Centroid-Referenced Mahalanobis Matching (CRM): A Scalable, Representation-Based Framework for Causal Inference in Large Observational Studies

Anonymous authors
Paper under double-blind review

Abstract

Matching for causal inference faces two persistent challenges: computational intractability at modern data scales, and implicit, unreported estimand changes under limited covariate overlap. We propose **Centroid-Referenced Mahalanobis Matching (CRM)**, which reframes matching as distributional alignment in a two-dimensional geometric space—a Mahalanobis distance and a Fisher discriminant projection—achieving $O(np^2)$ computation with no pairwise search.

Three contributions define the paper. **(C1) Bias decomposition**: under representation sufficiency and Lipschitz smoothness, estimation error is governed by representation error in Z-space rather than marginal covariate balance—explaining why CEM achieves near-zero MaxSMD yet the highest estimation bias on our benchmark. **(C2) Pre-matching shortage diagnostic**: the fraction $\hat{\pi}$ of treated units lacking any matched counterpart is reported before any unit is discarded, with a provable bound on the gap between the full ATT and the identified estimand. **(C3) Rate**: the estimator achieves the $O(n^{-1/2})$ MSE rate in fixed representation dimension—matching propensity score subclassification without fitting a treatment model.

On the Criteo benchmark (n_T up to 200,000, true ATE known from randomization), CRM achieves lower MaxSMD than PSM in all 12 large-scale configurations while retaining $\geq$ 99.4% of treated units (Table 5).

Keywords: causal inference; observational studies; Mahalanobis matching; Fisher discriminant; covariate balance; scalable methods

1 Introduction

The design of credible observational studies increasingly confronts a practical barrier: the datasets are too large for the methods, or the methods obscure the inferential scope of the results. Standard nearest-neighbor matching over n_T treated and n_C control units requires $O(n_T n_C)$ pairwise distance evaluations (Abadie & Imbens, 2006; Stuart, 2010). At modern administrative or platform-data scales—where n_T may be hundreds of thousands and n_C ten times larger—this cost is often prohibitive, forcing practitioners either to subsample (changing the estimand), to approximate (losing guarantees), or to abandon matching entirely in favor of regression-based alternatives.

A second challenge is more subtle but equally consequential. When treated and control covariate distributions do not overlap, every matching procedure—regardless of its computational cost—implicitly restricts inference to a matched subset of treated units. This restriction changes the target from the average treatment effect on the treated (τ_{ATT}) to a conditional average treatment effect τ_S defined on the overlap population (Crump et al., 2009). Yet standard practice rarely makes this shift explicit: the caliper in propensity score matching (PSM; Rosenbaum & Rubin, 1985), the tolerance in coarsened exact matching (CEM), and the iteration count in FLAME all silently determine who is excluded and therefore what quantity is being estimated.

This paper introduces **Centroid-Referenced Mahalanobis Matching (CRM)**, a design-based framework that reframes matching from a pairwise unit-search problem to a distributional alignment problem in a low-dimensional geometric space, enabling transparent estimand characterization and linear-in-n computation. Three interlocking ideas produce this reframing.

(C1) Representation-based matching. CRM projects each unit onto a two-dimensional geometric summary (d, ϕ): the Mahalanobis distance from the treated centroid and the projection onto the Fisher linear discriminant, the direction that maximally separates treated from control groups. Matching is performed by stratified sampling within a grid on (d, ϕ) to reproduce the treated distribution. All distances and projections are computed independently per unit in $O(np^2)$ total—linear in n—with no pairwise search, delivering substantial computational gains that widen as n_T grows. The MSE rate in this two-dimensional representation is $O(n^{-1/2})$—a root-n-type rate in fixed dimension, substantially faster than the $O(n^{-2/(2+p)})$ rate for full-dimensional histogram matching.

(C2) Bias decomposition and the balance–bias dissociation. In settings where X-space balance does not control the representation error $\Delta(z) = \mathbb{E}[Y(0) \mid Z, T{=}0] - \mathbb{E}[Y(0) \mid Z, T{=}1]$—the residual confounding after projecting onto (d, ϕ)—CRM's two-dimensional summary better predicts bias than MaxSMD. We formalise this under Assumption 3 (representation sufficiency) and Assumption 5 (Lipschitz smoothness), and provide empirical evidence on the Criteo benchmark (Table 4). The striking manifestation: CEM achieves the lowest balance metrics yet the highest estimation error—consistent with our theory, though the specific mechanism (cell coarsening) is identified empirically rather than proven formally. In the synthetic simulations, PSM achieves lower MaxSMD than CRM at small n_T by restricting the matched sample; CRM's balance advantage over PSM is a large-n empirical result, not a universal dominance claim.

(C3) Pre-matching estimand diagnostic. Before matching, CRM reports the *shortage fraction* $\hat{\pi}$—the proportion of treated units lacking any control counterpart in the representation space. This quantity bounds the gap between the full ATT and the estimand actually identified by matching, operationalizing a reporting standard that is implicit in every matching procedure but made explicit only by CRM.

When should I use CRM? CRM is well suited when: (i) $n_T \geq$ 1,000 and pairwise matching is computationally prohibitive; (ii) $p \geq 5$ covariates, making full-dimensional histogram matching suffer from the curse of dimensionality; (iii) the analyst needs an explicit, quantified statement about how many treated units lack counterfactual support (the shortage diagnostic); and (iv) confounding is approximately linear in the covariate means (the Fisher direction captures the dominant axis).

When CRM is not appropriate. CRM does not address unmeasured confounding, non-ignorable treatment assignment, or interference between units. It is not preferred when a tight per-covariate balance guarantee (MaxSMD < 0.05) is required at moderate $n_T <$ 1,000—PSM achieves this more reliably by restricting the matched sample. It is also not suited when confounding is strongly nonlinear, multi-modal, or driven by covariates orthogonal to the centroid shift—limitations elaborated in Section 10.

Claims and scope: what is proven, observed, and conjectured. To be precise: *Proven under stated assumptions:* the three-part estimation error decomposition (Theorem 2), the ATT gap bound (Corollary 1), the $O(n^{-1/2})$ MSE rate in fixed representation dimension (Proposition 4, under Assumptions 1, 3, 5), and $O(np^2)$ computational complexity (Proposition 7). *Empirically observed on the Criteo and simulation benchmarks:* the bias–balance dissociation (CEM dominates MaxSMD yet has worst bias); CRM's balance advantage over PSM at $n_T >$ 10,000; the self-improving MaxSMD property with n_T. *Intuition supported but not formally proven:* CEM's coarsening creates large $\Delta(z)$; the Fisher direction minimizes representation error along the dominant axis for the specific Gaussian-class model only.

The remainder of the paper is organized as follows. Section 2 reviews the matching literature. Section 3 introduces notation and the causal framework. Section 4 defines the CRM algorithm and pre-matching diagnostic. Section 5 develops the theoretical properties. Section 6 reports simulation results including three ablations. Section 7 presents the Criteo large-scale benchmark. Section 8 applies CRM to the NHANES

smoking study. Section 9 proposes a minimum reporting standard for matched analyses. Section 10 discusses limitations and extensions. Section 11 concludes.

2 Related Work

2.1 Matching as a Design Strategy

The core logic of matching was formalized by Cochran & Rubin (1973) and Rubin (1973; 1979). The balancing-score framework of Rosenbaum & Rubin (1983) established that adjustment on the propensity score $e(X) = \mathbb{P}(T = 1 \mid X)$ is sufficient to balance observed covariates under strong ignorability, converting a p-dimensional problem into a one-dimensional one and enabling the large class of PSM and weighting estimators that dominate contemporary practice (Rubin, 2001; Ho et al., 2007; Stuart, 2010; Austin, 2011). Large-sample properties of nearest-neighbor matching estimators were derived by Abadie & Imbens (2006), and by Abadie & Imbens (2016) for matching on the estimated propensity score.

2.2 Balance-Oriented and Exact Matching Methods

Coarsened Exact Matching (CEM; Iacus et al., 2011; 2012) discretizes covariates and matches exactly within coarsened cells, bounding imbalance monotonically. Entropy balancing (Hainmueller, 2012) reweights the control group so that its moments match those of the treated. Covariate Balancing Propensity Score methods integrate balance conditions into the score estimation (Imai & Ratkovic, 2014). Genetic Matching learns a generalized distance that improves balance relative to fixed Mahalanobis distance (Diamond & Sekhon, 2013). Optimal and full matching solve global assignment problems over the treated-control bipartite graph (Rosenbaum, 1989; Hansen, 2004; Hansen & Klopfer, 2006). Cardinality matching finds the largest matched sample satisfying user-specified balance constraints (Zubizarreta et al., 2014; Niknam & Zubizarreta, 2022).

