Cross Matching for Flexible Design of Observational Studies Based on Learned Scores

Yifan Zhao[1] and Ying Jin[2]

[1]University of Science and Technology of China
[2]Department of Statistics and Data Science, University of Pennsylvania

Work In Progress; November 26, 2026

Abstract

The use of data-adaptive balancing information has advanced weighting-based causal inference, but its integration into nearest-neighbor matching raises fundamental inferential questions. In particular, when propensity and prognostic scores are learned flexibly, what conditions ensure that their joint use leads to reliable matching-based inference? We address this problem through *Cross Matching*, a cross-fitted double-score matching framework. We establish a product condition governing score approximation error and characterize how matching dimension determines the need for regression adjustment. Under suitable conditions, the resulting estimators are asymptotically normal and semiparametrically efficient. Simulations and a real-data application demonstrate the practical relevance of these results.

1 Introduction

1.1 Background

Matching is a classical approach to causal inference in observational studies. By constructing comparable treated and control groups using pretreatment covariates, matching aims to reduce confounding without relying entirely on an outcome model (Rubin, 1973; Rosenbaum and Rubin, 1985; Rosenbaum, 2020). Among various matching procedures, nearest-neighbor (NN) matching is particularly appealing for its simplicity and transparency. However, its statistical properties depend strongly on the dimension and geometry of the variables used for matching. In particular, matching directly on a large set of covariates can leave substantial matching discrepancies and lead to non-negligible bias (Abadie and Imbens, 2006; Viel et al., 2025).

A natural way to address this difficulty is to match on lower-dimensional balancing scores. The propensity score $e(X) = \mathbb{P}(T = 1 \mid X)$ balances the covariate distributions across treatment groups (Rosenbaum and Rubin, 1983), while the prognostic score $m_0(X) = \mathbb{E}[Y(0) \mid X]$ summarizes covariate information relevant to the control outcome (Hansen, 2008). Their joint use, which we refer to as double-score matching (DSM), can improve adjustment by incorporating information about both treatment assignment and the outcome (Rubin and Thomas, 2000; Leacy and Stuart, 2014). This motivates the use of multiple candidate propensity and prognostic scores as a flexible, low-dimensional representation of the covariates.

1.2 Related Work

Existing theoretical work on DSM has established important robustness properties. Doubly robust matching estimators and extensions to multiple candidate score models have also derived large-sample results that account for nuisance parameter estimation (Antonelli et al., 2018; Yang and Zhang, 2023). These results provide a foundation for combining propensity and prognostic information, but their guarantees are based on correctly specified candidate models. In applications with many covariates or nonlinear relationships,

however, it can be difficult to specify suitable parametric models in advance. Flexible learners such as random forests and sieve estimators offer an alternative, but estimating the scores and matching on those same observations can make the matching features overly adapted to the sample.

Cross-fitting provides a natural way to separate these two steps. Sample splitting has been used to construct data-informed balancing features for subsequent weighting and analysis (Jin and Zubizarreta, 2025). This framework motivates the separation of feature construction and estimation, but does not study nearest-neighbor matching on learned double-score vectors. This distinction matters because matching generally achieves only local, rather than exact, balance.

A second relevant line of work studies the large-sample behavior of nearest-neighbor matching. Classical results establish the large-sample theory for matching with a fixed number of neighbors and its bias-corrected version (Abadie and Imbens, 2006, 2011). Allowing the number of matches to diverge yields a matching-based density-ratio estimator and can establish semiparametric efficiency under suitable smoothness and outcome-model conditions (Lin et al., 2023). Further results characterize how the geometry of the support and higher-order regularity affect matching rates (Viel et al., 2025). These results clarify the role of matching dimension and the number of neighbors, but do not directly cover matching on score vectors learned from flexible candidate libraries.

Recent work also illustrates the practical importance of this issue. For example, a recent survey-weighted double-score framework uses one propensity score and one prognostic score, corresponding to a two-dimensional matching representation, while focusing on the placement of survey weights and population-level inference (Zeng et al., 2026). The question of how a larger collection of learned scores affects matching and inference therefore remains relevant. More generally, recent results on inexact matching highlight that low-dimensional matching features do not by themselves guarantee negligible matching error (Jin, 2026).

These developments leave two questions for double-score matching with flexible learners:

1. **How can flexible score learning be incorporated into matching?** A practical method should allow data-adaptive propensity and prognostic estimators, while separating the construction of matching features from the observations used for matching.

2. **When does matching on learned scores permit valid inference?** In addition to nuisance approximation error, nearest-neighbor matching has a dimension-dependent local discrepancy. It is therefore necessary to determine when this discrepancy is negligible without adjustment, and when regression adjustment is needed.

1.3 Our Contributions

We propose *Cross Matching*, a cross-fitted double-score matching procedure, and study its consistency and asymptotic distribution for the average treatment effect on the treated (ATT). Our contributions are threefold:

1. **A cross-fitted framework for flexible double-score matching.** Cross Matching estimates candidate propensity and prognostic scores on one fold and performs nearest-neighbor matching on the other, then reverses the roles of the folds. The method accommodates flexible parametric and nonparametric learners while avoiding the direct reuse of the same observations for score construction and matching. A regression-adjustment function, when used, is also fitted on the training fold.

2. **Large-sample theory under score approximation.** We establish consistency and asymptotic normality without requiring an exactly correctly specified candidate score model. The nuisance approximation errors enter through a product condition, while the matching discrepancy is controlled separately through conditions on the learned score space. This provides an inference theory for cross-fitted double-score nearest-neighbor matching with flexible score estimators.

3. **Dimension-dependent inference and efficiency.** We show that raw matching admits root-n inference under our conditions when the score dimension is $d = 2$ or $d = 3$. For $d \geq 4$, our sufficient

conditions require regression adjustment to remove the leading matching discrepancy. By allowing the number of matches M_n to diverge, the resulting estimators attain the semiparametric efficiency bound under the corresponding regularity and rate conditions. We also examine a cross-fitted pairwise bootstrap in simulations; its finite-sample performance is studied separately from the asymptotic normality results.

The rest of the paper is organized as follows. Section 2 introduces the Cross Matching procedure and its assumptions. Section 3 establishes consistency, asymptotic normality, and semiparametric efficiency. Section 4 reports simulation studies, and Section 5 illustrates the method in an empirical application. Section 6 concludes.

2 Cross Matching

2.1 Problem Setup and Notation

Let $O_i = (X_i, T_i, Y_i)$, $i = 1, \dots, n$, be independent and identically distributed observations, where $X_i \in \mathcal{X} \subseteq \mathbb{R}^p$ denotes pretreatment covariates, $T_i \in \{0, 1\}$ is the treatment indicator, and $Y_i \in \mathbb{R}$ is the observed outcome. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control, respectively. We assume the stable unit treatment value assumption (SUTVA), so that $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$.

Define the propensity score by $e(x) := \mathbb{P}(T = 1 \mid X = x)$ and the control prognostic function by $m_0(x) := \mathbb{E}[Y(0) \mid X = x]$. Our target is the average treatment effect on the treated, $\tau_{\mathrm{ATT}} := \mathbb{E}[Y(1) - Y(0) \mid T = 1]$. Writing $\theta_0 := \mathbb{E}[Y(0) \mid T = 1]$ and $\pi_a := \mathbb{P}(T = a)$ for $a \in \{0, 1\}$, we have $\tau_{\mathrm{ATT}} = \mathbb{E}[Y \mid T = 1] - \theta_0$. Under unconfoundedness, the covariate density ratio of the treated group relative to the control group is $r_X(x) := \frac{dP_{X|T=1}}{dP_{X|T=0}}(x) = \frac{\pi_0 e(x)}{\pi_1 \{1 - e(x)\}}$, and hence $\theta_0 = \mathbb{E}[m_0(X) \mid T = 1] = \mathbb{E}[r_X(X) Y \mid T = 0]$.

For $a \in \{0, 1\}$, write $\mathbb{E}_a[f] := \mathbb{E}[f \mid T = a]$ and $P_a f := \mathbb{E}_a[f]$. We use $\| \cdot \|_2$ for the Euclidean norm. The notation $o_p(a_n)$ and $O_p(a_n)$ denotes convergence to zero and boundedness in probability, respectively, after division by a_n; $o(a_n)$ and $O(a_n)$ have their usual deterministic meanings. We write $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$, $a_n \lesssim b_n$ if $a_n \le C b_n$ for a constant C independent of n, and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The notation $\to_p$ and $\rightsquigarrow$ denotes convergence in probability and in distribution, respectively.

2.2 Cross Matching Estimator

We use two-fold cross-fitting. Let $\mathcal{I} := \{1, \dots, n\}$ be the sample index set and randomly partition it into two disjoint folds $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$ of equal size. For $k \in \{1, 2\}$, let $\mathcal{I}^{(-k)} := \mathcal{I}^{(3-k)}$ denote the training fold and $\mathcal{I}^{(k)}$ the evaluation fold. For $a \in \{0, 1\}$, define $\mathcal{I}_a := \{i \in \mathcal{I} : T_i = a\}$, $\mathcal{I}_a^{(k)} := \mathcal{I}_a \cap \mathcal{I}^{(k)}$, and $\mathcal{I}_a^{(-k)} := \mathcal{I}_a \cap \mathcal{I}^{(-k)}$. Let $n^{(k)} := |\mathcal{I}^{(k)}|$ and $n_a^{(k)} := |\mathcal{I}_a^{(k)}|$.

For each fold k, all nuisance models are estimated using only the training observations in $\mathcal{I}^{(-k)}$. Let $\{\widehat{e}_j^{(-k)} : 1 \le j \le J\}$ and $\{\widehat{m}_{0,\ell}^{(-k)} : 1 \le \ell \le L\}$ be candidate propensity and prognostic score estimators, respectively, where J and L are fixed positive integers. Candidate propensity estimates are clipped to a fixed interval $[c_e, 1 - c_e]$, with $0 < c_e < \eta$, before applying the logit transformation $\mathrm{logit}(u) := \log\{u/(1 - u)\}$. Let $\mathcal{A}^{(-k)}$ denote the coordinate-wise centering and scaling transformation estimated from the training fold. The learned score map $\widehat{\boldsymbol{s}}^{(k)} : \mathcal{X} \to \mathbb{R}^d$, with $d := J + L$, is

$$\widehat{\boldsymbol{s}}^{(k)}(x) = \mathcal{A}^{(-k)} \left(\mathrm{logit}\, \widehat{e}_1^{(-k)}(x), \dots, \mathrm{logit}\, \widehat{e}_J^{(-k)}(x); \widehat{m}_{0,1}^{(-k)}(x), \dots, \widehat{m}_{0,L}^{(-k)}(x) \right).$$

Thus, conditional on the training-fold sigma-field $\mathcal{F}_{\mathrm{tr}}^{(k)} := \sigma\{O_i : i \in \mathcal{I}^{(-k)}\}$, the map $\widehat{\boldsymbol{s}}^{(k)}$ is fixed. We distinguish this map from its evaluation-fold values $\widehat{\boldsymbol{S}}_i^{(k)} := \widehat{\boldsymbol{s}}^{(k)}(X_i)$ and from the support of the corresponding conditional score distribution. Specifically, write $\boldsymbol{S}^{(k)} := \widehat{\boldsymbol{s}}^{(k)}(X)$ and let $\mathcal{S}_a^{(k)} := \mathrm{supp}(\boldsymbol{S}^{(k)} \mid T = a, \mathcal{F}_{\mathrm{tr}}^{(k)})$. Under overlap, the two conditional covariate distributions are mutually absolutely continuous, so their images under the same score map have the same support. We therefore write their common support as $\mathcal{S}^{(k)}$.

The matching assignments in fold k are determined by the training information and the evaluation-fold score values and treatment indicators. We denote the corresponding sigma-field by

$$\mathcal{G}^{(k)} := \sigma\left(\mathcal{F}_{\mathrm{tr}}^{(k)}, \{(\widehat{\boldsymbol{S}}_i^{(k)}, T_i) : i \in \mathcal{I}^{(k)}\}\right).$$

In particular, $\mathcal{G}^{(k)}$ does not include the evaluation-fold covariates or outcomes beyond their score values and treatment indicators.

For a function h of covariates or score values, let $\|h\|_{0,2}^{(k)} := \{\mathbb{E}_0[h^2 \mid \mathcal{F}_{\mathrm{tr}}^{(k)}]\}^{1/2}$, with the argument of h understood from context. Define the best candidate approximation errors in fold k by $\epsilon_{e,n}^{(k)} := \min_{1\le j\le J} \|\widehat{e}_j^{(-k)} - e\|_{0,2}^{(k)}$ and $\epsilon_{m,n}^{(k)} := \min_{1\le \ell\le L} \|\widehat{m}_{0,\ell}^{(-k)} - m_0\|_{0,2}^{(k)}$. Their fold-uniform versions are $\epsilon_{e,n} := \max_{k=1,2} \epsilon_{e,n}^{(k)}$ and $\epsilon_{m,n} := \max_{k=1,2} \epsilon_{m,n}^{(k)}$. We do not impose separate convergence conditions on these two errors here; the product and dimension-dependent rate conditions are stated where they are needed in the asymptotic results.

