CI length. The bootstrap method's narrower intervals are artificially optimistic due to its failure to account for the true sampling variability induced by control unit dependencies.

J.2 Additional Simulation Results of the Che et al. (2024) DGP

This section provides supplementary simulation results that further validate our theoretical framework. We examine three key aspects: the accuracy of our V_E component estimation, verification of asymptotic bias patterns, and the behavior of effective sample sizes across different overlap scenarios.

Table 5: Additional Simulation Results: Variance Components and Bias Analysis

Degree of Overlap	**True** SE_E	**Est.** SE_E	**Coverage Rate**	**Coverage w/o Bias Corr.**	**Mean ESS**$_C$	**Mean** V/n_T
Very Low	0.183	0.184	95.0%	92.3%	8.41	0.130
Low	0.160	0.163	94.6%	92.4%	11.26	0.122
Medium	0.145	0.148	94.0%	93.0%	14.03	0.117
High	0.133	0.136	94.4%	93.6%	16.96	0.112
Very High	0.125	0.129	94.4%	92.4%	19.64	0.110

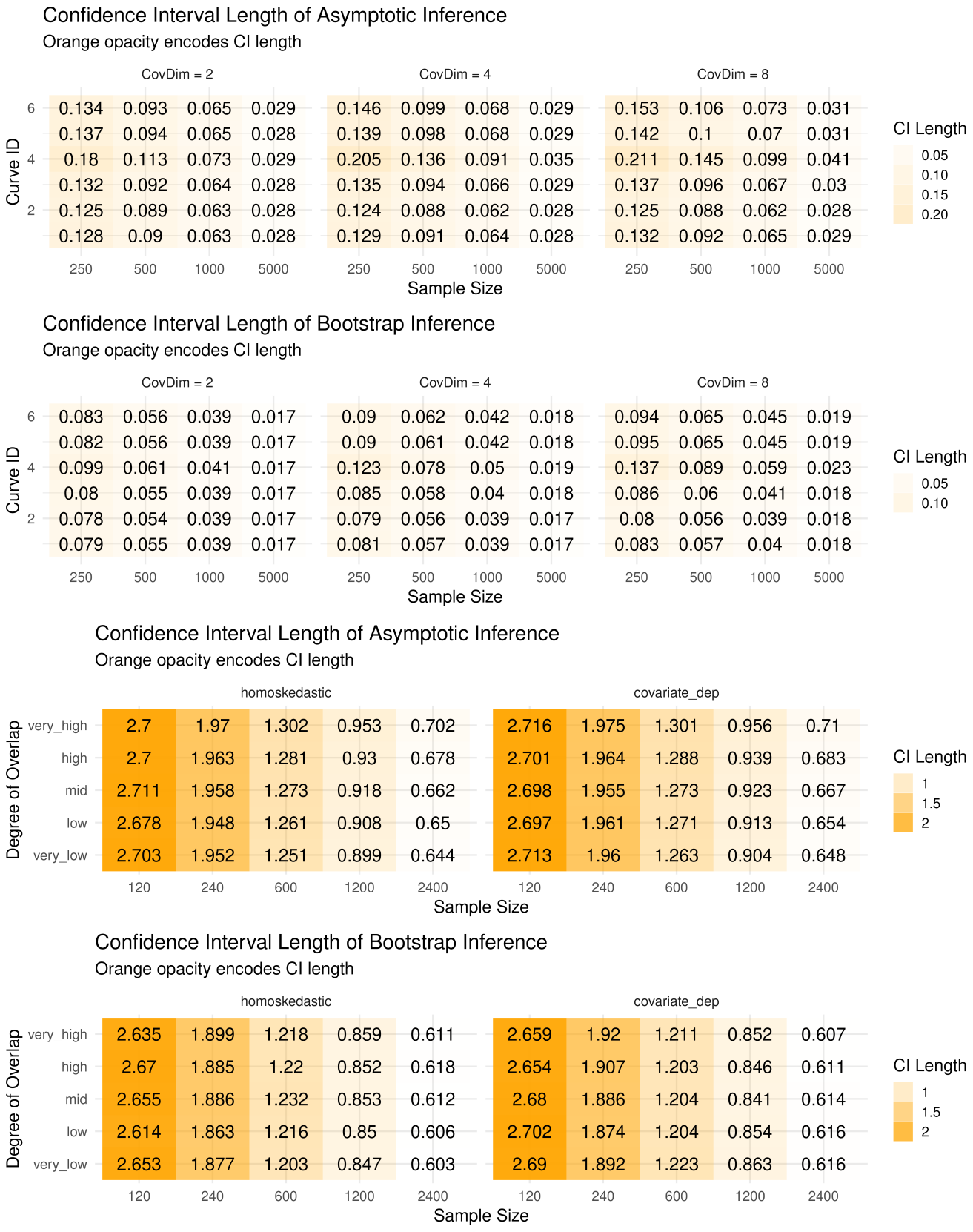

Figure 3: Confidence interval lengths, with orange opacity encoding interval width. Top: results for the Otsu-Rai data generating process across varying covariate dimensions ($K = 2, 4, 8$), sample sizes ($n = 250, 500, 1000, 5000$), and nonlinear outcome functions (curves 1–6). Bottom: results for the Che et al. data generating process across varying degrees of population overlap and two error variance structures.

J.2.1 Variance Component Estimation

Our estimator demonstrates excellent performance in estimating the V_E component, which captures the measurement error variance from residual outcome noise. Table 5 shows the close correspondence between the true SE_E (computed as the standard deviation of $\hat{\tau} - \text{SATT}$ across simulations) and our estimated SE_E values across all overlap scenarios. The differences are minimal, ranging from 0.001 to 0.004, indicating that our pooled variance estimator accurately captures this component of the total variance.

This accuracy is particularly important because the V_E component reflects how matching structure affects variance through control unit reuse. Unlike the bootstrap method, which does not decompose variance into interpretable components, our approach allows researchers to understand how different aspects of matching contribute to overall uncertainty.

J.2.2 Asymptotic Bias Verification

The simulation results provide clear evidence of asymptotic bias as predicted by our theoretical propositions. Comparing coverage rates with and without bias correction demonstrates the importance of the bias correction term B_n. Across all overlap scenarios, coverage without bias correction is systematically lower than with bias correction:

- Very Low overlap: 92.3% vs 95.0% (difference of 2.7 percentage points)
- Low overlap: 92.4% vs 94.6% (difference of 2.2 percentage points)
- Medium overlap: 93.0% vs 94.0% (difference of 1.0 percentage points)
- High overlap: 93.6% vs 94.4% (difference of 0.8 percentage points)
- Very High overlap: 92.4% vs 94.4% (difference of 2.0 percentage points)

This pattern confirms that bias correction is essential for achieving proper coverage, particularly in low-overlap scenarios where matching quality is poorer and bias is more substantial.

J.2.3 Effective Sample Size Analysis

The effective sample size of controls (ESS_C) shows an intuitive increasing pattern with the degree of overlap, ranging from 8.41 in very low overlap scenarios to 19.64 in very high overlap scenarios. This trend reflects that higher overlap allows for more efficient use of the control sample, as each control unit can contribute meaningfully to multiple matches without dramatically inflating variance through excessive reuse.

The mean V/n_T values (representing the estimated total variance scaled by sample size) show a corresponding decreasing pattern as overlap increases, from 0.130 to 0.110. This demonstrates that better overlap not only improves bias (through closer matches) but also reduces variance (through more efficient control utilization), confirming the bias-variance tradeoff in matching estimators discussed in the theoretical sections.