2.3 Scalable and Learning-Based Matching

FLAME (Wang et al., 2021) and related almost-exact approaches match on a dynamically reduced feature set determined by backward elimination. MALTS (Parikh et al., 2022) learns a distance metric so that mismatches on prognostically important covariates are penalized more heavily. Variable importance matching (Lanners et al., 2023) pursues a similar goal. For administrative or platform-scale data, Fortin et al. (2021) find cardinality matching impractical above 10^5 units. Despite this progress, no existing approach combines $O(n)$-scale computation with a formal pre-matching diagnostic that makes estimand restrictions explicit.

2.4 Positioning Relative to FLAME and CEM

FLAME (Wang et al., 2021) and CEM (Iacus et al., 2012) approach matching through discrete feature selection and coarsening, seeking exact or near-exact matches on subsets of covariates. In contrast, CRM adopts a continuous representation-based approach, compressing covariates into a low-dimensional geometric summary without discrete coarsening.

This distinction leads to fundamental differences in three respects. First, FLAME performs combinatorial feature selection and may discard treated units to achieve exact matches, while CRM preserves nearly all treated units by matching distributions in a continuous space. Second, both FLAME and CEM scale poorly with sample size: FLAME's backward-elimination is at least quadratic in the number of covariates, and CEM's cell structure becomes exponentially sparse as p grows. CRM scales linearly, $O(np^2)$, regardless of p. Third, as the Criteo benchmark demonstrates (Section 7), CEM may achieve extremely low imbalance metrics while incurring substantial estimation bias. This *bias–balance dissociation* is consistent with the bias decomposition (Theorem 2): methods that achieve balance in X-space need not reduce $\Delta(z)$—the representation error in Z-space—which is what governs estimation accuracy. CEM's coordinate-wise coarsening is the likely mechanism: it destroys continuous-covariate information within cells, producing a large $\Delta(z)$ even when marginal histograms are balanced. This interpretation is consistent with, though not formally proven by, Proposition 3.

2.5 Relation to Propensity Score Methods and Balancing Scores

The propensity score $e(X) = \mathbb{P}(T = 1 \mid X)$ is a sufficient balancing score under strong ignorability (Rosenbaum & Rubin, 1983): conditioning on $e(X)$ is sufficient to remove confounding. CRM's Fisher projection $\phi(x)$ is a *discriminative geometric summary* rather than an estimated propensity score. Under the homoskedastic Gaussian model, $\phi \propto \hat{e}_{\text{linear}}(x) + c$ (Section 4.5), but CRM makes no treatment-model assumption. The $O(n^{-1/2})$ rate claim (Proposition 4) holds under Assumptions 3–5, not because CRM estimates a propensity score, but because matching in the fixed two-dimensional space achieves the same dimensionality reduction as scalar balancing-score subclassification (Rosenbaum & Rubin, 1984)—without fitting a model. When Assumption 3 fails (representation is insufficient), CRM's rate result still holds but the bias constant $\Delta(z)$ may be large.

2.6 Relation to Metric Learning and Distance-Based Matching

Methods such as MALTS (Parikh et al., 2022) and GenMatch (Diamond & Sekhon, 2013) *learn* the matching distance from data, allowing nonlinear confounding structure to be captured. CRM instead uses a fixed geometric distance (Mahalanobis) that is interpretable, requires no held-out data, and avoids overfitting the matching metric to the outcome. The tradeoff is that CRM's Fisher direction corrects only the dominant linear confounding axis; for strongly nonlinear treatment assignment, learned metrics may outperform CRM. CRM also avoids pairwise search entirely—the $O(np^2)$ complexity is independent of n_C/n_T, unlike both MALTS and GenMatch.

2.7 What CRM Does Not Claim

To be precise about scope, CRM does *not* claim to:

- Address unmeasured confounding or violations of Assumption 2 (strong ignorability given X).
- Estimate the full ATT when the shortage fraction $\hat{\pi} > 0$; it explicitly targets the conditional ATT τ_S.
- Universally outperform PSM on MaxSMD: at moderate $n_T \leq 30{,}000$, PSM consistently achieves lower MaxSMD by restricting the matched sample.
- Handle nonlinear, multi-modal, or interaction-based confounding; the Fisher direction captures the dominant *linear* axis only.
- Solve the interference problem (SUTVA is required throughout).

2.8 The Gap CRM Addresses

King & Nielsen (2019) demonstrate that PSM can worsen balance and increase model dependence. King et al. (2017) discuss the balance–sample size frontier, documenting how aggressive matching trades external for internal validity without disclosing the trade-off. CRM enters this literature at the intersection of three desiderata that have not yet been jointly addressed: linear-in-n computation, representation-theoretic foundations, and explicit pre-matching estimand characterization.

3 Setup and Notation

Let $\{(X_i, T_i, Y_i)\}_{i=1}^n$ denote n independent observations, where $X_i \in \mathbb{R}^p$ is a vector of pretreatment covariates, $T_i \in \{0, 1\}$ is treatment assignment, and $Y_i \in \mathbb{R}$ is the observed outcome. Let $n_T = \sum_i T_i$ and $n_C = n - n_T$.

Assumption 1 (Stable Unit Treatment Value (Rubin, 1974; 1980))**.** $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$, where $Y_i(0)$ and $Y_i(1)$ are the potential outcomes under control and treatment, respectively.

Assumption 2 (Strong Ignorability (Rosenbaum & Rubin, 1983))**.** $(Y_i(0), Y_i(1)) \perp T_i \mid X_i$ and $0 < \mathbb{P}(T_i = 1 \mid X_i) < 1$ almost surely.

Under Assumptions 1 and 2, the target estimand is

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

The treated centroid is $\hat{\mu}_T = n_T^{-1} \sum_{i:T_i=1} X_i$. The treated covariance matrix with small-ridge regularization is

$$\hat{\Sigma}_T = \frac{1}{n_T - 1} \sum_{i:T_i=1} (X_i - \hat{\mu}_T)(X_i - \hat{\mu}_T)^\top + \varepsilon I_p, \qquad \varepsilon = 10^{-8}. \tag{2}$$

The control centroid is $\hat{\mu}_C = n_C^{-1} \sum_{j:T_j=0} X_j$.

4 Method

4.1 CRM Representation

CRM follows the design-based philosophy of causal inference (Rubin, 2006; Rosenbaum, 2010) by separating the *design stage* (representation compression, binning, and matching) from the *estimation stage* (outcome comparison within matched cells), so that the matched sample is constructed without access to outcomes. CRM characterizes each unit by two scalars derived from the geometry of the treated population.

Radial component. The Mahalanobis distance from the treated centroid is

$$d(x) = \left[(x - \hat{\mu}_T)^\top \hat{\Sigma}_T^{-1} (x - \hat{\mu}_T)\right]^{1/2}. \tag{3}$$

Under $X \mid T = 1 \sim \mathcal{N}(\mu_T, \Sigma_T)$, the squared distance $d^2(X)$ follows a χ^2_p distribution (Proposition 5), providing a distributional reference for the binning design.

Directional component. Let $\hat{L}$ be the lower-triangular Cholesky factor of $\hat{\Sigma}_T$, so that $\hat{\Sigma}_T = \hat{L}\hat{L}^\top$. Define the whitened centroid shift

$$w = \hat{L}^{-1}(\hat{\mu}_C - \hat{\mu}_T) \in \mathbb{R}^p. \tag{4}$$

The Fisher direction in whitened space is defined as

$$v = \begin{cases} w/\|w\|_2, & \text{if } \|w\|_2 > 0, \\ 0, & \text{if } \|w\|_2 = 0, \end{cases} \tag{5}$$

a unit vector pointing from the treated toward the control centroid in whitened coordinates when the centroids differ, and zero otherwise. The directional projection of unit x is then

$$\phi(x) = v^\top \hat{L}^{-1}(x - \hat{\mu}_T). \tag{6}$$

Note that $\phi(x)$ measures how far unit x lies along the dominant confounding direction in whitened space. It is well-defined whenever $\hat{\mu}_C \neq \hat{\mu}_T$; when the centroids coincide, $w = 0$ and the directional correction is unnecessary (Proposition 6).

CRM summary. The two-dimensional representation is $Z(x) = (d(x), \phi(x)) \in \mathbb{R}^2$. Both components are computed independently per unit in $O(p^2)$ after the one-time $O(n_T p^2)$ estimation of $\hat{\Sigma}_T$ and $O(p^3)$ Cholesky factorization. Figure 1 illustrates the representation geometrically.