For a function f of an observation, define the evaluation-fold empirical average $P_{a,n}^{(k)} f := \{n_a^{(k)}\}^{-1} \sum_{i\in\mathcal{I}_a^{(k)}} f(O_i)$. For each treated observation $i \in \mathcal{I}_1^{(k)}$, let $\mathcal{J}_{M_n}^{(k)}(i) \subseteq \mathcal{I}_0^{(k)}$ be the indices of its M_n nearest controls in the learned score space, where M_n is a positive integer that may depend on n. Matching is performed with replacement. For a control observation $j \in \mathcal{I}_0^{(k)}$, define its reuse count by $K_j^{(k)} := \sum_{i\in\mathcal{I}_1^{(k)}} \mathbf{1}\{j \in \mathcal{J}_{M_n}^{(k)}(i)\}$, and the matching-induced density-ratio estimate by $\widehat{r}_{M_n}^{(k)}(X_j) := \{n_0^{(k)}/(M_n n_1^{(k)})\}K_j^{(k)}$. The fold-specific raw estimator of the counterfactual mean is

$$\widehat{\theta}_0^{(k)} = \frac{1}{n_1^{(k)}} \sum_{i\in\mathcal{I}_1^{(k)}} \frac{1}{M_n} \sum_{j\in\mathcal{J}_{M_n}^{(k)}(i)} Y_j = P_{0,n}^{(k)}\{\widehat{r}_{M_n}^{(k)} Y\}.$$

Matching on the learned scores reduces the dimension of the matching variables, but does not generally eliminate the remaining discrepancy between treated observations and their matched controls. We therefore also consider a regression-adjusted version. Define the score-space regression function by $\mu_S^{(k)}(s) := \mathbb{E}_0[m_0(X) \mid \boldsymbol{S}^{(k)} = s, \mathcal{F}_{\mathrm{tr}}^{(k)}]$. We estimate it by a regression learner $\widehat{g}^{(k)}$, fitted on the training-fold controls $\mathcal{I}_0^{(-k)}$ using their outcomes and learned score values as response and covariates. The learner may be parametric or nonparametric, but is trained using the training fold only. The fold-specific regression-adjusted estimator is

$$\widehat{\theta}_{0,\mathrm{RA}}^{(k)} = P_{1,n}^{(k)}\{\widehat{g}^{(k)}(\widehat{\boldsymbol{S}})\} + P_{0,n}^{(k)}\left[\widehat{r}_{M_n}^{(k)}\{Y - \widehat{g}^{(k)}(\widehat{\boldsymbol{S}})\}\right].$$

The final two-fold ATT estimators are

$$\widehat{\tau}_{\mathrm{ATT}} = \frac{1}{2}\sum_{k=1}^{2}\left\{P_{1,n}^{(k)} Y - \widehat{\theta}_0^{(k)}\right\}, \qquad \widehat{\tau}_{\mathrm{ATT,RA}} = \frac{1}{2}\sum_{k=1}^{2}\left\{P_{1,n}^{(k)} Y - \widehat{\theta}_{0,\mathrm{RA}}^{(k)}\right\}.$$

For completeness, let P denote the distribution of $O = (X, T, Y)$ and P_n its full-sample empirical measure, so that $P_n f = n^{-1}\sum_{i=1}^n f(O_i)$. The empirical process is denoted by $\mathbb{G}_n := \sqrt{n}(P_n - P)$. For a measurable set $A \subseteq \mathbb{R}^d$, define $\mathrm{dist}(s, A) := \inf_{u\in A} \|s - u\|_2$, and let $B(s, r) := \{u \in \mathbb{R}^d : \|u - s\|_2 \le r\}$.

2.3 Assumptions

We impose the following basic identification, moment, and matching regularity conditions.

Assumption 2.1(Identification, overlap, and moments). (a) Treatment assignment is unconfounded given the pretreatment covariates, that is, $T \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$. (b) There exists $\eta \in (0, 1/2)$ such that $\eta \le e(X) \le 1 - \eta$ almost surely. (c) $\mathbb{E}[Y(1)^2] < \infty$ and there exists a constant $C_Y < \infty$ such that $\mathbb{E}[Y(0)^2 \mid X] \le C_Y$ almost surely.

For each fold k, the fold proportions satisfy $n^{(k)} = n/2$ and $n_a^{(k)}/n^{(k)} \to_p \pi_a$ for $a \in \{0, 1\}$.

Assumption 2.2(Score-space regularity). Conditional on $\mathcal{F}_{\mathrm{tr}}^{(k)}$, the common support $\mathcal{S}^{(k)}$ is compact and the distribution of $\boldsymbol{S}^{(k)}$ given $T = a$ has a density $f_a^{(k)}$ with respect to d-dimensional Lebesgue measure,

for $a \in \{0,1\}$. The densities satisfy the following conditions uniformly over the two folds: (a) There exist constants $0 < c_f < C_f < \infty$ such that $c_f \le f_a^{(k)}(s) \le C_f$ for all $s \in \mathcal{S}^{(k)}$ and $a \in \{0,1\}$. (b) Both $f_0^{(k)}$ and $f_1^{(k)}$ are Lipschitz continuous on $\mathcal{S}^{(k)}$, with a common Lipschitz constant. (c) The support has uniformly non-vanishing local volume, and satisfies the boundary condition

$$\inf_{\substack{s\in\mathcal{S}^{(k)}\\ 0<r\le \mathrm{diam}(\mathcal{S}^{(k)})}} \frac{\mathrm{Leb}_d\{B(s,r)\cap \mathcal{S}^{(k)}\}}{\mathrm{Leb}_d\{B(s,r)\}} > 0, \qquad \sup_{\lambda>0} \lambda^{1/d} \int_{\mathcal{S}^{(k)}} \exp\{-\lambda \delta_{\mathcal{S}^{(k)}}(s)^d\} f_1^{(k)}(s)\, ds < \infty,$$

where Leb_d is d-dimensional Lebesgue measure, $\mathrm{diam}(\mathcal{S}^{(k)})$ is the diameter of the support, and $\delta_{\mathcal{S}^{(k)}}(s) := \mathrm{dist}(s, (\mathcal{S}^{(k)})^c)$. The boundary condition is stated with respect to the treated score density because treated observations are the target points in the matching procedure.

Assumption 2.3(Diverging number of matches). The number of matched controls M_n satisfies, for each $k \in \{1,2\}$, $M_n \log n_0^{(k)}/n_0^{(k)} \to_p 0$ and $M_n/\log n_0^{(k)} \to_p \infty$.

Assumption 2.1 gives the standard identification and moment conditions. Assumption 2.2 imposes regularity on the distributions of the learned scores, rather than on the original covariates; its density and support conditions are of the type used in nearest-neighbor matching theory (Abadie and Imbens, 2006; Viel et al., 2025). The diverging-neighbor perspective allows the reuse counts to provide an estimate of the density ratio between the treated and control groups in the matching space (Lin et al., 2023). The conditions on score approximation and on the regression learner are not imposed as blanket assumptions here. They enter the results in Section 3 through the product $\epsilon_{e,n}\epsilon_{m,n}$ and, depending on the matching dimension, through additional rates for raw matching or regression adjustment.

3 Theoretical Results

3.1 Main Results

We first establish the consistency and asymptotic normality of the fold-specific counterfactual mean estimators. Throughout this section, $\varepsilon := Y - m_0(X)$ denotes the control outcome residual, and all results are understood for either evaluation fold $k \in \{1,2\}$.

Theorem 3.1 (Consistency and asymptotic normality)**.** *Suppose Assumptions 2.1–2.3 hold. Then, for each evaluation fold k, the following results hold.*

(a) ***Consistency.*** *If $\epsilon_{e,n}\epsilon_{m,n} = o_p(1)$ and $\epsilon_{m,n} = O_p(1)$, then $\widehat{\theta}_0^{(k)} \to_p \theta_0$.*

(b) ***Asymptotic normality.*** *Suppose $\epsilon_{e,n} = o_p(1)$ and $\epsilon_{e,n}\epsilon_{m,n} = o_p(n^{-1/2})$.*

(i) ***Raw matching.*** *If $d \in \{2,3\}$, assume in addition that*

$$\begin{cases} M_n^2/n \to 0, & d=2,\\ M_n^4/n \to 0, & d=3, \end{cases} \qquad (M_n/n)^{1/d}\epsilon_{m,n} = o_p(n^{-1/2}).$$

Then

$$\sqrt{n^{(k)}}\{\widehat{\theta}_0^{(k)} - \theta_0\} \rightsquigarrow N(0, V_0).$$

(ii) ***Regression-adjusted matching.*** *If $d \ge 4$, assume that $\max_{k=1,2} \|m_0 - \mu_S^{(k)}\|_{0,2}^{(k)} = o_p(1)$ and*

$$(M_n/n)^{1/d} \max_{k=1,2} \|\mu_S^{(k)} - \widehat{g}^{(k)}\|_{0,2}^{(k)} = o_p(n^{-1/2}).$$

Then

$$\sqrt{n^{(k)}}\{\widehat{\theta}_{0,\mathrm{RA}}^{(k)} - \theta_0\} \rightsquigarrow N(0, V_0).$$

Here

$$V_0 = \frac{\mathrm{Var}\{m_0(X) \mid T=1\}}{\pi_1} + \frac{\mathbb{E}_0[r_X(X)^2\varepsilon^2]}{\pi_0}.$$

The consistency result in Theorem 3.1(a) allows either score family to contain a consistent candidate, provided that the approximation error of the other family remains bounded. More generally, it only requires the product of the two best approximation errors to vanish. We refer to this property as *multiple consistency*. Unlike the exact-model conditions in existing multiply robust matching procedures, it does not require a candidate score model to be correctly specified at every sample size. It is important, however, to distinguish consistency from inference: the product condition at rate $o_p(1)$ does not by itself guarantee root-n normality.

The two normality regimes in part (b) reflect the dimension dependence of nearest-neighbor matching. For $d = 2$ and $d = 3$, raw matching is sufficient under the stated restrictions on M_n and the prognostic-score approximation error. In particular, the three-dimensional result goes beyond the usual implementation using one propensity score and one prognostic score. For $d \geq 4$, the leading local matching discrepancy need not be negligible at the root-n scale. The regression-adjusted estimator instead requires the score-space regression learner to approximate $\mu_S^{(k)}$ sufficiently accurately. In both regimes, the product condition controls the error due to estimating the matching scores, while the additional conditions control the matching discrepancy.

The theorem also shows the role of allowing M_n to diverge. Following the density-ratio representation of nearest-neighbor matching in Lin et al. (2023), the reuse counts estimate the relative density of treated and control observations in the score space. Under our conditions, this leads to the same first-order limit in both regimes, rather than an additional variance component from a fixed number of matches.

Corollary 3.2 (Asymptotic normality and efficiency of the ATT)**.** *Under Theorem 3.1(b)(i), the raw two-fold estimator*

$$\widehat{\tau}_{\mathrm{ATT}} = \frac{1}{2}\sum_{k=1}^{2}\left\{P_{1,n}^{(k)}Y - \widehat{\theta}_0^{(k)}\right\}$$

satisfies

$$\sqrt{n}(\widehat{\tau}_{\mathrm{ATT}} - \tau_{\mathrm{ATT}}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_\tau(O_i) + o_p(1).$$

Under Theorem 3.1(b)(ii), the same result holds for the regression-adjusted estimator

$$\widehat{\tau}_{\mathrm{ATT,RA}} = \frac{1}{2}\sum_{k=1}^{2}\left\{P_{1,n}^{(k)}Y - \widehat{\theta}_{0,\mathrm{RA}}^{(k)}\right\}.$$

In both cases, the influence function is

$$\phi_\tau(O) = \frac{T}{\pi_1}\{Y - m_0(X) - \tau_{\mathrm{ATT}}\} - \frac{1-T}{\pi_0}r_X(X)\{Y - m_0(X)\},$$

and the asymptotic variance is $V_{\mathrm{ATT}} = \mathbb{E}[\phi_\tau(O)^2]$.

To see why the two-fold estimator has this influence function, Theorem 3.1 gives, in either normality regime,

$$\widehat{\theta}_{0,\bullet}^{(k)} - \theta_0 = (P_{1,n}^{(k)} - P_1)m_0 + P_{0,n}^{(k)}(r_X\varepsilon) + o_p\{(n^{(k)})^{-1/2}\},$$

where $\bullet$ denotes the corresponding raw or regression-adjusted estimator. Subtracting this expansion from the treated sample mean yields the fold-specific ATT expansion. Since the two evaluation folds partition the full sample and have equal sizes, averaging the two expansions gives the full-sample representation in Corollary 3.2; in particular, no additional factor of two appears.

For completeness, let $m_1(x) := \mathbb{E}[Y(1) \mid X = x]$, $\tau(x) := m_1(x) - m_0(x)$, and $\sigma_a^2(x) := \mathrm{Var}\{Y(a) \mid X = x\}$ for $a \in \{0, 1\}$. The variance in the corollary can equivalently be written as

$$V_{\mathrm{ATT}} = \frac{1}{\pi_1^2}\mathbb{E}\left[e(X)\left\{\sigma_1^2(X) + (\tau(X) - \tau_{\mathrm{ATT}})^2\right\} + \frac{e(X)^2}{1-e(X)}\sigma_0^2(X)\right].$$

This is the semiparametric efficiency bound for the ATT in the nonparametric unconfoundedness model (Hahn, 1998). Thus, both versions of Cross Matching attain the same efficiency bound in their respective dimension regimes.

3.2 Dimension and Practical Implications

The distinction between raw and regression-adjusted matching is not merely a technical feature of the proof. For a smooth score-space regression function, the leading local matching discrepancy is generally of second order in the nearest-neighbor radius, which is of order $(M_n/n)^{1/d}$. Its root-n order is therefore governed by $\sqrt{n}(M_n/n)^{2/d}$. This term can be controlled under appropriate restrictions when $d = 2$ or $d = 3$, but it need not vanish when $d \geq 4$ and $M_n \to \infty$. The following example shows that this issue can persist even when the score space contains the exact propensity and prognostic information.