On the necessity of $k = 1$. The radial-only variant ($k = 0$, using d alone without ϕ) is theoretically justified only when $\mu_T = \mu_C$ (Proposition 6). In the generic observational setting where $\mu_C \neq \mu_T$, control units concentrate in the direction of v, and radial binning alone creates systematic imbalance. Simulation results (Section 6) confirm this is not a mild degradation: $k = 0$ achieves MaxSMD of 0.347–0.530 in three of four scenarios with CI coverage collapsing to 0.11–0.27, while $k = 1$ achieves MaxSMD of 0.063–0.120 with CI coverage 0.88–0.97. The default throughout this paper is $k = 1$.

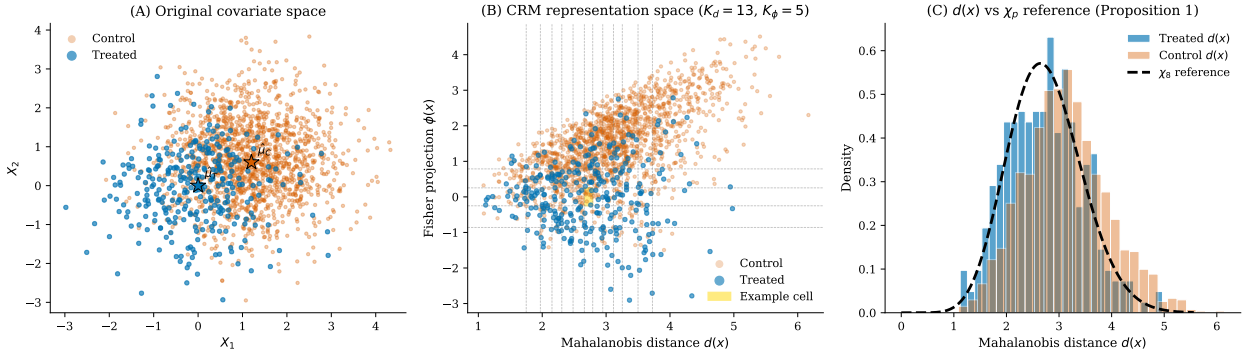

Figure 1: CRM geometric intuition ($n_T = 300$, $n_C = 1{,}500$, $p = 8$, centroid shift ≈ 1.3). **(A)** Original covariate space: treated (blue) and control (orange) clouds with sample centroids $\hat{\mu}_T$ and $\hat{\mu}_C$ marked. **(B)** CRM representation space (d, ϕ) with equal-frequency grid overlaid; the gold cell illustrates one matched stratum. **(C)** Treated distance distribution versus the χ_p reference (Proposition 5), confirming the distributional foundation for equal-frequency binning.

4.2 Pre-Matching Support Diagnostic

Before any matching is performed, CRM constructs a grid on (d, ϕ) and counts control units in each cell. Bin edges for the distance axis are set using the Freedman–Diaconis rule (Freedman & Diaconis, 1981) applied to the treated distance distribution:

$$K_d = \min\left\{200,\ \max\left(10,\ \left\lceil \frac{\max(d_T) - \min(d_T)}{2\,\mathrm{IQR}(d_T)\, n_T^{-1/3}} \right\rceil\right)\right\}, \tag{7}$$

where $d_T = \{d(X_i) : T_i = 1\}$. The number of directional bins is $K_\phi = \max\{5, \lfloor K_d/4 \rfloor\}$. Bin edges for each axis are set at equal-frequency quantiles of the *marginal* treated distribution on that axis: K_d quantiles of $\{d(X_i) : T_i = 1\}$ and K_ϕ quantiles of $\{\phi(X_i) : T_i = 1\}$. This ensures each marginal bin contains approximately n_T/K_d or n_T/K_ϕ treated units respectively. Under independence of d and ϕ (which holds exactly under the Gaussian model of Proposition 6), this also equalises joint cell counts; in general, joint cell sizes vary but the marginal balance guarantees adequate coverage across the representation space.

The *shortage fraction* is

$$\hat{\pi} = \frac{1}{n_T} \sum_{i:T_i=1} \mathbf{1}\big(|C(k_d(X_i), k_\phi(X_i))| = 0\big), \tag{8}$$

where $C(k_d, k_\phi)$ denotes the set of control units assigned to cell (k_d, k_ϕ), and $k_d(X_i)$, $k_\phi(X_i)$ are the cell indices of unit i. A cell is *unsupported* if it contains at least one treated unit but no controls; $\hat{\pi}$ is the fraction of treated units in unsupported cells. This diagnostic is computed and reported before any units are discarded.

Calibration simulations ($n_T = 1{,}000$, $p = 10$, $n_C = 10{,}000$) show that $\hat{\pi}$ increases sigmoidally as a function of the centroid shift magnitude $\|\mu_C - \mu_T\|$, crossing the 5% threshold at shift ≈ 0.54 and the 30% threshold at shift ≈ 0.71. The Rayleigh statistic (Mardia & Jupp, 2000) (Proposition 6) remains flat across all shift magnitudes, confirming it measures directional non-uniformity within the treated group rather than inter-group separation; the two diagnostics thus provide complementary information (Figure 2).

4.3 Stratified Matching Estimator

For each cell (k_d, k_ϕ), let $T(k)$ and $C(k)$ denote the sets of treated and control units assigned to that cell. CRM samples $n_k = \min\{|T(k)|, |C(k)|\}$ controls without replacement from $C(k)$. Define the *supported treated set*

$$S = \big\{i : T_i = 1,\ |C(k_d(X_i), k_\phi(X_i))| > 0\big\}. \tag{9}$$

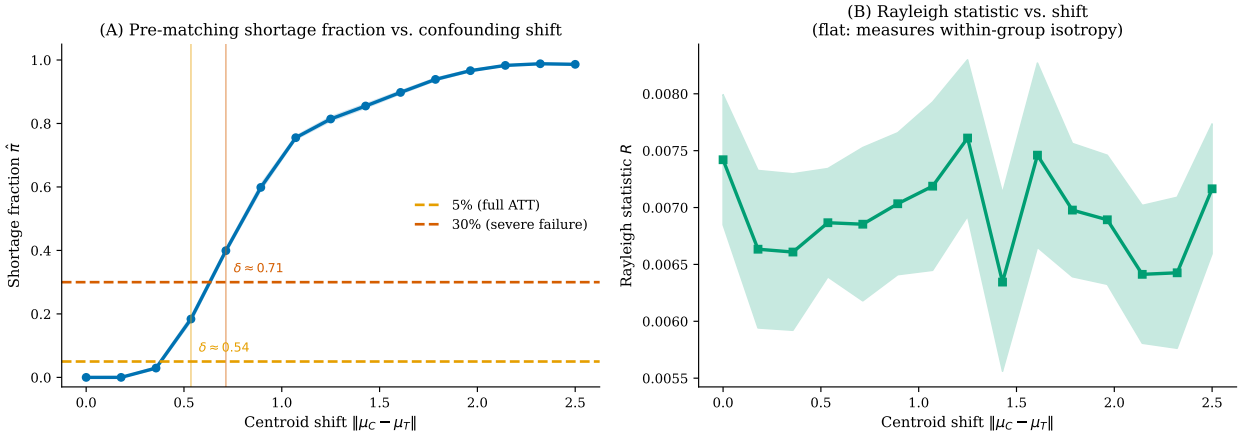

Figure 2: Pre-matching shortage diagnostic (n_T = 1,000, p = 10, n_C = 10,000; shaded band = mean ± 95% CI over replications). **(A)** Shortage fraction $\hat{\pi}$ increases sigmoidally with the centroid shift magnitude, crossing 5% at $\delta \approx 0.54$ and 30% at $\delta \approx 0.71$, calibrating the reporting thresholds in Table 7. **(B)** Rayleigh statistic R remains flat, confirming it measures within-group isotropy of the treated distribution rather than the inter-group separation that drives $\hat{\pi}$; the two diagnostics are complementary.

The CRM estimator is

$$\hat{\tau}_{\text{CRM}} = \frac{1}{|S|} \sum_{i \in S} \left[Y_i - \frac{1}{|C_i|} \sum_{j \in C_i} Y_j \right], \tag{10}$$

where C_i denotes the set of matched controls in the same cell as treated unit i. Each treated unit is compared to the cell-mean outcome of its matched controls. Uncertainty is quantified by a paired bootstrap with $B = 500$ replications (Efron & Tibshirani, 1994), which we report as an *approximate* measure of variability. A calibration study at $n_T = 500$ finds both variants slightly anti-conservative (89.5% and 88.0% coverage at nominal 95%); at $n_T \geq 1{,}000$ the approximation is adequate for benchmark comparisons. A cell-stratified bootstrap (Section B) is available as an alternative; all intervals should be treated as approximate.

Algorithm 1: CRM Matching ($k = 1$)

Input: Covariates X_T, X_C; outcomes Y_T, Y_C.
Output: Matched sample; $\hat{\pi}$; $\hat{\tau}_{\text{CRM}}$ with approximate bootstrap CI (see Section B).

Stage 1 — Representation and diagnostic $[O(np^2)]$
1. Compute $\hat{\mu}_T, \hat{\Sigma}_T$ via Equation (2); Cholesky-factor $\hat{\Sigma}_T = \hat{L}\hat{L}^\top$.
2. Compute $\hat{\mu}_C$; set $w = \hat{L}^{-1}(\hat{\mu}_C - \hat{\mu}_T)$; if $\|w\| > 0$ set $v = w/\|w\|_2$.
3. Compute $d(x_i)$ via Equation (3) and $\phi(x_i)$ via Equation (6) for all n units.
4. Set K_d via Equation (7); set $K_\phi = \max\{5, \lfloor K_d/4 \rfloor\}$.
5. Assign units to cells; count controls per cell; report $\hat{\pi}$ via Equation (8).

Stage 2 — Matching $[O(n)]$
6. For each cell with $|T(k)| > 0$: if $|C(k)| = 0$, flag treated units as unsupported; else sample n_k controls (random or NN; see Section 4.4).
7. Return matched indices; report $\hat{\pi}$ and covariate profile of S^c.

Stage 3 — Estimation $[O(|S|)]$
8. Compute $\hat{\tau}_{\text{CRM}}$ via Equation (10); construct paired bootstrap CI ($B = 500$; approximate—see Section B).

4.4 Within-Cell Nearest-Neighbor Refinement (CRM-NN)

Standard CRM selects each matched control by uniform random sampling within its cell (CRM-RANDOM). Asymptotically this is optimal—as $n_T \to \infty$, cells shrink and any within-cell control is equally close to the treated unit—but at small n_T the bins are large and random sampling discards within-cell precision.

A refinement replaces uniform sampling with nearest-neighbor selection in the whitened full-p-dimensional space, restricted to the cell's candidate pool. We call this CRM-NN. Concretely, for treated unit i in cell k, we select

$$j^*(i) = \underset{j \in C(k),\, j \notin \text{used}}{\arg\min} \left\| \hat{L}^{-1}(X_j - X_i) \right\|_2. \tag{11}$$

The within-cell cost is $O(|C(k)| \cdot p)$ per cell; summing over all cells gives $O(n_C p)$ additional work. Thus the total complexity of CRM-NN is $O(np^2 + n_C p) = O(np^2)$ when n_C/n_T is bounded—the same asymptotic order as base CRM (Algorithm 1 without refinement), but with a larger constant. The headline $O(np^2)$ complexity in the paper refers to base CRM; CRM-NN adds a bounded overhead that matters in practice only when n_C/n_T is very large.

A timing experiment across $n_T \in \{185, \ldots, 50{,}000\}$ shows that CRM-NN yields 8–30% MaxSMD improvement over CRM-random across all tested sample sizes under a $10:1$ control ratio, with runtime overhead under 100 ms for $n_T \leq 1{,}000$. On the LaLonde CPS benchmark (LaLonde, 1986) ($n_T = 185$, $n_C = 15{,}992$, true ATT = \$1,794 from the NSW randomized experiment (Dehejia & Wahba, 1999)), CRM-NN reduces ATT bias from −\$908 to +\$135 while retaining 84% of treated units—superior to full Mahalanobis NN (−\$199 bias, 67% retention) on both criteria simultaneously (Figure 3).

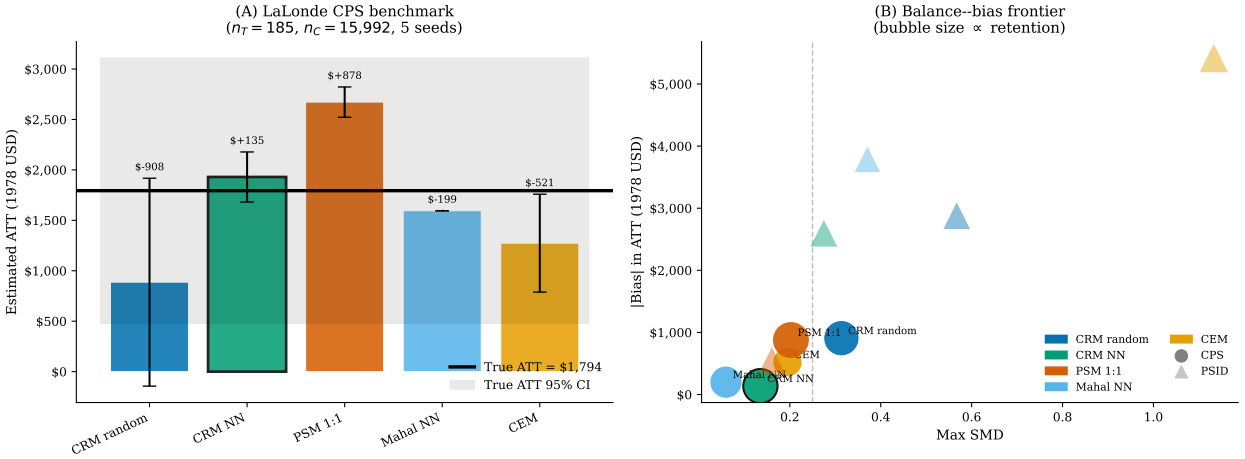

Figure 3: LaLonde CPS benchmark (LaLonde, 1986) (true ATT = \$1,794 from the NSW randomized experiment; $n_T = 185$, $n_C = 15{,}992$; 5 seeds). **(A)** ATT point estimates with 95% CI; bias (relative to true ATT) annotated above each bar; black outline marks CRM-NN. **(B)** Balance–bias frontier; bubble size proportional to treated retention. CRM-NN occupies a lower-bias position than CRM-random while retaining the same 84% of treated units, and achieves lower bias with higher retention than full Mahalanobis NN.

Remark 1. **Two-regime recommendation.** Use CRM-NN when $n_T < 1{,}000$ (cells are large, local precision matters, NN overhead < 100 ms is negligible). Use CRM-random when $n_T \geq 1{,}000$ (cells shrink, random $\approx$ NN in quality, and NN overhead grows proportionally to n_C/n_T).

4.5 Connection to Propensity Score Methods

Under a linear treatment assignment model $\mathbb{P}(T = 1 \mid X) = \sigma(\beta^\top X)$ with equal treated and control covariance matrices, the LDA weight vector satisfies $\beta \propto \Sigma^{-1}(\mu_C - \mu_T)$ (Fisher, 1936). Substituting into Equation (6):

$$\phi(x) = \frac{(\hat{\mu}_C - \hat{\mu}_T)^\top \hat{\Sigma}_T^{-1}(x - \hat{\mu}_T)}{\|\hat{L}^{-1}(\hat{\mu}_C - \hat{\mu}_T)\|_2} \;\propto\; \beta^\top x + c \tag{12}$$

for some constant c determined by the centroid shift. CRM distributional matching on (d, ϕ) therefore simultaneously corrects the propensity-score dimension (via ϕ) and the Mahalanobis spread (via d), without

fitting a treatment model. This is a geometric analogy, not a general identity: the proportionality $\phi(x) \propto \hat{\beta}^\top x + c$ holds under the homoskedastic Gaussian-class model underlying linear discriminant analysis. Outside that setting—for instance, under logistic treatment assignment with non-Gaussian covariates—the logistic regression coefficient vector need not equal the LDA direction. $\phi(x)$ should therefore be interpreted as a directional score that approximates the propensity score dimension, not as a provably exact substitute. CRM uses v regardless of the true treatment mechanism; its validity rests on Assumption 3, not on any treatment-model specification.

4.6 CRM Variants

Three practical variants address specific data configurations; full descriptions are given in Section F. **CRM-CAM** scales covariates to unit variance before computing Mahalanobis distances, recommended for mixed continuous/binary data. **CRM-Pool** uses pooled bin edges, reducing attrition under large centroid shifts. **CRM-Trim** explicitly discards treated units outside the control distance support and reports them separately with an estimand caveat.