Proposition 3.3 (A four-dimensional oracle-score example)**.** *There exists a data-generating process satisfying the identification and score-space regularity conditions of Section 2 for which $\epsilon_{e,n} = 0$, $m_0(X)$ is measurable with respect to the oracle score vector, and $\epsilon_{e,n}\epsilon_{m,n} = 0$, but the raw matching estimator has a non-negligible root-n bias when $d = 4$ and $M_n \to \infty$.*

Construction and argument. Let $E \sim \text{Unif}[1/4, 3/4]$, Z_1 have density $q_\gamma(z) = c_\gamma \exp(\gamma z)$ on $[-1, 1]$ for some $\gamma > 0$, and let Z_2, Z_3 be independent $\text{Unif}[-1, 1]$ variables, all mutually independent. Here c_γ is the normalizing constant. Set $X = (E, Z_1, Z_2, Z_3)$ and $\mathbb{P}(T = 1 \mid X) = e(X) = E$, so that $\pi_1 = \pi_0 = 1/2$. Let $Y(0) = h(Z_1) + \varepsilon$, where $h(z) = z - z^3/3$, $\mathbb{E}[\varepsilon \mid X, T] = 0$, and $\mathbb{E}[\varepsilon^2] < \infty$.

Consider the four-dimensional score vector $\boldsymbol{S} = (\text{logit}(E), Z_1, Z_2, Z_3)$. It contains the oracle propensity score and a one-to-one representation of the oracle prognostic score $m_0(X) = h(Z_1)$; the remaining two coordinates are additional candidate features. In particular, $m_0(X)$ is exactly measurable with respect to $\boldsymbol{S}$, so that $\mu_S(\boldsymbol{S}) = m_0(X)$. The propensity score approximation error is zero. The score support is a compact rectangle, and its treated and control densities are bounded, Lipschitz, and bounded away from zero.

For notational simplicity, we express the score coordinates in their standardized population units. The sample standardization used by Cross Matching converges to a fixed nonsingular coordinate-wise affine transformation. Such a transformation changes the leading bias constant through the induced local metric, but does not change its order or the non-vanishing property established below.

Write $s = (t, z_1, z_2, z_3)$, $e(t) = \text{expit}(t)$, and let f_1 and f_0 denote the treated and control score densities. A direct change of variables gives $f_1(s) = e(t)^2\{1-e(t)\}q_\gamma(z_1)$ and $f_0(s) = e(t)\{1-e(t)\}^2 q_\gamma(z_1)$. Consequently, $\partial_{z_1} \log f_0(s) = \gamma$.

For an interior treated score s, let $R_j(s)$ be the distance to its jth nearest control score and let Δ_j denote the corresponding displacement. The local nearest-neighbor expansion gives $\mathbb{E}[\Delta_{j,1} \mid R_j, s] = \gamma R_j^2/6 + o(R_j^2)$ and $\mathbb{E}[\Delta_{j,1}^2 \mid R_j, s] = R_j^2/6 + o(R_j^2)$. Since $h'(z) = 1 - z^2$ and $h''(z) = -2z$, a Taylor expansion yields

$$\mathbb{E}[h(z_1 + \Delta_{j,1}) - h(z_1) \mid R_j, s] = \frac{R_j^2}{6}\left\{\gamma h'(z_1) + \frac{1}{2}h''(z_1)\right\} + o(R_j^2).$$

Moreover, with $v_4 = \pi^2/2$ the volume of the unit ball in $\mathbb{R}^4$, the average squared radius of the first M_n neighbors satisfies

$$\frac{1}{M_n}\sum_{j=1}^{M_n}\mathbb{E}[R_j(s)^2 \mid s] = \frac{2}{3}\left\{\frac{M_n}{n_0 v_4 f_0(s)}\right\}^{1/2}\{1 + o(1)\}.$$

Averaging over the treated score distribution therefore gives

$$\mathbb{E}[\widehat{\theta}_0^{(k)} - \theta_0] = C_\gamma\left(\frac{M_n}{n_0^{(k)}}\right)^{1/2} + o\left\{\left(\frac{M_n}{n_0^{(k)}}\right)^{1/2}\right\},$$

where

$$C_\gamma = \frac{1}{9\sqrt{v_4}}\int_{\mathcal{S}}\frac{f_1(s)}{\sqrt{f_0(s)}}\left\{\gamma h'(z_1) + \frac{1}{2}h''(z_1)\right\}ds > 0.$$

Indeed, the integral over the z_1 coordinate is proportional to

$$\int_{-1}^{1} \sqrt{q_\gamma(z)} \left\{ \gamma h'(z) + \frac{1}{2} h''(z) \right\} dz = \frac{3\gamma}{4} \int_{-1}^{1} \sqrt{q_\gamma(z)} h'(z)\, dz$$
$$> 0,$$

where integration by parts uses $h'(-1) = h'(1) = 0$. The boundary strips contribute only a smaller-order term: the outcome regression is flat to first order at the two z_1 endpoints, and the other faces of the rectangular support are orthogonal to the z_1 direction.

It follows that $\sqrt{n^{(k)}}\, \mathbb{E}[\widehat{\theta}_0^{(k)} - \theta_0] \asymp \sqrt{M_n} \to \infty$. Thus the raw estimator cannot have a centered root-n limit in this example. In contrast, the oracle regression adjustment $\widehat{g}^{(k)} = \mu_S^{(k)}$ removes this leading local discrepancy. This establishes that the need for regression adjustment in dimension four is not caused by score estimation error: it can arise solely from the geometry of matching. □

Proposition 3.3 clarifies the role of the dimension restriction in Theorem 3.1.It shows that score accuracy and even oracle prognostic information do not, by themselves, guarantee root-n inference in four dimensions. Regression adjustment provides a direct way to remove this source of bias.

The results also have implications for the construction of the candidate library. Propensity scores and prognostic scores capture complementary features: the former describe treatment assignment, while the latter describe outcome risk. Their joint use can therefore improve the quality of matches relative to using either type of score alone. This does not mean that adding more candidates is always beneficial. A larger library can reduce $\epsilon_{e,n}$ or $\epsilon_{m,n}$, but it also increases the matching dimension d and can worsen the local matching discrepancy. In practice, the library should therefore be chosen to balance score approximation and matching dimension, rather than expanded indiscriminately. The simulations in Section 4 examine this trade-off directly.

Cross-fitting is what makes this use of flexible learners tractable. Each score map is estimated on a training fold and then treated as fixed when matching on the complementary fold. This separation accommodates data-adaptive learners while avoiding the direct use of the same observations to construct and match on the scores.

Finally, the choice of M_n is relevant to both efficiency and resampling inference. The ordinary nonparametric bootstrap can fail for fixed-neighbor matching, even when the estimator is root-n consistent and asymptotically normal (Abadie and Imbens, 2008). Our use of a diverging M_n leads to a different asymptotic regime and, under Theorem 3.1, to an asymptotically linear ATT estimator. This provides a basis for studying resampling procedures. Recent results obtain bootstrap validity for other matching designs, including one-to-one matching without replacement in low dimensions (Jin, 2026), and a linearization-based bootstrap for survey-weighted double-score matching (Zeng et al., 2026). These procedures have different matching and sampling structures from Cross Matching. We therefore examine a cross-fitted pairwise bootstrap in Section 4, while leaving a formal bootstrap validity theorem for the present estimator to future work.

4 Simulation Studies

We use two-fold cross-fitting throughout: candidate scores are estimated on one fold, and matching is conducted on the other. The fold-specific estimates are then averaged, with $M_n = \lceil n^{1/10} \rceil$.

Data-generating process. For the score-library simulations, let $X_1, \ldots, X_{10}$ be independent Unif$[-1, 1]$ variables and define

$$Z_j = \frac{f_j(X_j) - \mu_j}{s_j}, \qquad j = 1, \ldots, 10,$$

where μ_j and s_j are the mean and standard deviation of $f_j(U)$ for $U \sim \text{Unif}[-1, 1]$. We use

$$(f_1, \ldots, f_{10})(x) = (\sin(\pi x), x^2, |x|, \tanh(2x), \cos(\pi x),$$
$$x^3, \exp(x), \text{expit}(3x), \log(x + 1.5), \sin(2x)).$$

For $p > 10$, the additional covariates are independent noise variables. Set

$$\boldsymbol{\alpha}_0 = 0.4(1,1,1,0,0,-1,-1,0,0,0)^\top, \quad \boldsymbol{\beta}_0 = 0.5(1,1,1,1,1,-1,-1,-1,-1,-1)^\top.$$

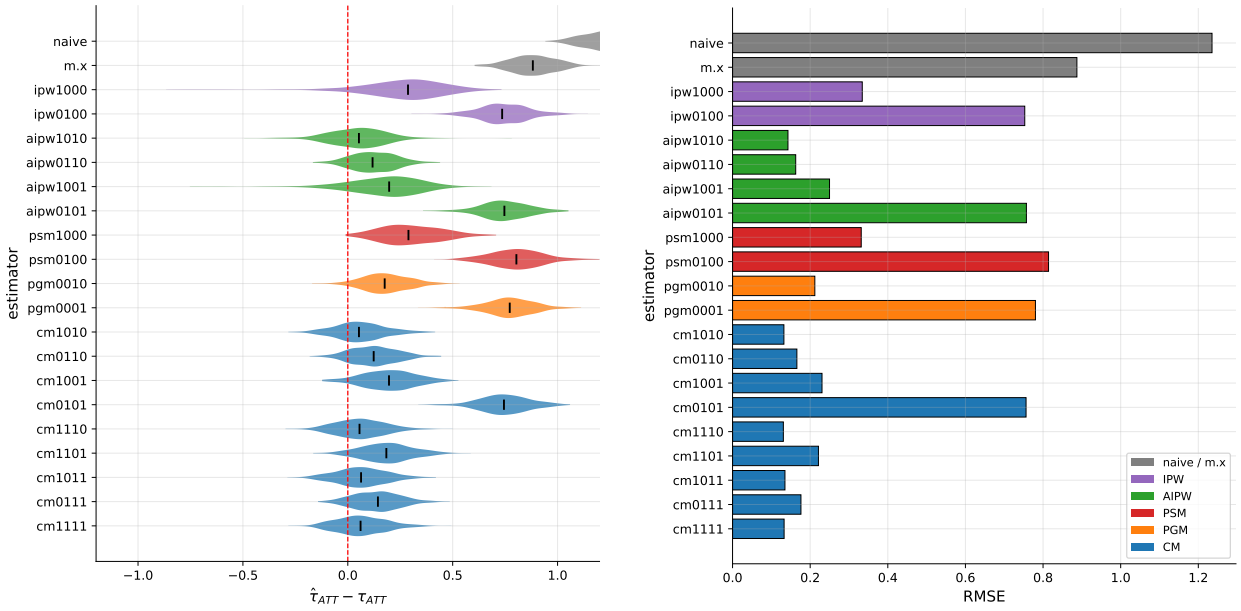
The common design is

$$\begin{aligned} e(X) &= \operatorname{expit}\{\psi \boldsymbol{\alpha}_0^\top Z\}, & m_0(X) &= \boldsymbol{\beta}_0^\top Z, \\ T \mid X &\sim \text{Bernoulli}\{e(X)\}, & Y(0) &= m_0(X) + \varepsilon_0, \\ Y(1) &= m_0(X) + \varepsilon_1, & \varepsilon_0, \varepsilon_1 &\overset{\text{iid}}{\sim} N(0,1), \end{aligned}$$

with errors independent of (X, T). Thus, the treatment effect is zero. We truncate the propensity score to $[10^{-3}, 1 - 10^{-3}]$ numerically and set $\psi = 1.75$ for the finite-sample comparison and $\psi = 1$ otherwise.

The candidate library consists of a flexible propensity estimator $\widehat{e}_1$ and prognostic estimator $\widehat{m}_{0,1}$, both based on additive cubic B-splines of the ten active covariates, and linear working models $\widehat{e}_2$ and $\widehat{m}_{0,2}$ based on the raw covariates. A four-digit code indicates which candidates enter the score, in the order $(\widehat{e}_1, \widehat{e}_2, \widehat{m}_{0,1}, \widehat{m}_{0,2})$. For example, $(1,1,1,0)$ refers to $\{\widehat{e}_1, \widehat{e}_2, \widehat{m}_{0,1}\}$. The spline specifications use fixed knots and penalties, so these simulations illustrate finite-sample behavior rather than a growing-sieve consistency argument.

Finite-sample comparison. We compare Cross Matching with the difference in means, covariate matching, IPW, AIPW, propensity-score matching, and prognostic-score matching, using $n = 1{,}000$ and 500 replications. Figure 1 shows that Cross Matching specifications incorporating flexible candidates have substantially smaller errors. In particular, *cm*1010, *cm*1110, *cm*1011 and *cm*1111 are among the best-performing specifications and has smallest RMSE. The results emphasize the importance of containing both flexible candidate propensity scores and prognostic scores for matching, and Cross Matching achieves the strongest robustness among the methods because it allows multiple candidate models in the library. Performance is guaranteed whenever a score model well estimated.



(a) Estimation-error distributions.

(b) RMSE.

Figure 1: Finite-sample comparison under nonlinear confounding ($n = 1{,}000$, 500 replications, $\psi = 1.75$).

Multiple consistency. Figure 2 reports the absolute bias over sample sizes for all nonempty candidate combinations. Bias decreases when at least one flexible candidate is included, whether it comes from the

propensity or prognostic library; it remains large when both are omitted. Using flexible candidates from both libraries generally gives the fastest decay. This pattern agrees with the product structure in Theorem 3.1(a): accuracy in either nuisance component can compensate for error in the other.