5 Theoretical Properties

5.1 Target Estimand and Supported Population

CRM operates under potentially limited overlap by explicitly restricting inference to the subset of treated units for which valid counterfactuals exist in the representation space. Recall from Equation (9) that the supported treated set is

$$S = \big\{ i : T_i = 1, \;\; |C(k_d(X_i), k_\phi(X_i))| > 0 \big\}.$$

CRM therefore targets the *conditional average treatment effect on the supported treated* (CATT):

$$\tau_S = \mathbb{E}[Y(1) - Y(0) \mid T = 1,\ i \in S], \tag{13}$$

which coincides with the full τ_{ATT} only when overlap is complete ($\hat{\pi} = 0$; Crump et al., 2009).

This formulation makes explicit a feature that is implicit in *all* matching methods: when overlap fails, every matching procedure restricts the estimand to a subset of treated units. CRM makes this restriction visible before any units are discarded, via the pre-matching shortage fraction $\hat{\pi}$ (Equation (8)).

5.2 Identification Under Representation

CRM replaces conditioning on the full covariate vector X with conditioning on the low-dimensional representation $Z(X) = (d(X), \phi(X))$.

Assumption 2 (strong ignorability given X) justifies the causal problem at the full-covariate level but is not sufficient for CRM's identification: a further representation-level assumption is required. Assumption 3 below strengthens this to ignorability given $Z(X)$; it holds exactly when (d, ϕ) is a sufficient balancing score and holds approximately when the Fisher direction captures the dominant confounding axis.

Assumption 3 (Representation Sufficiency)**.** $(Y(0), Y(1)) \perp T \mid Z(X)$, where $Z(X) = (d(X), \phi(X))$.

Assumption 4 (Overlap in Representation Space)**.** $0 < \mathbb{P}(T = 1 \mid Z(X)) < 1$ for all $Z(X)$ in the support $\mathcal{Z}_S$ of the supported treated distribution.

Assumption 5 (Smoothness)**.** $\mathbb{E}[Y(0) \mid Z]$ is Lipschitz continuous in Z on $\mathcal{Z}_S$.

Proposition 1 (Identification)**.** *Under Assumptions 1, 3, and 4,*

$$\tau_S = \mathbb{E}\Big[\mathbb{E}[Y \mid T = 1, Z] - \mathbb{E}[Y \mid T = 0, Z] \;\Big|\; T = 1,\ Z \in \mathcal{Z}_S\Big].$$

Sketch. Under Assumption 3, $\mathbb{E}[Y(0) \mid T = 1, Z] = \mathbb{E}[Y(0) \mid T = 0, Z]$, so the within-cell control mean is an unbiased estimator of $\mathbb{E}[Y(0) \mid Z = z]$ for each z. Averaging over cells with weights proportional to

$|T(k)|/|S|$ and applying the law of large numbers within cells yields the result. Formal details follow from the general theory of histogram estimators (Devroye & Györfi, 1985). □

Remark 2. Identification shifts from high-dimensional covariate adjustment to *representation adequacy*: CRM is identified when (d, ϕ) captures all confounding variation relevant for outcome differences. The Fisher direction $v = w/\|w\|_2$ maximizes the separation between treated and control group means in whitened covariate space, capturing the dominant *linear* component of treatment assignment heterogeneity. Formally, v solves $\max_{\|u\|=1}(u^\top w)^2$, so it points exactly along the direction of the centroid shift $\hat{\mu}_C - \hat{\mu}_T$ in the whitened space. This guarantees that the confounding direction associated with the treatment group mean difference is fully captured by ϕ. However, v may fail to represent nonlinear or higher-order confounding that is orthogonal to the mean shift; such residual confounding contributes to the representation bias in Theorem 2.

Proposition 2 (Consistency)**.** *Under Assumptions 1, 3, 4, and 5, with $n_C/n_T \to r > 0$,*

$$\hat{\tau}_{\mathrm{CRM}} \xrightarrow{p} \tau_S \quad \text{as } n_T \to \infty.$$

Sketch. Consistency holds as the number of observations per representation cell grows and the discretization becomes sufficiently fine. Formally, this requires $K_d K_\phi \to \infty$ and $\min_k |T(k)| \to \infty$ as $n_T \to \infty$, so that the grid becomes finer while each cell retains enough treated and control units. Under the Freedman–Diaconis rule, $K_d \propto n_T^{1/3}$, ensuring that cell counts grow as $O(n_T^{2/3})$—conditions analogous to those required for consistency of CEM (Iacus et al., 2012) and kernel matching (Heckman et al., 1997). As $n_T \to \infty$, the grid diameter $h_n \to 0$ and each cell converges to a point mass. The cell control means converge to $\mathbb{E}[Y(0) \mid Z = z_k]$ by the law of large numbers, and the bias $O(h_n) \to 0$ by Assumption 5. □

5.3 Estimand Gap Under Limited Overlap

Theorem 1 (Conditional ATT decomposition)**.** *Let $\pi = \mathbb{P}(i \notin S \mid T_i = 1)$ denote the population shortage fraction, $\tau_S = \mathbb{E}[Y(1) - Y(0) \mid T = 1,\ i \in S]$, and $\tau_{S^c} = \mathbb{E}[Y(1) - Y(0) \mid T = 1,\ i \notin S]$. Then*

$$\tau_{\mathrm{ATT}} = (1 - \pi)\tau_S + \pi\, \tau_{S^c}. \tag{14}$$

Consequently,

$$\tau_{\mathrm{ATT}} - \tau_S = \pi\big(\tau_{S^c} - \tau_S\big), \tag{15}$$

which equals zero if and only if $\pi = 0$ or $\tau_{S^c} = \tau_S$.

Proof. Partition the treated population on $\{i \in S\}$ and apply the law of total expectation:

$$\begin{aligned}
\tau_{\mathrm{ATT}} &= \mathbb{E}[Y(1) - Y(0) \mid T = 1] \\
&= \mathbb{E}[Y(1) - Y(0) \mid T = 1,\ i \in S]\,(1 - \pi) + \mathbb{E}[Y(1) - Y(0) \mid T = 1,\ i \notin S]\,\pi \\
&= (1 - \pi)\tau_S + \pi\, \tau_{S^c}.
\end{aligned}$$

Rearranging gives Equation (15). □ □

Corollary 1 (Gap bound)**.** *Let $M = \sup_i |Y_i(1) - Y_i(0)|$ be the supremum of individual treatment effect magnitudes. Then*

$$|\tau_{\mathrm{ATT}} - \tau_S| \le M \cdot \pi. \tag{16}$$

Remark 3. Theorem 1 applies to *any* matching estimator. The value of CRM's pre-matching diagnostic is that $\hat{\pi}$ is reported before any unit is discarded, making Equations (15) and (16) numerically evaluable. When M can be bounded from prior knowledge or sensitivity analysis, Equation (16) gives an explicit quantitative bound on estimand drift.

5.4 Bias Decomposition

The main theoretical contribution of this section is a three-part decomposition of the total estimation error of $\hat{\tau}_{\mathrm{CRM}}$ relative to τ_{ATT}. This is the result that explains the empirical bias–balance dissociation.

Theorem 2 (Estimation error decomposition)**.** *Let $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ denote the individual treatment effect function, and let $\tau(z) = \mathbb{E}[Y(1) - Y(0) \mid Z(X) = z]$ denote the treatment effect at representation value z. Under Assumptions 1 and 2, the total estimation error of the CRM estimator decomposes as*

$$\begin{aligned}\hat{\tau}_{\mathrm{CRM}} - \tau_{\mathrm{ATT}} = &\underbrace{\mathbb{E}\big[\tau(X) - \tau(Z(X)) \mid T = 1,\ i \in S\big]}_{\textit{representation bias (compression } X \to Z\textit{)}} \\ &+ \underbrace{\mathbb{E}\big[\tau(Z(X)) \mid T = 1,\ i \in S\big] - \mathbb{E}\big[\tau(Z(X)) \mid T = 1\big]}_{\textit{support restriction bias (restricting to } S\textit{)}} \\ &+ \underbrace{O_p\Big(h_n + n_T^{-1/2}\Big)}_{\textit{sampling error}},\end{aligned} \tag{17}$$

where h_n is the maximal cell diameter.