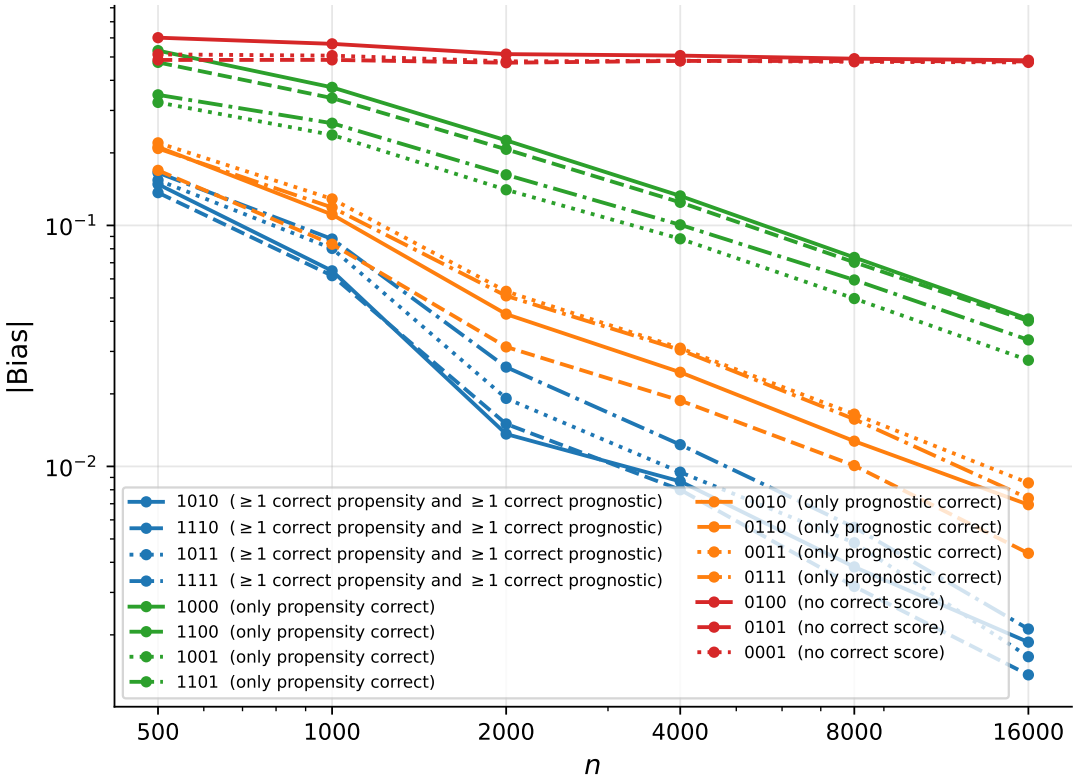
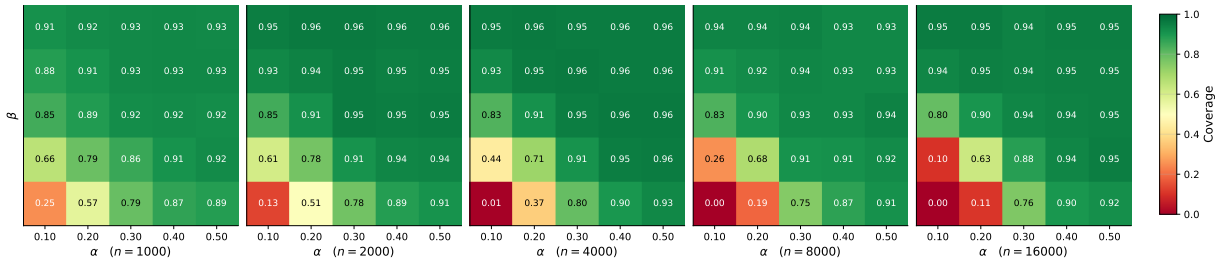


Figure 2: Absolute Monte Carlo bias across sample sizes and candidate-score combinations.

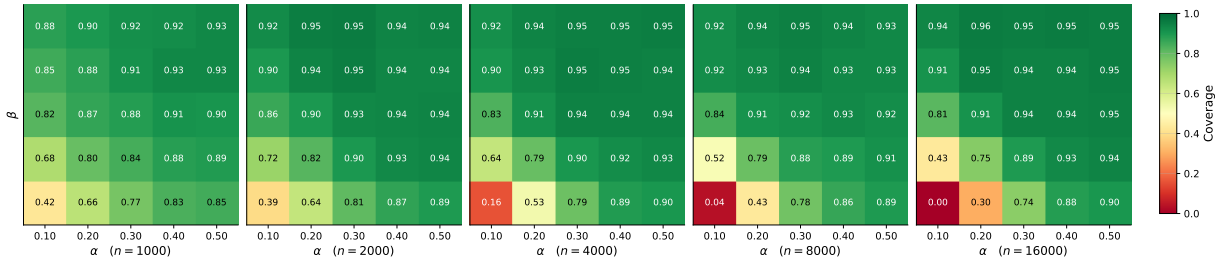
Product-rate condition. To examine Theorem 3.1(b)(i), we perturb the first propensity and prognostic candidates at rates $n^{-\alpha}$ and $n^{-\beta}$, respectively, with $(\alpha, \beta) \in \{0.1, 0.2, 0.3, 0.4, 0.5\}^2$. The perturbations share a nonlinear component and have magnitude 2.8. We consider $d = 2$, with score $(\operatorname{logit} \widehat{e}_1, \widehat{m}_{0,1})$, and $d = 3$, which additionally includes the misspecified logit $\widehat{e}_2$. Figure 3 shows that coverage approaches the nominal level in the region predicted by the theory. For $d = 2$ and $d = 3$, the displayed transitions are governed by $\alpha + \beta > 1/2$.

Dimension and regression adjustment. We use the oracle-score design: $X \sim \mathrm{Unif}([0,1]^d)$, $d \in \{2, 3, 4, 5\}$, $e(X) = 0.1 + 0.3X_1$, and $\boldsymbol{S} = (e(X), X_2, \ldots, X_d)$. The outcome regression is quadratic in the score coordinates, and the errors are heteroscedastic Gaussian. Figure 4a compares raw matching with quadratic regression adjustment. The two methods perform similarly for $d = 2, 3$, whereas raw matching deteriorates in dimensions four and five, increasingly so with sample size. Regression adjustment maintains coverage close to the nominal level in all dimensions. This supports the dimension-dependent role of adjustment in Theorem 3.1(b). The intervals use the theoretical variance under the data-generating process.

Bootstrap comparison. Figure 4b compares the linearization-based bootstrap standard error with the analytic standard error for the two-dimensional score specification under the common design. Their coverage is similar and remains close to the nominal 95% level over the displayed sample sizes. This provides a finite-sample calibration check.

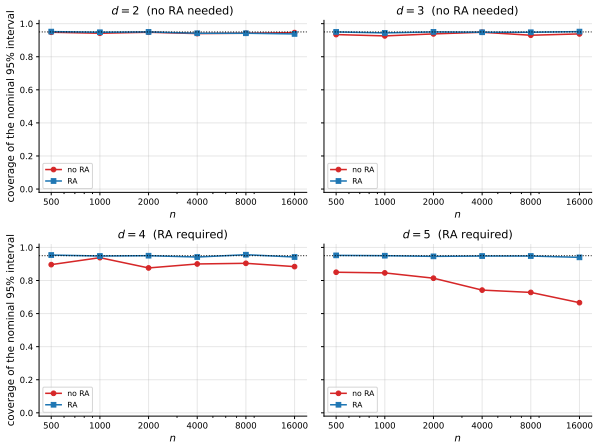


(a) $d = 2$: $\boldsymbol{S} = (\operatorname{logit} \widehat{e}_1, \widehat{m}_{0,1})$.

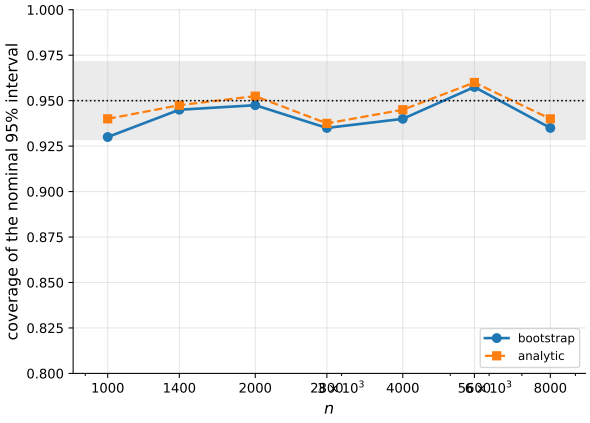


(b) $d = 3$: $\boldsymbol{S} = (\operatorname{logit} \widehat{e}_1, \operatorname{logit} \widehat{e}_2, \widehat{m}_{0,1})$.

Figure 3: Coverage of nominal 95% intervals under controlled nuisance-estimation rates.



(a) Dimension-dependent coverage with and without regression adjustment.



(b) Bootstrap and analytic standard-error coverage.

Figure 4: Additional inference diagnostics.

Additional simulations with ambient covariate dimension increasing from 20 to 1,000 are reported in Appendix B.

5 Real Data Application

We illustrate Cross Matching using the `binge` data from the `iTOS` R package, which are based on the NHANES 2017–March 2020 pre-pandemic sample (Rosenbaum, 2023). The variable `AlcGroup` classifies participants into binge drinkers (B), never-binge drinkers (N), and past binge drinkers (P). Following the data construction in our original analysis, we exclude the P group and compare B with N. The original data contain 206 B, 3,919 N, and 502 P participants. After excluding P, the analysis sample has $n = 4{,}125$ observations: 206 treated individuals (5.0%) and 3,919 controls (95.0%), a roughly 19-to-1 imbalance. Thus, $T = 1$ indicates binge-drinking status (B), whereas $T = 0$ denotes the never-binge group (N).

The outcome is `bpCombined`, a composite blood-pressure measure formed by summing systolic and diastolic blood pressure after each has been standardized using its median and median absolute deviation in the reference NHANES population. It therefore measures blood pressure on a standardized composite scale, rather than in mmHg. We adjust for nine pre-treatment characteristics: age, sex, education, body mass index (BMI), waist-to-hip ratio, vigorous physical activity, current smoking status, blood-pressure medication use, and smoking cessation status. These variables capture important demographic, socioeconomic, anthropometric, and health-behavior factors that may be associated with both drinking status and blood pressure. For the balance diagnostic, the three levels of current smoking (no, every day, and some days) are displayed separately; together with the other eight terms, this gives eleven covariate indicators in the Love plot.

Our primary targets are the counterfactual mean among binge drinkers,

$$\theta_0 = \mathbb{E}[Y(0) \mid T = 1],$$

and the average treatment effect on the treated,

$$\tau_{\mathrm{ATT}} = \mathbb{E}[Y(1) - Y(0) \mid T = 1] = \mathbb{E}[Y \mid T = 1] - \theta_0.$$

Thus, the ATT describes the average change in the composite blood-pressure measure for the observed binge-drinking group relative to the counterfactual in which these same individuals belonged to the never-binge group.

Figure 5 compares the ATT and counterfactual-mean estimates from five procedures: Cross Matching with $d = 2$, Cross Matching with $d = 4$ and regression adjustment, propensity-score matching (PSM), prognostic-score matching (PGM), and nearest-neighbor matching on the full covariate vector (m.x). The point estimates are similar across procedures: the ATT estimates lie between approximately 0.92 and 0.95, while the counterfactual means lie between approximately 0.80 and 0.84. Their confidence intervals overlap substantially, indicating that the estimated effects are not particularly sensitive to the choice among these five procedures.

Figure 6 compares the distributions of the estimated propensity-score logit among treated individuals, all controls, and matched controls for the two Cross Matching specifications. The original control group differs visibly from the treated group, whereas the distribution among matched controls closely tracks that of the treated individuals in both $d = 2$ and $d = 4$. This illustrates how the induced matching ratios select and weight controls to reproduce the treated group's score distribution.

The covariate balance results are shown in Figure 7. Before matching, several of the eleven displayed covariate indicators have substantial standardized mean differences. After matching, all absolute SMDs are below 0.1 for both score dimensions. The improvement across demographic, health, and behavioral characteristics indicates that Cross Matching substantially reduces the observed differences between the treated group and its matched controls.

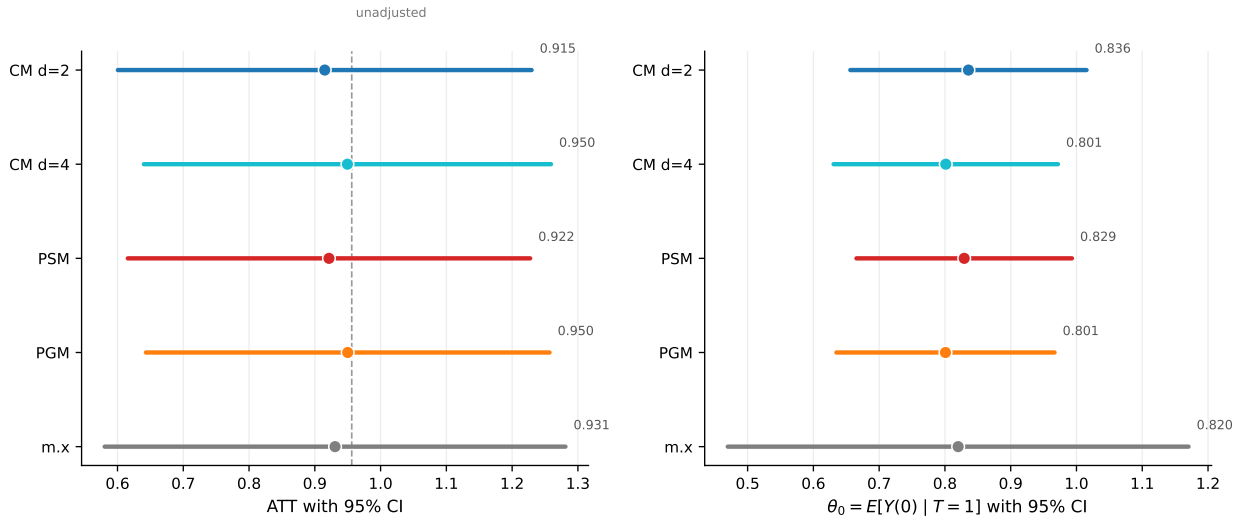


Figure 5: ATT and counterfactual-mean estimates with 95% confidence intervals for the five procedures.

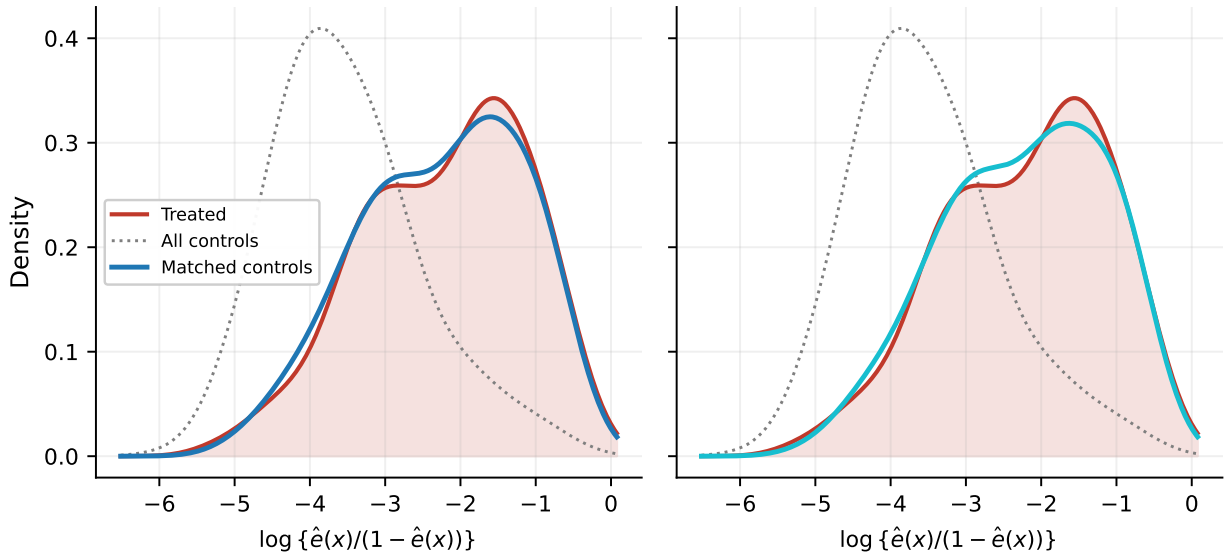


Figure 6: Distributions of the estimated propensity-score logit for treated, all control, and matched-control observations under Cross Matching with $d = 2$ and $d = 4$.

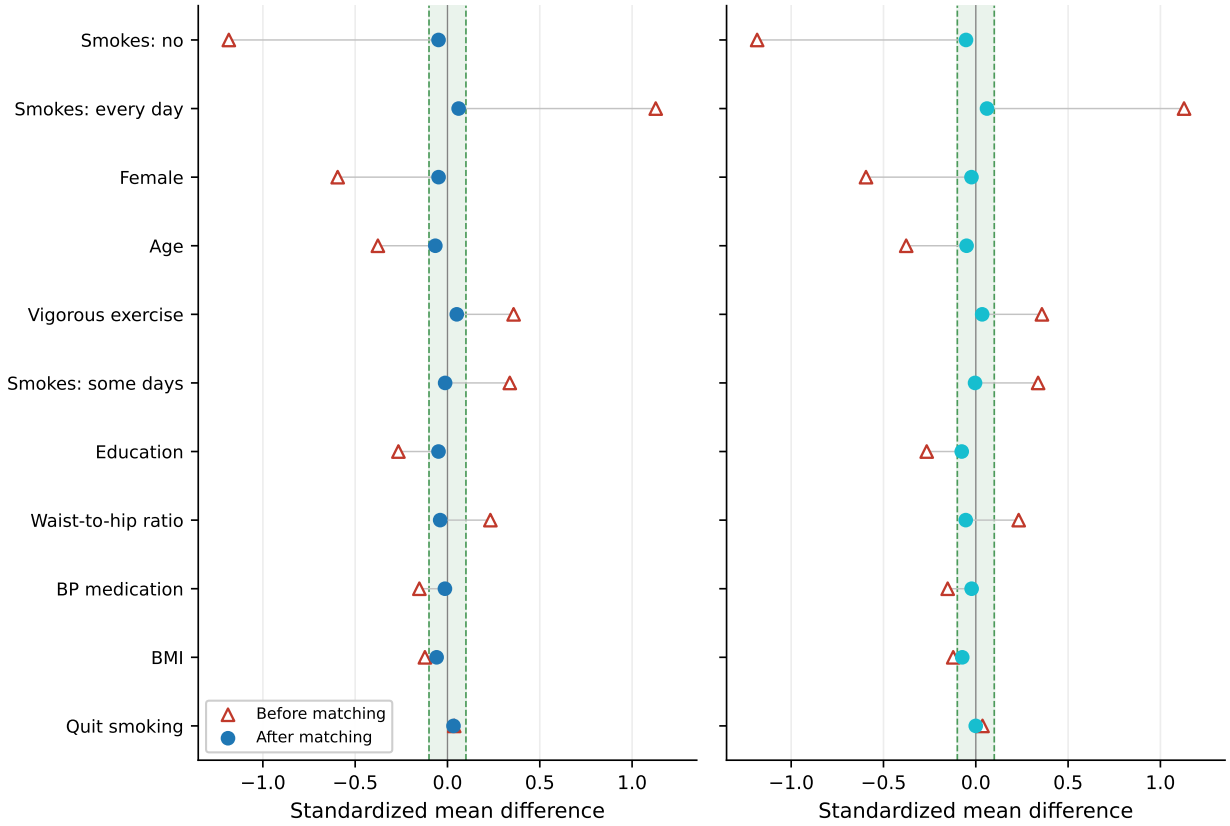


Figure 7: Covariate balance before and after Cross Matching. Points show standardized mean differences for the eleven displayed covariate indicators under $d = 2$ and $d = 4$.

6 Discussion

This paper develops Cross Matching to address two challenges that arise when nearest-neighbor matching is combined with flexible, data-adaptive score estimation: the scores must be learned without reusing the same observations for matching, and the remaining matching discrepancy must be controlled in a dimension-dependent way. The two-fold construction separates score learning from matching, while the theory separates the resulting error into a score-projection component and a local matching component. The former is governed by the product of the best propensity- and prognostic-score approximation errors, yielding a multiple-consistency result that does not require an exactly specified candidate model. The latter determines the appropriate inferential procedure: under the stated rate conditions, raw matching admits root-n inference in two- and three-dimensional score spaces, whereas regression adjustment provides a corresponding result in dimensions four and above. In both regimes, the ATT estimator is asymptotically linear and attains the semiparametric efficiency bound. The four-dimensional oracle-score construction further clarifies that accurate scores alone do not eliminate the geometric bias of nearest-neighbor matching; the need for adjustment can remain even when the propensity and prognostic information are represented exactly.

The simulation results support these theoretical distinctions. Flexible candidates from either score library can improve performance relative to specifications based only on misspecified working models, and including useful candidates from both libraries generally gives the strongest results. Under controlled nuisance perturbations, the observed coverage transitions align with the product-rate condition and, in dimension three, with the additional prognostic-score restriction. The dimension experiment shows that raw matching becomes unreliable as the score dimension increases, while regression adjustment restores coverage close to the nominal level. In the ambient-dimension experiment, Cross Matching with flexible scores maintains comparatively low RMSE as irrelevant covariates are added, although its error increases gradually and prognostic screening becomes less accurate. The real-data analysis gives a complementary illustration: five procedures produce similar ATT and counterfactual-mean estimates, while Cross Matching makes the matched-control score distribution closely track that of the treated group and brings all eleven displayed covariate SMDs below 0.1. These diagnostics indicate robustness to the choice among the examined estimators and substantial improvement in observed balance; they do not, by themselves, establish the absence of unmeasured confounding.

Several questions remain open. First, our theory uses a fixed finite collection of candidate scores and regularity conditions on the learned score distributions, including overlap, compact support, and density smoothness. It would be useful to allow the candidate libraries and score dimension to grow, and to develop guarantees for more general data-adaptive screening and representation-learning procedures. A related direction is to weaken the support and overlap conditions, which may be restrictive in applications with limited common support. Second, the dimension-dependent results are sufficient conditions rather than a complete characterization of the boundary between raw and regression-adjusted matching. Sharper bias expansions could clarify when adjustment is necessary, how the choice of the number of matches affects the trade-off, and whether adaptive choices of matching dimension and neighbor count can retain valid inference. Third, the bootstrap experiment is preliminary: the implemented resampling scheme fixes the learned scores, matching assignments, and control reuse counts. A formal bootstrap theory would need to account for their sampling variation, potentially by re-estimating the scores and repeating the matching step or by developing a suitable multiplier approximation. Finally, applications should complement covariate-balance diagnostics with sensitivity analyses for unmeasured confounding and assessments of overlap. These extensions would broaden the practical scope of Cross Matching while preserving its central goal: combining flexible score learning with transparent matching and valid uncertainty quantification.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267.

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

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of business & economic statistics*, 29(1):1–11.

Antonelli, J., Cefalu, M., Palmer, N., and Agniel, D. (2018). Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488.

Jin, Y. (2026). When does inexact matching ensure balance and inference without adjustment?

Jin, Y. and Zubizarreta, J. (2025). Cross-balancing for data-informed design and efficient analysis of observational studies. *arXiv preprint arXiv:2511.15896*.

Leacy, F. P. and Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in medicine*, 33(20):3488–3508.

Lin, Z., Ding, P., and Han, F. (2023). Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*, 91(6):2187–2217.

Rosenbaum, P. R. (2020). *Design of Observational Studies*. Springer Series in Statistics. Springer, Cham, 2 edition.

Rosenbaum, P. R. (2023). A second evidence factor for a second control group. *Biometrics*, 79(4):3968–3980.

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

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183.

Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585.

Viel, S., Truquet, L., and Yamane, I. (2025). Convergence rate for nearest neighbour matching: Geometry of the domain and higher-order regularity. *Bernoulli*. Accepted for publication; arXiv:2504.21633.

Yang, S. and Zhang, Y. (2023). Multiply robust matching estimators of average and quantile treatment effects. *Scandinavian Journal of Statistics*, 50(1):235–265.

Zeng, Y., Tong, G., Tong, J., Lu, H., Mukherjee, B., and Li, F. (2026). Where to weight? estimating population causal effects with weighted double score matching in complex surveys.

A Proofs and Additional Theoretical Results

A.1 Proof of Theorem 3.1

We first establish the consistency of the raw matching estimator, and then derive its asymptotic linear representation for $d \in \{2, 3\}$. Finally, we consider the regression-adjusted estimator for $d \geq 4$.

Fix an evaluation fold k and condition throughout on its training-fold sigma-field $\mathcal{F}_{\text{tr}}^{(k)}$. For notational simplicity, we suppress the fold superscript whenever no confusion arises. Write $n_a = n_a^{(k)}$, $n_* = n^{(k)}$, and $M = M_n$. Let P_a denote expectation conditional on $T = a$, and let $P_{a,n} f = n_a^{-1} \sum_{i \in \mathcal{I}_a^{(k)}} f(X_i)$ be the corresponding empirical average.

Let $S = \widehat{\boldsymbol{s}}^{(k)}(X)$ be the learned score vector. Define the score-space density ratio and prognostic regression function by

$$r_S(S) := \mathbb{E}_0[r_X(X) \mid S, \mathcal{F}_{\text{tr}}], \qquad \mu_S(S) := \mathbb{E}_0[m_0(X) \mid S, \mathcal{F}_{\text{tr}}],$$

and set

$$\Theta_0(S) := P_0\{r_S(S) m_0(X)\}.$$

The identity $P_0\{r_S(S) q(S)\} = P_1\{q(S)\}$ holds for every integrable S-measurable function q. In particular, $\Theta_0(S)$ is the counterfactual mean induced by projecting the covariate density ratio onto the learned score space.

For any square-integrable function $f(X)$, define

$$\begin{aligned} \Delta_M(f) := P_{0,n}\{(\widehat{r}_M - r_S) f\} \\ + (P_{0,n} - P_0)(r_S f) - (P_{1,n} - P_1) f. \end{aligned} \tag{A.1}$$

If f is measurable with respect to S, then $P_0(r_S f) = P_1 f$, and hence

$$\Delta_M(f) = P_{0,n}(\widehat{r}_M f) - P_{1,n} f. \tag{A.2}$$

Here and below, $\widehat{r}_M$ denotes the matching-induced density-ratio estimator defined in Section 2.

Let $\varepsilon = Y - m_0(X)$ for control observations. The following decomposition is the main step in the proof. It separates the error due to projecting the covariate density ratio onto the learned score space from the error due to nearest-neighbor matching.

Lemma A.1 (Basic decomposition)**.** *Let $\widehat{\theta}_{0,\text{raw}}$ and $\widehat{\theta}_{0,\text{RA}}$ be the fold-specific raw and regression-adjusted estimators. Then*

$$\begin{aligned} \widehat{\theta}_{0,\text{raw}} - \theta_0 = \{\Theta_0(S) - \theta_0\} + \Delta_M(m_0) \\ + (P_{1,n} - P_1) m_0 + P_{0,n}(\widehat{r}_M \varepsilon), \end{aligned} \tag{A.3}$$

and

$$\begin{aligned} \widehat{\theta}_{0,\text{RA}} - \theta_0 = \{\Theta_0(S) - \theta_0\} + \Delta_M(m_0 - \widehat{g}) \\ + (P_{1,n} - P_1) m_0 + P_{0,n}(\widehat{r}_M \varepsilon). \end{aligned} \tag{A.4}$$

Proof. For the raw estimator, write $Y = m_0(X) + \varepsilon$ in the control group. Then

$$\widehat{\theta}_{0,\text{raw}} = P_{0,n}(\widehat{r}_M m_0) + P_{0,n}(\widehat{r}_M \varepsilon).$$

By the definition of Δ_M,

$$P_{0,n}(\widehat{r}_M m_0) = \Delta_M(m_0) + P_{1,n} m_0 + \Theta_0(S) - \theta_0.$$

Substitution gives (A.3).