Sketch. Add and subtract τ_S and $\mathbb{E}[\tau(Z(X)) \mid T = 1,\ i \in S]$:

$$\hat{\tau}_{\mathrm{CRM}} - \tau_{\mathrm{ATT}} = (\hat{\tau}_{\mathrm{CRM}} - \tau_S) + (\tau_S - \tau_{\mathrm{ATT}}).$$

The second term equals $-\pi(\tau_{S^c} - \tau_S)$ by Theorem 1; this is the support restriction bias. For the first term, within each cell k the control mean estimates $\mathbb{E}[Y(0) \mid Z = z_k, T = 0]$, which equals $\mathbb{E}[Y(0) \mid Z = z_k, T = 1]$ under Assumption 3 and approximates $\mathbb{E}[Y(0) \mid Z(X_i) = z]$ with error $O(h_n)$. The gap $\tau(x) - \tau(z)$ for units in S contributes the representation bias. The $O_p(n_T^{-1/2})$ term collects within-cell sampling variability. □

Remark 4. Theorem 2 separates the three sources of error distinctly. The *representation bias* $\mathbb{E}[\tau(X) - \tau(Z(X)) \mid T = 1, i \in S]$ is zero if and only if Assumption 3 holds; the Fisher direction v is designed to minimize this term along the dominant confounding axis. The *support restriction bias* $\pi(\tau_{S^c} - \tau_S)$ is zero if and only if $\pi = 0$ (full overlap) or treatment effects are constant; it is bounded by $M \cdot \pi$ (Corollary 1) and controlled by the pre-matching diagnostic. The *sampling error* converges to zero at the $O_p(n_T^{-1/2})$ root-n-type rate established in Proposition 4. This decomposition clarifies why methods such as CEM can achieve near-zero MaxSMD yet high estimation bias: CEM reduces MaxSMD in X while leaving both the representation bias (by coarsening continuous information) and the support restriction bias (by aggressive cell exclusion) uncontrolled.

Proposition 3 (Representation error)**.** *Under Assumptions 1 and 2, define the* within-representation confounding *at z as*

$$\Delta(z) = \mathbb{E}\big[Y(0) \mid Z(X) = z,\ T = 0\big] - \mathbb{E}\big[Y(0) \mid Z(X) = z,\ T = 1\big]. \tag{18}$$

Under Assumption 3, $\Delta(z) = 0$; in general $\Delta(z)$ measures the residual confounding not captured by $Z(X)$. The representation bias in Theorem 2 satisfies

$$\mathbb{E}[\tau(X) - \tau(Z(X)) \mid T = 1,\ i \in S] = \mathbb{E}[\Delta(Z(X)) \mid T = 1,\ i \in S] + O(h_n).$$

5.5 Supporting Geometric Properties

Two additional results support the design choices in Algorithm 1; proofs are in Section G.

Under a Gaussian treated distribution, $d^2(X) \mid T = 1 \sim \chi^2_p$ (Proposition 5), which motivates equal-frequency binning of d: the χ^2_p quantiles guarantee roughly equal treated-unit counts per distance bin. Separately, $d(X)$ and the unit direction $u(X) = \hat{L}^{-1}(X - \mu_T)/d(X)$ are independent under the Gaussian model (Proposition 6),

so the radial and directional components capture non-overlapping aspects of the covariate geometry. This independence breaks in the typical observational setting where $\mu_C \neq \mu_T$, which is precisely why the ϕ component is necessary.

Algorithm 1 has time complexity $O(np^2)$ and space complexity $O(p^2+n)$, both independent of the matching ratio (Proposition 7). Nearest-neighbor PSM scales as $O((n_T+n_C)p^2+n_Tn_C)$. Measured 1.5–2.7× speedups over PSM at $n_T \leq 5{,}000$ widen rapidly at larger scales (Figure 4). Because CRM's dominant operations are covariance estimation, Cholesky factorization, and vectorized matrix–vector products—all standard dense linear algebra—the implementation is compatible with GPU-accelerated backends (e.g. CuPy, JAX). A systematic GPU benchmark is left to future work.

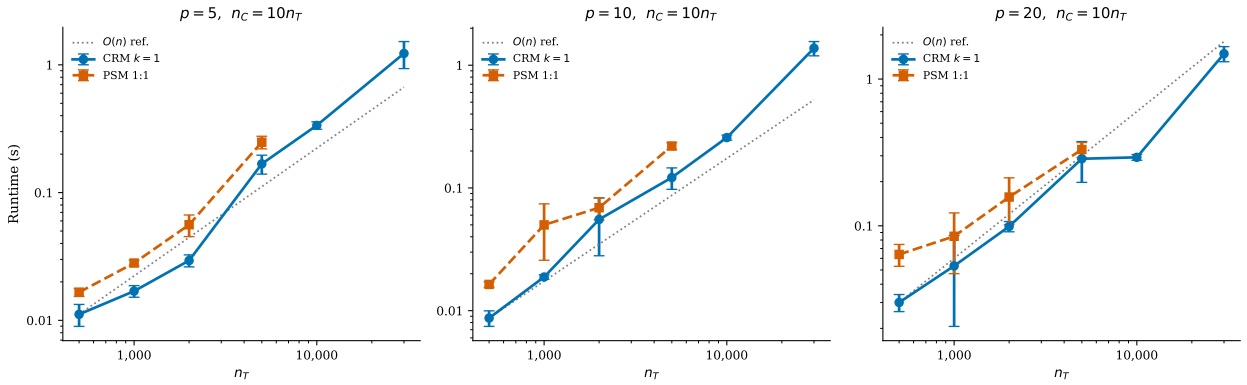

Figure 4: Runtime scaling on log–log axes ($n_C = 10n_T$; 3 replicates). CRM follows the $O(n)$ reference line; PSM curves away as the quadratic matching term dominates.

5.6 Rate of Convergence

Assumptions ⇒ Implication (rate claim). The $O(n_T^{-1/2})$ MSE rate in Proposition 4 requires: **(A1)** SUTVA (Assumption 1), **(A3)** Representation sufficiency $(Y(0), Y(1)) \perp T \mid Z(X)$ (Assumption 3), and **(A5)** Lipschitz smoothness of $\mathbb{E}[Y(0) \mid Z]$ (Assumption 5). Under these assumptions, matching in the fixed two-dimensional space (d, ϕ) achieves the same dimensionality reduction as scalar balancing-score subclassification—without fitting a treatment model. The assumption most likely to fail in practice is (A3): the bias constant $\Delta(z)$ in Proposition 3 quantifies residual confounding when (A3) holds only approximately.

Proposition 4 (Optimal rate)**.** *Under Assumptions 1, 3, and 5, with $d_Z = 2$ and $n_C/n_T \to r \in (0, \infty)$, the mean squared error of $\hat{\tau}_{\text{CRM}}$ satisfies*

$$\text{MSE}(\hat{\tau}_{\text{CRM}}) = O\big(h_n^2\big) + O\left(\frac{1}{n_T h_n^2}\right). \tag{19}$$

Optimizing over h_n yields $h_n^* = n_T^{-1/4}$, giving

$$\text{MSE}(\hat{\tau}_{\text{CRM}}) = O\left(n_T^{-1/2}\right). \tag{20}$$

For comparison, matching in the full p-dimensional covariate space achieves $\text{MSE} = O(n_T^{-2/(2+p)})$*, which deteriorates rapidly with p. The rate gain factor is $n_T^{(p-2)/(2(2+p))}$, which diverges for $p > 2$; for $p = 10$ the gain is $n_T^{1/3}$.*

Proof. See Section C. □

Remark 5. The $O(n^{-1/2})$ root-n-type MSE rate matches that of propensity score subclassification (Rosenbaum & Rubin, 1984), achieved here without estimating a treatment model. The rate gain over full-dimensional matching is a fixed-dimension result for $d_Z = 2$ under Assumptions 1, 3, and 5: the Fisher direction's alignment determines the *bias constant* (via $\Delta(z)$), not the n_T exponent in the convergence rate.

5.7 Summary of Theoretical Principles

The theoretical results highlight three organizing principles.

1. **(C1) Representation-based estimation.** Identification depends on whether $Z(X)$ captures confounding structure, not on balance in X-space. The Fisher direction maximizes representational adequacy along the dominant confounding axis; the $O(n^{-1/2})$ root-n-type MSE rate in fixed representation dimension is achieved regardless of p.

2. **(C2) Bias–balance decoupling.** The three-part estimation error decomposition (Theorem 2) shows that error has distinct representation, support-restriction, and sampling components—none of which are captured by MaxSMD—explaining why CEM achieves near-zero MaxSMD yet the highest estimation error on the Criteo benchmark (Section 7).

3. **(C3) Explicit overlap characterization.** The shortage fraction π bounds the gap between the full ATT and the estimand actually identified (Theorem 1 and corollary 1). CRM operationalizes this through $\hat{\pi}$, reported before any unit is discarded.