For the regression-adjusted estimator, use

$$\widehat{\theta}_{0,\text{RA}} = P_{1,n} \widehat{g} + P_{0,n}\{\widehat{r}_M (m_0 - \widehat{g})\} + P_{0,n}(\widehat{r}_M \varepsilon).$$

Since $\widehat{g}$ is measurable with respect to the learned score vector,

$$P_0\{(r_S - r_X)\widehat{g}(S)\} = 0.$$

Also,

$$P_0\{(r_S - r_X)m_0\} = \Theta_0(S) - \theta_0.$$

It follows from the definition of Δ_M that

$$P_{0,n}\{\widehat{r}_M(m_0 - \widehat{g})\} = \Delta_M(m_0 - \widehat{g}) + P_{1,n}(m_0 - \widehat{g}) + \Theta_0(S) - \theta_0.$$

Adding $P_{1,n}\widehat{g}$ proves (A.4). □

Lemma A.2 (Score projection error)**.** *Under Assumption 2.1, for each fold k,*

$$\|r_X - r_S\|_{0,2}^{(k)} \le C_\eta \epsilon_{e,n}^{(k)}, \qquad C_\eta := \frac{\pi_0}{\pi_1 \eta^2}. \tag{A.5}$$

Moreover,

$$|\Theta_0(S) - \theta_0| \le C_\eta \epsilon_{e,n}^{(k)} \epsilon_{m,n}^{(k)}. \tag{A.6}$$

Proof. For any square-integrable function $q(S)$, the density-ratio identity gives

$$P_0\{q(S) r_X(X)\} = P_1\{q(S)\} = P_0\{q(S) r_S(S)\}.$$

Therefore,

$$r_S(S) = \mathbb{E}_0[r_X(X) \mid S, \mathcal{F}_{\text{tr}}], \tag{A.7}$$

so r_S is the $L_2(P_0)$ projection of r_X onto the functions measurable with respect to S.

For a propensity-score candidate $\widehat{e}_j$, define

$$q_j(X) := \frac{\pi_0 \widehat{e}_j(X)}{\pi_1\{1 - \widehat{e}_j(X)\}}.$$

Because $\widehat{e}_j(X)$ is a coordinate of S, $q_j(X)$ is S-measurable. The projection property in (A.7) yields

$$\|r_X - r_S\|_{0,2} \le \|r_X - q_j\|_{0,2}.$$

For $u, v \in [\eta, 1 - \eta]$,

$$\left| \frac{u}{1-u} - \frac{v}{1-v} \right| = \frac{|u - v|}{(1-u)(1-v)} \le \frac{|u - v|}{\eta^2}.$$

Consequently,

$$\|r_X - q_j\|_{0,2} \le \frac{\pi_0}{\pi_1 \eta^2} \|\widehat{e}_j - e\|_{0,2}.$$

Taking the minimum over j proves (A.5).

Next, since $P_0(r_X m_0) = \theta_0$,

$$\Theta_0(S) - \theta_0 = P_0\{(r_S - r_X)m_0\}.$$

For every prognostic-score candidate $\widehat{m}_{0,\ell}$, the function $\widehat{m}_{0,\ell}(X)$ is S-measurable. Thus, by (A.7),

$$P_0\{(r_S - r_X)\widehat{m}_{0,\ell}\} = 0.$$

It follows that

$$\Theta_0(S) - \theta_0 = P_0\{(r_S - r_X)(m_0 - \widehat{m}_{0,\ell})\}.$$

Applying the Cauchy–Schwarz inequality and (A.5),

$$|\Theta_0(S) - \theta_0| \le C_\eta \|\widehat{e}_j - e\|_{0,2} \|m_0 - \widehat{m}_{0,\ell}\|_{0,2}.$$

Taking the minimum over both candidate libraries proves (A.6). □

Lemma A.3 (Nearest-neighbor matching discrepancy)**.** *Suppose Assumptions 2.2 and 2.3 hold. Let* $a_n := (M_n/n_0)^{1/d}$*, where* n_0 *is the number of controls in the evaluation fold.*

For an S*-measurable function* q *with finite second moment,*

$$\Delta_{M_n}(q) = O_p\{(a_n + n^{-1/2})\|q\|_{0,2}\}. \tag{A.8}$$

For a function h *satisfying* $\mathbb{E}_0[h(X) \mid S, \mathcal{F}_{\text{tr}}] = 0$*,*

$$\Delta_{M_n}(h) = O_p(n^{-1/2}\|h\|_{0,2}). \tag{A.9}$$

In addition, if $d \in \{2, 3\}$ *and* q *is twice continuously differentiable on* $\mathcal{S}$*, then*

$$\sqrt{n_*}\Delta_{M_n}(q) = o_p(1) \tag{A.10}$$

provided that $M_n^2/n \to 0$ *for* $d = 2$*, or* $M_n^4/n \to 0$ *for* $d = 3$*.*

Proof. We first consider an S-measurable function q. By (A.2),

$$\Delta_{M_n}(q) = P_{0,n}(\widehat{r}_{M_n} q) - P_{1,n} q.$$

Let $\bar{r}_{M_n}(s) := \mathbb{E}[\widehat{r}_{M_n}(X_j) \mid S_j = s, \mathcal{F}_{\text{tr}}]$ denote the conditional mean of the matching-induced density-ratio estimate. The matching density-ratio result gives the following bound under its matching-count and density regularity conditions (Lin et al., 2023):

$$\|\bar{r}_{M_n} - r_S\|_\infty \le C \left(\frac{M_n}{n_0}\right)^{1/d}.$$

The same matching-ratio theory controls the empirical fluctuation of the reuse-count weights. In particular,

$$P_{0,n}\{(\widehat{r}_{M_n} - \bar{r}_{M_n})q\} = O_p(n^{-1/2}\|q\|_{0,2}).$$

Decompose

$$\begin{aligned}\Delta_{M_n}(q) = {} & P_{0,n}\{(\widehat{r}_{M_n} - \bar{r}_{M_n})q\} \\ & + P_{0,n}\{(\bar{r}_{M_n} - r_S)q\} \\ & + (P_{0,n} - P_0)(r_S q) - (P_{1,n} - P_1)q.\end{aligned}$$

The second term is bounded by

$$|P_{0,n}\{(\bar{r}_{M_n} - r_S)q\}| \le \|\bar{r}_{M_n} - r_S\|_\infty P_{0,n}|q| = O_p(a_n\|q\|_{0,2}).$$

The last two terms are ordinary empirical fluctuations and are $O_p(n^{-1/2}\|q\|_{0,2})$, since r_S is bounded under overlap. Combining these bounds proves (A.8).

Now let h satisfy $\mathbb{E}_0[h \mid S, \mathcal{F}_{\text{tr}}] = 0$. Using the definition in (A.1), we have

$$\begin{aligned}\Delta_{M_n}(h) = {} & P_{0,n}\{(\widehat{r}_{M_n} - r_S)h\} \\ & + (P_{0,n} - P_0)(r_S h) - (P_{1,n} - P_1)h.\end{aligned}$$

The conditional mean-zero property eliminates the population component of the first term. The matching-ratio second-moment bound controls its variance at the order $n^{-1}\|h\|_{0,2}^2$. The second term has variance bounded by $Cn_0^{-1}\|h\|_{0,2}^2$, because r_S is bounded. Finally, strong overlap implies

$$P_1(h^2) = P_0(r_X h^2) \le \|r_X\|_\infty P_0(h^2) \le C\|h\|_{0,2}^2,$$

so the third term is also $O_p(n^{-1/2}\|h\|_{0,2})$. This proves (A.9).

For (A.10), we invoke a low-dimensional bias result for nearest-neighbor matching (Viel et al., 2025). It applies under the support-thickness, density regularity, and boundary conditions imposed in Assumption 2.2, together with twice continuous differentiability of the regression function on the matching space. Under these conditions, the matching discrepancy is $o_p(n_*^{-1/2})$ when $d = 2$ and $M_n^2/n \to 0$, or when $d = 3$ and $M_n^4/n \to 0$. The twice continuous differentiability of the regression function on the matching space holds because $\widehat{m}_{0,\ell_*}$ is a coordinate of the learned score vector. This gives (A.10). □

Lemma A.4 (Control of the outcome residual)**.** *Suppose the conditional second-moment condition on the control outcome in Assumption 2.1 holds. If $\epsilon_{e,n} = o_p(1)$, then*

$$P_{0,n}\{(\widehat{r}_{M_n} - r_X)\varepsilon\} = o_p(n^{-1/2}). \tag{A.11}$$

Proof. Write

$$P_{0,n}\{(\widehat{r}_{M_n} - r_X)\varepsilon\} = P_{0,n}\{(\widehat{r}_{M_n} - r_S)\varepsilon\} + P_{0,n}\{(r_S - r_X)\varepsilon\}. \tag{A.12}$$

Let

$$\mathcal{H} := \mathcal{G}^{(k)} \vee \sigma\{X_j : j \in \mathcal{I}_0^{(k)}\}.$$

Conditional on $\mathcal{H}$, the matching weights and the control covariates are fixed, while $\mathbb{E}[\varepsilon_j \mid X_j, T_j = 0] = 0$. By the conditional second-moment bound,

$$\begin{aligned} &\mathrm{Var}\left(P_{0,n}\{(\widehat{r}_{M_n} - r_S)\varepsilon\} \mid \mathcal{H}\right) \\ &\quad \le \frac{C}{n_0^2} \sum_{j \in \mathcal{I}_0^{(k)}} \{\widehat{r}_{M_n}(X_j) - r_S(S_j)\}^2. \end{aligned} \tag{A.13}$$

The L_2 consistency of the matching-induced density-ratio estimator follows from the density-ratio result under Assumptions 2.2 and 2.3 (Lin et al., 2023), giving

$$P_{0,n}\{(\widehat{r}_{M_n} - r_S)^2\} = o_p(1).$$

Consequently, the conditional variance in (A.13) is $o_p(n_0^{-1})$, and conditional Chebyshev's inequality yields

$$P_{0,n}\{(\widehat{r}_{M_n} - r_S)\varepsilon\} = o_p(n^{-1/2}). \tag{A.14}$$

For the second term in (A.12), conditional on the control covariates and the training fold, its mean is zero and

$$\begin{aligned} &\mathrm{Var}\left(P_{0,n}\{(r_S - r_X)\varepsilon\} \mid X_j : j \in \mathcal{I}_0^{(k)}, \mathcal{F}_{\mathrm{tr}}\right) \\ &\quad \le \frac{C}{n_0} P_{0,n}\{(r_S - r_X)^2\}. \end{aligned} \tag{A.15}$$

By Lemma A.2,

$$P_{0,n}\{(r_S - r_X)^2\} = O_p(\epsilon_{e,n}^2) + o_p(1).$$

Since $\epsilon_{e,n} = o_p(1)$, the right-hand side of (A.15) is $o_p(n_0^{-1})$. Another application of conditional Chebyshev's inequality gives

$$P_{0,n}\{(r_S - r_X)\varepsilon\} = o_p(n^{-1/2}). \tag{A.16}$$

Combining (A.14) and (A.16) proves (A.11). □

Proof of Theorem 3.1(a): Consistency. By Lemmas A.1 and A.2,

$$\widehat{\theta}_{0,\mathrm{raw}} - \theta_0 = \{\Theta_0(S) - \theta_0\} + \Delta_{M_n}(m_0) + (P_{1,n} - P_1)m_0 + P_{0,n}(\widehat{r}_{M_n}\varepsilon). \tag{A.17}$$

The first term is $o_p(1)$ under $\epsilon_{e,n}\epsilon_{m,n} = o_p(1)$.

To control the matching discrepancy, choose a candidate $\widehat{m}_{0,\ell_*}$ attaining the minimum defining $\epsilon_{m,n}^{(k)}$. Since it is a coordinate of the learned score vector, it is S-measurable. Decompose

$$m_0 = \widehat{m}_{0,\ell_*} + \{\mu_S - \widehat{m}_{0,\ell_*}\} + (m_0 - \mu_S). \tag{A.18}$$

The conditional expectation defining μ_S is an $L_2(P_0)$ projection. Therefore,

$$\|\mu_S - \widehat{m}_{0,\ell_*}\|_{0,2} \le \epsilon_{m,n}, \qquad \|m_0 - \mu_S\|_{0,2} \le \epsilon_{m,n}.$$

Applying (A.8) to the first two terms in (A.18), and (A.9) to the last term, gives

$$\begin{aligned}\Delta_{M_n}(m_0) = O_p\{(a_n + n^{-1/2})\}\|\widehat{m}_{0,\ell_*}\|_{0,2} \\ + O_p\{(a_n + n^{-1/2})\}\epsilon_{m,n} + O_p(n^{-1/2}\epsilon_{m,n}).\end{aligned} \tag{A.19}$$

Because $\epsilon_{m,n} = O_p(1)$ and $\|\widehat{m}_{0,\ell_*}\|_{0,2} = O_p(1)$, while $a_n = (M_n/n_0)^{1/d} \to 0$, (A.19) implies $\Delta_{M_n}(m_0) = o_p(1)$.

The empirical fluctuation $(P_{1,n} - P_1)m_0$ converges to zero by the law of large numbers. Moreover,

$$P_{0,n}(\widehat{r}_{M_n}\varepsilon) = P_{0,n}\{(\widehat{r}_{M_n} - r_X)\varepsilon\} + P_{0,n}(r_X\varepsilon).$$

The first term is $o_p(1)$ by the matching-ratio consistency and the boundedness of $\epsilon_{e,n}$ implied by clipping; the second converges to zero by the law of large numbers and $\mathbb{E}_0[r_X\varepsilon] = 0$. Hence all terms in (A.17) are $o_p(1)$, proving

$$\widehat{\theta}^{(k)}_{0,\mathrm{raw}} \xrightarrow{p} \theta_0.$$

Proof of Theorem 3.1(b)(i): Raw matching for $d = 2, 3$. By Lemmas A.1 and A.2, it suffices to show

$$\Theta_0(S) - \theta_0 = o_p(n^{-1/2}), \qquad \Delta_{M_n}(m_0) = o_p(n^{-1/2}), \tag{A.20}$$

and then apply Lemma A.4.

The projection term satisfies

$$|\Theta_0(S) - \theta_0| \le C_\eta \epsilon_{e,n}\epsilon_{m,n} = o_p(n^{-1/2}) \tag{A.21}$$

by the product-rate condition.

For the matching discrepancy, use the decomposition (A.18). The selected candidate score $\widehat{m}_{0,\ell_*}$ is a coordinate function of S and hence is twice continuously differentiable as a function on the score space. Since the candidate library is finite, the low-dimensional matching result (A.10) applies uniformly over its elements. Thus,

$$\Delta_{M_n}(\widehat{m}_{0,\ell_*}) = o_p(n^{-1/2}). \tag{A.22}$$

For the remaining two terms, (A.8) and (A.9) give

$$\begin{aligned}&\Delta_{M_n}(\mu_S - \widehat{m}_{0,\ell_*}) + \Delta_{M_n}(m_0 - \mu_S) \\ &\quad = O_p\{(a_n + n^{-1/2})\epsilon_{m,n}\} + O_p(n^{-1/2}\epsilon_{m,n}).\end{aligned} \tag{A.23}$$

The assumed rate $a_n\epsilon_{m,n} = o_p(n^{-1/2})$ also implies $\epsilon_{m,n} = o_p(1)$ in dimensions two and three. Indeed,

$$\sqrt{n}\, a_n \asymp \begin{cases} M_n^{1/2}, & d = 2, \\ n^{1/6}M_n^{1/3}, & d = 3, \end{cases}$$

and both quantities diverge under Assumption 2.3. Consequently, the right-hand side of (A.23) is $o_p(n^{-1/2})$. Combining (A.22) and (A.23) proves the second assertion in (A.20).

Finally, Lemma A.4 and the exact decomposition (A.3) yield

$$\widehat{\theta}_{0,\mathrm{raw}} - \theta_0 = (P_{1,n} - P_1)m_0 + P_{0,n}(r_X\varepsilon) + o_p(n^{-1/2}). \tag{A.24}$$

The two empirical terms are independent across treatment groups. Since $n_a/n_* \to_p \pi_a$, the ordinary central limit theorem gives

$$\sqrt{n_*}(P_{1,n} - P_1)m_0 \rightsquigarrow N\left(0, \frac{\mathrm{Var}\{m_0(X) \mid T = 1\}}{\pi_1}\right) \tag{A.25}$$

and

$$\sqrt{n_*}P_{0,n}(r_X\varepsilon) \rightsquigarrow N\left(0, \frac{\mathbb{E}_0[r_X(X)^2\varepsilon^2]}{\pi_0}\right).$$

Therefore, by independence and (A.24),

$$\sqrt{n_*}\{\widehat{\theta}_{0,\mathrm{raw}} - \theta_0\} \rightsquigarrow N(0, V_0),$$

where

$$V_0 = \frac{\mathrm{Var}\{m_0(X) \mid T = 1\}}{\pi_1} + \frac{\mathbb{E}_0[r_X(X)^2\varepsilon^2]}{\pi_0}. \tag{A.26}$$

Proof of Theorem 3.1(b)(ii): Regression-adjusted matching for $d \geq 4$. We again start from the decomposition in Lemma A.1. The projection term satisfies (A.21). For the matching discrepancy, write

$$m_0 - \widehat{g} = (\mu_S - \widehat{g}) + (m_0 - \mu_S).$$

The first component is S-measurable, whereas $\mathbb{E}_0[m_0 - \mu_S \mid S, \mathcal{F}_{\mathrm{tr}}] = 0$. By (A.8) and (A.9),

$$\begin{aligned}\Delta_{M_n}(m_0 - \widehat{g}) = O_p\{(a_n + n^{-1/2})\|\mu_S - \widehat{g}\|_{0,2}\} \\ + O_p(n^{-1/2}\|m_0 - \mu_S\|_{0,2}).\end{aligned} \tag{A.27}$$

Projection optimality and the definition of $\epsilon_{m,n}$ imply

$$\|m_0 - \mu_S\|_{0,2} \leq \epsilon_{m,n} = o_p(1).$$

The assumed regression rate is $a_n\|\mu_S - \widehat{g}\|_{0,2} = o_p(n^{-1/2})$. Furthermore, Assumption 2.3 implies

$$\sqrt{n}\, a_n \asymp n^{1/2-1/d} M_n^{1/d} \longrightarrow \infty \qquad (d \geq 4),$$

so that $\|\mu_S - \widehat{g}\|_{0,2} = o_p(1)$. It follows that the right-hand side of (A.27) is $o_p(n^{-1/2})$. Thus,

$$\Delta_{M_n}(m_0 - \widehat{g}) = o_p(n^{-1/2}).$$

The product-rate condition gives $\Theta_0(S) - \theta_0 = o_p(n^{-1/2})$, while Lemma A.4 gives $P_{0,n}\{(\widehat{r}_{M_n} - r_X)\varepsilon\} = o_p(n^{-1/2})$. Substituting these results into (A.4) establishes

$$\widehat{\theta}_{0,\mathrm{RA}} - \theta_0 = (P_{1,n} - P_1)m_0 + P_{0,n}(r_X\varepsilon) + o_p(n^{-1/2}). \tag{A.28}$$

The same central limit argument as in (A.25)–(A.26) then yields

$$\sqrt{n_*}\{\widehat{\theta}_{0,\mathrm{RA}} - \theta_0\} \rightsquigarrow N(0, V_0).$$

This completes the proof of Theorem 3.1. □

A.2 Proof of Corollary 3.2: ATT Expansion and Semiparametric Efficiency

We prove the result for either of the two estimators in its corresponding dimension regime. Let $\widehat{\theta}_{0,\bullet}^{(k)}$ denote the raw estimator when $d \in \{2, 3\}$ and the regression-adjusted estimator when $d \geq 4$. By (A.24) and (A.28), for each fold k,

$$\widehat{\theta}_{0,\bullet}^{(k)} - \theta_0 = (P_{1,n}^{(k)} - P_1)m_0 + P_{0,n}^{(k)}(r_X\varepsilon) + o_p\{(n^{(k)})^{-1/2}\}. \tag{A.29}$$

Let $\widehat{\tau}_{\mathrm{ATT}}^{(k)} := P_{1,n}^{(k)}Y - \widehat{\theta}_{0,\bullet}^{(k)}$ be the fold-specific ATT estimator. Since $\tau_{\mathrm{ATT}} = P_1 Y - \theta_0$, (A.29) gives

$$\begin{aligned}\widehat{\tau}_{\mathrm{ATT}}^{(k)} - \tau_{\mathrm{ATT}} = (P_{1,n}^{(k)} - P_1)(Y - m_0) \\ - P_{0,n}^{(k)}(r_X\varepsilon) + o_p\{(n^{(k)})^{-1/2}\}.\end{aligned}$$

Because $P_1(Y - m_0) = \tau_{\mathrm{ATT}}$, the leading term can be written as a sum over the observations in fold $\mathcal{I}^{(k)}$:

$$\begin{aligned}\widehat{\tau}^{(k)}_{\mathrm{ATT}} - \tau_{\mathrm{ATT}} = \frac{1}{n^{(k)}} \sum_{i \in \mathcal{I}^{(k)}} \Bigg[& \frac{T_i}{\pi_1}\{Y_i - m_0(X_i) - \tau_{\mathrm{ATT}}\} \\ & - \frac{1 - T_i}{\pi_0} r_X(X_i)\{Y_i - m_0(X_i)\} \Bigg] \\ & + o_p\{(n^{(k)})^{-1/2}\}.\end{aligned} \tag{A.30}$$

The replacement of $n_a^{(k)}/n^{(k)}$ by π_a in this expression is justified by $n_a^{(k)}/n^{(k)} \to_p \pi_a$ and the fact that the corresponding centered empirical averages are $O_p\{(n^{(k)})^{-1/2}\}$.

Define

$$\phi_\tau(O) := \frac{T}{\pi_1}\{Y - m_0(X) - \tau_{\mathrm{ATT}}\} - \frac{1 - T}{\pi_0} r_X(X)\{Y - m_0(X)\}.$$

Then (A.30) becomes

$$\widehat{\tau}^{(k)}_{\mathrm{ATT}} - \tau_{\mathrm{ATT}} = \frac{1}{n^{(k)}} \sum_{i \in \mathcal{I}^{(k)}} \phi_\tau(O_i) + o_p\{(n^{(k)})^{-1/2}\}. \tag{A.31}$$

Since the two folds partition the full sample and have equal size, averaging (A.31) over $k = 1, 2$ gives

$$\sqrt{n}(\widehat{\tau}_{\mathrm{ATT}} - \tau_{\mathrm{ATT}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_\tau(O_i) + o_p(1),$$

where $\widehat{\tau}_{\mathrm{ATT}}$ denotes the corresponding two-fold raw or regression-adjusted estimator. In particular, the factor $1/2$ in the two-fold average is exactly canceled by $n^{(k)} = n/2$.

It remains to calculate the variance of the influence function. Let $m_1(x) := \mathbb{E}[Y(1) \mid X = x]$, $\tau(x) := m_1(x) - m_0(x)$, and $\sigma_a^2(x) := \mathrm{Var}\{Y(a) \mid X = x\}$ for $a \in \{0, 1\}$. Since $T(1 - T) = 0$, the cross-product in $\mathbb{E}[\phi_\tau(O)^2]$ vanishes. Therefore,

$$\begin{aligned}V_{\mathrm{ATT}} = & \frac{1}{\pi_1^2}\mathbb{E}\left[T\{Y - m_0(X) - \tau_{\mathrm{ATT}}\}^2\right] \\ & + \frac{1}{\pi_0^2}\mathbb{E}\left[(1 - T) r_X(X)^2\{Y - m_0(X)\}^2\right].\end{aligned} \tag{A.32}$$

For the treated component, iterated expectation gives

$$\begin{aligned}&\mathbb{E}[T\{Y - m_0(X) - \tau_{\mathrm{ATT}}\}^2] \\ &\quad = \mathbb{E}\left[e(X)\left\{\sigma_1^2(X) + (\tau(X) - \tau_{\mathrm{ATT}})^2\right\}\right].\end{aligned}$$

For the control component, using $r_X(X) = \pi_0 e(X)/\{\pi_1(1 - e(X))\}$,

$$\begin{aligned}&\frac{1}{\pi_0^2}\mathbb{E}[(1 - T) r_X(X)^2\{Y - m_0(X)\}^2] \\ &\quad = \frac{1}{\pi_1^2}\mathbb{E}\left[\frac{e(X)^2}{1 - e(X)}\sigma_0^2(X)\right].\end{aligned} \tag{A.33}$$

Combining (A.32)–(A.33), we obtain

$$V_{\mathrm{ATT}} = \frac{1}{\pi_1^2}\mathbb{E}\left[e(X)\left\{\sigma_1^2(X) + (\tau(X) - \tau_{\mathrm{ATT}})^2\right\} + \frac{e(X)^2}{1 - e(X)}\sigma_0^2(X)\right].$$

This is the semiparametric efficiency bound for the ATT in the nonparametric unconfoundedness model (Hahn, 1998, Theorem 2). Hence the two-fold Cross Matching estimator attains the semiparametric efficiency bound in both dimension regimes under the conditions of Theorem 3.1. □

A.3 A Four-Dimensional Oracle-Score Counterexample

We construct a data-generating process in which the learned score space contains the exact propensity information and the true prognostic function, but raw nearest-neighbor matching still has a non-negligible root-n bias in dimension four.

Let $E \sim \mathrm{Unif}[1/4, 3/4]$, $Z_1 \sim q_\gamma$, and $Z_2, Z_3 \sim \mathrm{Unif}[-1, 1]$ be mutually independent, where $\gamma > 0$ and

$$q_\gamma(z) := \frac{\gamma e^{\gamma z}}{e^\gamma - e^{-\gamma}}, \qquad -1 \le z \le 1.$$

Set $X = (E, Z_1, Z_2, Z_3)$ and let $T \mid X \sim \mathrm{Bernoulli}(E)$. The potential outcomes are

$$Y(0) = h(Z_1) + \xi_0, \qquad Y(1) = Y(0) + \tau, \qquad h(z) := z - \frac{z^3}{3},$$

where ξ_0 is independent of X and has mean zero and finite variance. Thus, the true propensity score is $e(X) = E$, and the control prognostic function is $m_0(X) = h(Z_1)$.

Consider the candidate score collection

$$\widehat{e}_1(X) = E, \qquad \widehat{e}_2(X) = \mathrm{expit}(Z_2), \qquad \widehat{m}_{0,1}(X) = Z_1, \qquad \widehat{m}_{0,2}(X) = Z_3.$$

After the logit transformation of the propensity coordinates, the resulting four-dimensional score vector, up to fixed coordinate-wise standardization, is

$$S = (V, Z_2, Z_1, Z_3), \qquad V := \mathrm{logit}(E).$$

The first coordinate is the exact propensity score on the logit scale. Moreover, $m_0(X) = h(Z_1)$ is a known smooth function of the third score coordinate. Hence $\mu_S(S) = m_0(X)$, and there is no score projection residual. Also, $\epsilon_{e,n} = 0$, so the product approximation condition is satisfied exactly.