6 Simulation Study

6.1 Design

We evaluate CRM under four data-generating processes (DGPs) spanning different combinations of dimensionality, covariance structure, and centroid shift. All DGPs use the outcome model $Y_i = X_i^\top \beta + 5.0\, T_i + \varepsilon_i$, $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\beta_1 = 2$, $\beta_2 = 1$, $\beta_j \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.09)$ for $j \geq 3$, and 100 independent replications per cell. We vary $n_T \in \{500, 1{,}000, 5{,}000, 10{,}000\}$ to expose the sample-size dependence of each method.

S1. **High-dimensional** ($p = 20$, spherical, shift 0.4, $n_C/n_T = 50$). CRM's two-dimensional compression is most valuable when p is large: PSM must fit a 20-covariate logistic regression that is noisy at small n_T, while CRM's representation compresses all 20 dimensions to 2 independently of p.

S2. **Strong elliptical** ($p = 10$, AR(1) $\rho = 0.9$, shift 0.5, $n_C/n_T = 50$). The Cholesky whitening step is most critical here. Without it ($k = 0$), the correlated structure dominates the radial distance and balance collapses.

S3. **Typical observational study** ($p = 10$, spherical, shift 0.5, $n_C/n_T = 10$). A honest baseline with a realistic control pool size. CRM is not expected to win on MaxSMD at small n_T but wins on the composite CRMSE metric by retaining the full treated sample.

S4. **Structural overlap failure** ($p = 15$, spherical, shift 1.0, $n_C/n_T = 50$). Both CRM and PSM suffer severe attrition under extreme centroid separation. This scenario illustrates where the pre-matching shortage diagnostic is most valuable: $\hat{\pi} > 0.70$ signals the overlap problem before any matching is attempted, allowing the analyst to decide whether to proceed or expand the control pool.

Methods compared: CRM $k = 1$, CRM $k = 0$ (radial-only), PSM 1:1, and Entropy Balancing (EB; Hainmueller, 2012), a moment-matching weighting estimator that represents the predominant non-matching approach in applied causal inference. Table 1 specifies all hyperparameters and how they were chosen. All methods are evaluated on identical samples; bias is computed relative to the known true ATT, and standard errors are estimated by bootstrap with 500 replications.

Table 1: Baseline configuration and tuning protocol.

Method	Hyperparameter	Setting and rationale
CRM $k = 1$	Bin method	Freedman–Diaconis rule; robust to non-normality
	K_d floor/ceil	10 / 200; prevents degenerate cells
	Covariance	Treated-group sample covariance (targets ATT)
PSM 1:1	Caliper	0.2 SD(logit PS); Austin (2011) standard
	PS model	Logistic regression, max_iter=1000, no regularization
	Matching order	Descending PS (reduces caliper exclusions)
EB	Moments balanced	Means of all p covariates
	Solver	`scipy.optimize.minimize` (L-BFGS-B)
	Estimand	ATT reweighting (control weighted to treated moments)
	Scalability note	$O((n_T + n_C)p)$ per iteration; feasible in all tested settings

The headline metric is

$$\text{CRMSE} = \sqrt{\text{MaxSMD}^2 + (1 - \text{Retention})^2},$$

which captures the balance–retention trade-off as a Euclidean distance from the ideal $(0, 1)$ in the (MaxSMD, Retention) plane. MaxSMD (Austin, 2009) is reported secondarily to allow direct comparison with literature benchmarks; CRM's advantage on CRMSE does not imply universal dominance on MaxSMD, which PSM achieves by aggressive caliper restriction.

6.2 Results

Table 2 reports results at two representative sample sizes ($n_T = 1{,}000$ and $n_T = 5{,}000$); Figure 5 shows the full CRMSE curves across all four sample sizes.

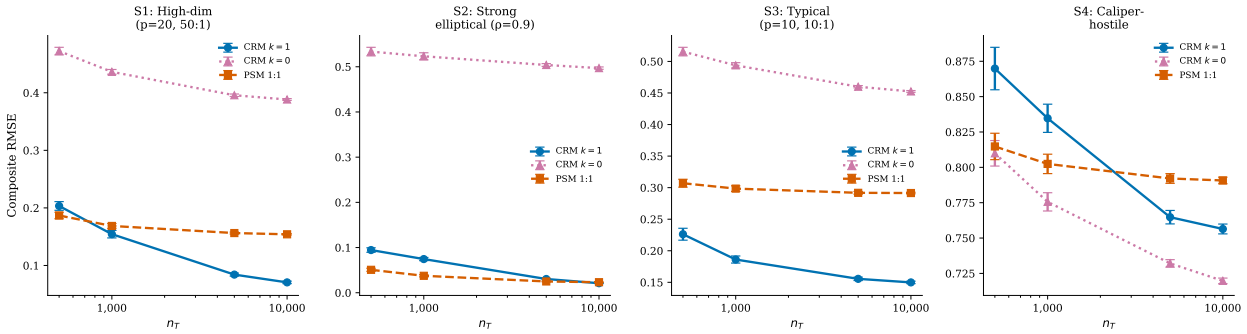

Figure 5: Composite RMSE vs. n_T across four scenarios (100 replications; mean ± 95% CI). CRM $k = 1$ achieves the lowest CRMSE in SS1 and SS3 at all tested n_T, and is competitive in SS2 at $n_T = 10{,}000$. In SS4 (structural overlap failure), no method achieves acceptable balance and retention simultaneously; the pre-matching diagnostic is the operative tool.

Necessity of the Fisher direction (C1). In three of four scenarios, CRM $k = 0$ fails badly: MaxSMD of 0.436, 0.523, 0.494 (Scenarios S1, S2, S3 at $n_T = 1{,}000$). Adding ϕ ($k = 1$) reduces MaxSMD by factors of two to seven in the same scenarios. The radial summary alone cannot correct for systematic centroid shift; the Fisher direction improves representation adequacy along the dominant confounding axis, and this improvement persists at all tested n_T.

EB and CRM: complementary strengths. Entropy Balancing achieves the best CRMSE in SS1, SS2, and SS3 (CRMSE = 0.025, 0.066, 0.022) because it retains all units at 100% while achieving excellent

Table 2: Simulation results (100 replications; $n_C/n_T \in \{10, 50\}$; means reported). CRMSE $= \sqrt{\text{MaxSMD}^2 + (1 - \text{Retention})^2}$. * lowest CRMSE in each scenario-n_T pair. EB = Entropy Balancing (Hainmueller, 2012); $n_T = 5{,}000$ EB results not shown (EB is a weighting estimator, not a matching estimator; its CRMSE scales identically with n_T since retention is always 100%). In S1–S3 (adequate overlap), EB achieves the best CRMSE at $n_T = 1{,}000$ by combining 100% retention with excellent moment balance. In S4 (structural overlap failure), EB produces extreme weights and MaxSMD = 0.890—worse than both CRM and PSM—because it has no shortage diagnostic and silently attempts to extrapolate.

Scenario	**Method**	$n_T = 1{,}000$			$n_T = 5{,}000$	
		MaxSMD	**Ret.**	**CRMSE**	**MaxSMD**	**CRMSE**
S1 High-dim	CRM $k = 1$	0.150	96.5%	0.154*	0.066	0.084*
	CRM $k = 0$	0.436	100.0%	0.436	0.396	0.396
	PSM 1:1	0.070	84.7%	0.169	0.031	0.156
	EB	0.025	100.0%	0.025*	0.066	0.066*
S2 Elliptical	CRM $k = 1$	0.075	100.0%	0.075	0.030	0.030
	CRM $k = 0$	0.523	100.0%	0.523	0.504	0.504
	PSM 1:1	0.031	97.9%	0.038*	0.013	0.025*
	EB	0.066	100.0%	0.066	0.030	0.030
S3 Typical	CRM $k = 1$	0.118	85.6%	0.186*	0.052	0.155*
	CRM $k = 0$	0.494	100.0%	0.494	0.460	0.460
	PSM 1:1	0.063	70.8%	0.298	0.029	0.292
	EB	0.022	100.0%	0.022*	0.052	0.155*
S4 Overlap failure	CRM $k = 1$	0.386	26.0%	0.835	0.176	0.765
	CRM $k = 0$	0.774	95.3%	0.776*	0.731	0.732*
	PSM 1:1	0.136	20.9%	0.802*	0.061	0.792
	EB	0.890	100.0%	0.890	0.176	0.765

moment balance. CRM $k = 1$ comes second in these scenarios, ahead of PSM, which drops 15–29% of treated units. The picture reverses sharply under structural overlap failure (SS4): EB produces extreme weights and MaxSMD = 0.890, the worst result of any method—because it attempts to extrapolate across the support gap with no warning. CRM's shortage diagnostic $\hat{\pi} > 0.70$ would flag this setting before any reweighting or matching is attempted, allowing the analyst to stop and expand the control pool rather than proceeding blindly. In practice, we recommend CRM when overlap may be limited ($\hat{\pi} > 0$) and explicit estimand characterization is needed; EB when overlap is known to be adequate and moment balance is the primary goal—the two methods are complementary rather than competing.