Let $e(v) := \mathrm{expit}(v)$ and

$$\mathcal{S} = [\mathrm{logit}(1/4), \mathrm{logit}(3/4)] \times [-1, 1]^3.$$

Since $\pi_1 = \pi_0 = 1/2$, the conditional score densities of treated and control observations are

$$f_1(s) = e(v)^2\{1 - e(v)\}q_\gamma(z_1), \qquad f_0(s) = e(v)\{1 - e(v)\}^2 q_\gamma(z_1).$$

Both densities are bounded, bounded away from zero, and Lipschitz on the compact rectangular support. In particular,

$$\frac{\partial}{\partial z_1} \log f_0(s) = \gamma. \tag{A.34}$$

We next calculate the bias of raw matching. Write $d = 4$ and let $v_4 = \pi^2/2$ be the volume of the unit ball in $\mathbb{R}^4$. For a treated score point s in the interior of $\mathcal{S}$, denote by $R_j(s)$ the distance to its jth nearest control score and by $\Delta_j = (\Delta_{j,1}, \ldots, \Delta_{j,4})$ the corresponding score displacement. Conditional on $R_j(s) = r$, the direction Δ_j/r lies on the unit sphere. Its conditional density is proportional to $f_0(s + r\omega)$, where $\omega \in \mathbb{S}^3$.

A first-order expansion of this density around s gives

$$\mathbb{E}\left[\frac{\Delta_j}{r} \,\middle|\, R_j(s) = r, S_i = s\right] = \frac{r}{4}\nabla \log f_0(s) + o(r),$$

and, for the third coordinate corresponding to Z_1,

$$\mathbb{E}\left[\left(\frac{\Delta_{j,3}}{r}\right)^2 \,\middle|\, R_j(s) = r, S_i = s\right] = \frac{1}{4} + o(1).$$

Consequently, using (A.34),

$$\mathbb{E}[\Delta_{j,3} \mid R_j(s) = r, S_i = s] = \frac{\gamma r^2}{4} + o(r^2), \tag{A.35}$$

and

$$\mathbb{E}[\Delta_{j,3}^2 \mid R_j(s) = r, S_i = s] = \frac{r^2}{4} + o(r^2). \tag{A.36}$$

The factor $1/4$ in these expressions is the second directional moment on the three-dimensional unit sphere in $\mathbb{R}^4$.

Since $h'(z) = 1 - z^2$ and $h''(z) = -2z$, Taylor expansion gives

$$h(z_1 + \Delta_{j,3}) - h(z_1) = h'(z_1)\Delta_{j,3} + \frac{1}{2}h''(z_1)\Delta_{j,3}^2 + o(r^2).$$

Taking conditional expectation and applying (A.35)–(A.36),

$$\begin{aligned} &\mathbb{E}[h(z_1 + \Delta_{j,3}) - h(z_1) \mid R_j(s) = r, S_i = s] \\ &\quad = \frac{r^2}{4}\left\{\gamma h'(z_1) + \frac{1}{2}h''(z_1)\right\} + o(r^2). \end{aligned} \tag{A.37}$$

It remains to average over the nearest-neighbor radii. In the interior of the support, the local point process of control scores is asymptotically homogeneous with intensity $n_0 f_0(s)$. Hence the standard nearest-neighbor order-statistic approximation gives

$$\mathbb{E}[R_j(s)^2 \mid S_i = s] = \frac{\Gamma(j + 1/2)}{\Gamma(j)}\{n_0 v_4 f_0(s)\}^{-1/2}\{1 + o(1)\}.$$

Using $\Gamma(j + 1/2)/\Gamma(j) \sim j^{1/2}$ and $\sum_{j=1}^{M_n} j^{1/2} \sim (2/3)M_n^{3/2}$, we obtain

$$\frac{1}{M_n}\sum_{j=1}^{M_n}\mathbb{E}[R_j(s)^2 \mid S_i = s] = \frac{2}{3}\left\{\frac{M_n}{n_0 v_4 f_0(s)}\right\}^{1/2}\{1 + o(1)\}. \tag{A.38}$$

Combining (A.37) and (A.38), the conditional bias averaged over the M_n matched controls is

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{M_n}\sum_{j=1}^{M_n}\{h(Z_{1,(j)}) - h(z_1)\}\middle| S_i = s\right] \\ &\quad = \frac{1}{6\sqrt{v_4}}\left(\frac{M_n}{n_0}\right)^{1/2}\frac{\gamma h'(z_1) + \frac{1}{2}h''(z_1)}{\sqrt{f_0(s)}} + o\left\{\left(\frac{M_n}{n_0}\right)^{1/2}\right\}. \end{aligned} \tag{A.39}$$

We now average over the treated score distribution. The support is a product rectangle and the outcome regression depends only on z_1. The faces perpendicular to the other three coordinates preserve symmetry in the z_1 direction, while $h'(-1) = h'(1) = 0$ makes the contribution from the two z_1 boundary layers of smaller order. Thus, the boundary contribution is negligible relative to $(M_n/n_0)^{1/2}$, and integration of (A.39) yields

$$\mathbb{E}[\widehat{\theta}_{0,\mathrm{raw}} - \theta_0] = C_\gamma\left(\frac{M_n}{n_0}\right)^{1/2} + o\left\{\left(\frac{M_n}{n_0}\right)^{1/2}\right\},$$

where

$$\begin{aligned} C_\gamma &= \frac{1}{6\sqrt{v_4}}\int_{\mathcal{S}}\frac{f_1(s)}{\sqrt{f_0(s)}}\left\{\gamma h'(z_1) + \frac{1}{2}h''(z_1)\right\}ds \\ &= \frac{4}{6\sqrt{v_4}}\left\{\int_{\mathrm{logit}(1/4)}^{\mathrm{logit}(3/4)} e(v)^{3/2}\,dv\right\} \\ &\qquad \times\left\{\int_{-1}^{1}\sqrt{q_\gamma(z)}\left[\gamma(1 - z^2) - z\right]dz\right\}. \end{aligned}$$

To verify that this constant is strictly positive, note that $q_\gamma'(z) = \gamma q_\gamma(z)$ and hence $(\sqrt{q_\gamma})' = (\gamma/2)\sqrt{q_\gamma}$. Integration by parts, using $h'(-1) = h'(1) = 0$, gives

$$\begin{aligned}
&\int_{-1}^{1} \sqrt{q_\gamma(z)} \left\{ \gamma h'(z) + \frac{1}{2} h''(z) \right\} dz \\
&\quad = \gamma \int_{-1}^{1} \sqrt{q_\gamma(z)} h'(z)\, dz - \frac{\gamma}{4} \int_{-1}^{1} \sqrt{q_\gamma(z)} h'(z)\, dz \\
&\quad = \frac{3\gamma}{4} \int_{-1}^{1} \sqrt{q_\gamma(z)} (1 - z^2)\, dz > 0.
\end{aligned}$$

Therefore $C_\gamma > 0$.

Finally, since $n_0/n_* \to_p \pi_0$ and $M_n \to \infty$,

$$\begin{aligned}
\sqrt{n_*}\, \mathbb{E}[\widehat{\theta}_{0,\mathrm{raw}} - \theta_0] &\asymp \sqrt{n_*} \left(\frac{M_n}{n_0} \right)^{1/2} \\
&\asymp \sqrt{M_n} \longrightarrow \infty.
\end{aligned}$$

Thus the raw matching estimator does not have a centered root-n limit in this example, despite the exact propensity score and the absence of score projection residual.

In contrast, the regression-adjusted estimator with the oracle choice $\widehat{g}(S) = \mu_S(S) = m_0(X)$ removes the leading matching discrepancy. Indeed, its matching term is $\Delta_{M_n}(m_0 - \widehat{g}) = 0$. This shows that the need for regression adjustment in dimension four can arise solely from the geometry of nearest-neighbor matching, even when the score representation itself is exact. □

B Additional Experimental Results

B.1 Increasing ambient covariate dimension

We examine how the estimators behave as irrelevant covariates are added to the learning problem. We use the nonlinear-confounding design in Section 4, with $n = 1{,}000$, $\psi = 1.75$, and 500 Monte Carlo replications. The ambient dimension is $P \in \{20, 50, 100, 200, 500, 1000\}$. In each replication, the data are generated once at $P = 1000$, and the designs at smaller dimensions use nested prefixes of the same covariate matrix. Thus, the treatment assignments and outcomes are held fixed as P increases; only additional independent noise covariates are introduced. The first ten coordinates contain all signal, so every coordinate beyond the tenth is irrelevant to both treatment assignment and the outcome.

For each training fold, the flexible propensity and prognostic candidates, $\widehat{e}_1$ and $\widehat{m}_{0,1}$, are estimated using all P available covariates through a two-stage procedure. We first rank covariates by the improvement in fit from a univariate cubic B-spline regression and retain the top 20; we then fit an additive cubic B-spline logistic regression for $\widehat{e}_1$ and an additive cubic B-spline ridge regression for $\widehat{m}_{0,1}$ on the selected covariates. The prognostic screening and regression use controls only. All screening and model fitting are conducted within the training fold. The working-model candidates $\widehat{e}_2$ and $\widehat{m}_{0,2}$ remain logistic and linear regressions on the raw covariates, respectively. The four-digit codes follow the ordering $(\widehat{e}_1, \widehat{e}_2, \widehat{m}_{0,1}, \widehat{m}_{0,2})$ used in Section 4. For example, `cm1010` uses $(\widehat{e}_1, \widehat{m}_{0,1})$, while `cm0101` uses only the two linear working-model candidates. The multi-score Cross Matching estimators include a linear regression adjustment in their score spaces; the single-score PSM and PGM comparators are unadjusted. Thus, matching uses at most four learned score coordinates, rather than all P raw covariates.

Table 1 reports the RMSE across ambient dimensions, computed as $\{500^{-1} \sum_{r=1}^{500} (\widehat{\tau}_{\mathrm{ATT}}^{(r)} - \tau_{\mathrm{ATT}})^2\}^{1/2}$ with $\tau_{\mathrm{ATT}} = 0$. Among the Cross Matching specifications, those including both flexible candidates have the smallest RMSEs throughout the comparison. For example, the RMSE of `cm1010` increases from 0.193 at $P = 20$ to 0.614 at $P = 1000$, while that of `cm1111` increases from 0.221 to 0.604. The AIPW specification is competitive at $P = 20$ and $P = 50$, but its RMSE rises sharply at larger dimensions. Several IPW/AIPW

specifications using the linear propensity candidate show similarly strong deterioration. By comparison, the difference-in-means and full-covariate matching baselines remain near an RMSE of 1.2. The specification `cm0101`, which uses only the two misspecified linear working models, approaches these baselines.

Table 1: RMSE by ambient covariate dimension.

Estimator	$P = 20$	$P = 50$	$P = 100$	$P = 200$	$P = 500$	$P = 1000$
Augmented inverse probability weighting						
`aipw0101`	0.7517	0.7884	1.5198	39.9326	43.5855	25.6787
`aipw0110`	0.1338	0.2481	0.9134	18.6697	32.3348	20.4579
`aipw1001`	3.5246	4.5324	5.2972	7.2752	8.2495	9.7410
`aipw1010`	1.5270	2.6506	3.5474	4.8555	6.3008	7.7746
Cross Matching						
`cm0101`	0.7556	0.7985	0.8929	1.0882	1.1427	1.1668
`cm0110`	0.2544	0.4139	0.5536	0.7124	0.8800	0.9658
`cm0111`	0.3294	0.4460	0.5590	0.7030	0.8686	0.9558
`cm1001`	0.3625	0.4506	0.5325	0.6576	0.6998	0.7377
`cm1010`	0.1930	0.2811	0.3633	0.4506	0.5469	0.6142
`cm1011`	0.2036	0.2816	0.3560	0.4417	0.5392	0.6078
`cm1101`	0.3945	0.4661	0.5313	0.6523	0.6982	0.7335
`cm1110`	0.1861	0.2814	0.3550	0.4397	0.5461	0.6156
`cm1111`	0.2210	0.2905	0.3531	0.4375	0.5363	0.6042
Inverse probability weighting						
`ipw0100`	0.7396	0.7487	1.2313	19.3018	59.4427	33.7764
`ipw1000`	3.3842	4.3582	5.0355	5.7698	5.8823	6.4212
Single-score matching and baselines						
`m.x`	1.0246	1.1159	1.1606	1.1845	1.2060	1.2184
`naive`	1.2302	1.2302	1.2302	1.2302	1.2302	1.2302
`pgm0001`	0.7876	0.8669	0.9735	1.1556	1.1691	1.1892
`pgm0010`	0.3319	0.5118	0.6423	0.7571	0.8999	0.9866
`psm0100`	0.8545	0.9589	1.0540	1.1563	1.2083	1.2147
`psm1000`	0.5557	0.6338	0.6690	0.7104	0.7406	0.7692

We further report the variable-screening diagnostics in Table 2. Recall and precision are computed relative to the nonzero coordinates of the corresponding data-generating coefficient vector, averaged over folds and replications; the selected-set size is fixed at 20. The propensity index has five active coordinates, whereas the prognostic index has ten. Accordingly, when $P = 20$ all variables are retained, giving recall one and precision $5/20 = 0.25$ for propensity screening and $10/20 = 0.50$ for prognostic screening. As P increases, propensity screening continues to recover almost all active coordinates: its recall is 0.987 at $P = 1000$. Prognostic screening is less accurate in the largest designs, with recall decreasing to 0.672 at $P = 1000$. Thus, the flexible learners are not given the true active sets; they must screen among all available covariates. Although prognostic screening becomes imperfect as noise variables are added, Cross Matching with both flexible scores retains comparatively low RMSE. This suggests that perfect recovery of the prognostic support is not necessary for favorable performance in this particular design.

Table 2: Variable-screening diagnostics for the full-dimensional learners.

P	Propensity screening			Prognostic screening		
	Recall	Precision	Size	Recall	Precision	Size
20	1.000	0.250	20	1.000	0.500	20
50	1.000	0.250	20	0.945	0.473	20
100	1.000	0.250	20	0.889	0.444	20
200	0.998	0.250	20	0.829	0.415	20
500	0.994	0.248	20	0.740	0.370	20
1000	0.987	0.247	20	0.672	0.336	20