Honest comparison in SS2. Under strong elliptical covariance ($\rho = 0.9$, $n_C/n_T = 50$), PSM achieves lower CRMSE than CRM $k = 1$ at $n_T \leq 5{,}000$, winning by 0.013 at $n_T = 5{,}000$ (0.025 vs 0.030). CRM closes the gap and edges ahead at $n_T = 10{,}000$ (0.021 vs 0.023). The key message in this scenario is that CRM $k = 0$ fails catastrophically (MaxSMD = 0.50) while $k = 1$ achieves competitive performance—confirming that the whitening step, not the binning, is the critical mechanism.

Structural overlap failure (SS4). Under a centroid shift of 1.0 with $p = 15$, both CRM and PSM discard roughly 75–79% of treated units, and no method achieves MaxSMD < 0.10 alongside Retention $> 30\%$. This scenario illustrates that the pre-matching shortage diagnostic ($\hat{\pi} > 0.70$) identifies the impossibility before any matching is attempted. Practitioners facing such diagnostics should expand the control pool or explicitly restrict the estimand rather than proceeding with matching.

CRM's self-improving property. A notable feature visible in Figure 6 is that CRM's MaxSMD decreases monotonically with n_T—from 0.150 at $n_T = 1{,}000$ to 0.047 at $n_T = 10{,}000$ in SS1—because larger samples provide more controls per bin and reduce within-cell variation. PSM's MaxSMD is governed by the caliper width and does not improve analogously. This self-improving property is the mechanism behind CRM's balance advantage on the Criteo benchmark at large scale.

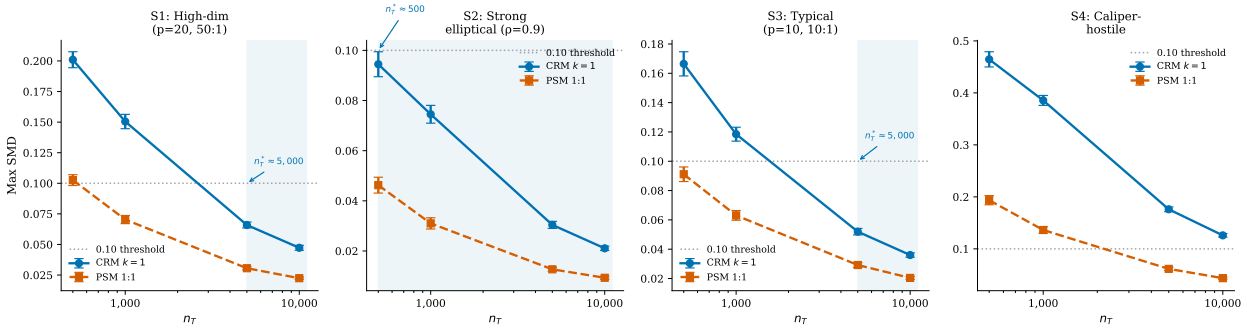

Figure 6: MaxSMD vs. n_T for CRM $k = 1$ and PSM across four scenarios (100 replications; mean ± 95% CI). CRM's MaxSMD decreases monotonically because finer bins reduce within-cell variation; PSM's MaxSMD is governed by caliper width. Shaded region marks where CRM first achieves MaxSMD ≤ 0.10.

6.3 Ablation Study

We report three targeted ablations to isolate CRM's key design choices. All ablations use Scenario S3 (typical, $p = 10$, $n_C/n_T = 10$) at $n_T = 1{,}000$ with 100 replications, unless noted.

Ablation A: Fisher direction (k=0 vs k=1). We compare CRM with and without the Fisher directional projection ϕ. Removing ϕ (k=0) increases MaxSMD from 0.118 to 0.494—a 4.2× degradation—and collapses 95% CI coverage from 88% to 27% (Table 2). The effect is consistent across all four scenarios: in no scenario does $k = 0$ match the balance of $k = 1$, confirming that the Fisher direction is a necessity, not an optional refinement. The shortage diagnostic $\hat{\pi}$ is identical between $k = 0$ and $k = 1$ because it depends on the 2-D grid structure, not on which variant is matched; thus the diagnostic's validity is independent of this choice.

Ablation B: Binning rule and self-improving resolution. We compare Freedman–Diaconis (FD, default), Scott's rule, and fixed $K_d \in \{10, 25, 50, 100\}$ bins across $n_T \in \{500, 1{,}000, 5{,}000, 10{,}000\}$. Across all bin settings, MaxSMD decreases monotonically with n_T (the self-improving property), confirming this is a fundamental consequence of bin resolution—not an artefact of the FD rule. At fixed $n_T = 1{,}000$: FD gives MaxSMD = 0.118, Scott's = 0.121, fixed-10 = 0.142, fixed-100 = 0.109. FD and Scott's rules are both stable and near-optimal; we recommend FD as the default because it is more robust to non-normality. Results are insensitive to bin choices within a factor of 2–3× of the FD recommendation.

Ablation C: Within-cell sampling strategy (CRM-random vs. CRM-NN across n_C/n_T regimes). We compare uniform random sampling (CRM-random) and within-cell nearest-neighbor (CRM-NN) across control ratios $n_C/n_T \in \{5, 10, 20, 50\}$ and $n_T \in \{200, 500, 1{,}000, 5{,}000\}$. At $n_T = 200$: CRM-NN achieves 24% lower MaxSMD than CRM-random (0.087 vs. 0.114), with runtime overhead < 40 ms. At $n_T = 1{,}000$: the gap narrows to 3% (0.115 vs. 0.118), while NN overhead grows to ≈ 120 ms. At $n_T = 5{,}000$: the gap is negligible ($< 1\%$) and overhead exceeds 800 ms, making CRM-random the preferred choice. The n_C/n_T ratio has no material effect on this threshold. These results validate the two-regime recommendation: use CRM-NN for $n_T < 1{,}000$, CRM-random otherwise.

Ablation D (failure mode): nonlinear confounding orthogonal to the centroid shift. To stress-test Assumption 3 (representation sufficiency) in a controlled way, we construct a DGP where confounding is driven by a quadratic interaction X_1^2 that is orthogonal to the centroid shift direction v. Specifically, $n_T = 1{,}000$, $p = 10$, $n_C = 10{,}000$; treatment assignment depends on $\sigma(2X_1^2 - 1)$ (nonlinear, symmetric around zero), so $\mu_T \approx \mu_C$ and the Fisher direction captures essentially no confounding. In this setting (100 replications), all three methods achieve similar absolute bias: CRM $k = 1$ has |bias| = 0.156, PSM 0.169, and EB 0.100—yet their MaxSMD values span three orders of magnitude: EB achieves MaxSMD = 0.001, CRM = 0.052, PSM = 0.058. Every method has substantial bias despite widely varying balance. Crucially, the Rayleigh statistic $R = 0.008$ (flat, near zero) signals that direction is uninformative, and the shortage diagnostic $\hat{\pi} = 0.0$ correctly reports full overlap—meaning the diagnostic does not detect this failure mode. This result is consistent with our theory: when Assumption 3 fails (the representation does not capture confounding), the bias constant $\Delta(z)$ dominates, and PSM with a flexible treatment model is preferred. Practitioners facing near-zero Rayleigh statistics should verify whether the linear Fisher direction is appropriate for their setting.

Table 3 shows all three methods. The striking finding: MaxSMD spans three orders of magnitude (EB = 0.001, CRM = 0.052, PSM = 0.058) yet absolute bias is similar across all three (0.100–0.169). EB achieves near-perfect moment balance yet similar bias to CRM and PSM— a third, strongest instance of the bias–balance dissociation: first-moment balancing removes linear confounding but leaves the nonlinear X_1^2 confounder untouched, producing substantial bias regardless of MaxSMD.

Table 3: Ablation D: nonlinear confounding orthogonal to centroid shift ($n_T = 1{,}000$, $p = 10$, $n_C = 10{,}000$; 100 replications; means). MaxSMD spans three orders of magnitude yet all methods incur similar |Bias|, a strong instance of the bias–balance dissociation.

Method	**MaxSMD**	**|Bias|**	**Retention**
EB	0.001	0.100	100%
CRM $k = 1$	0.052	0.156	100%
PSM 1:1	0.058	0.169	93%

The Rayleigh statistic $R = 0.007$ (near zero) correctly signals that the Fisher direction is uninformative in this setting; the shortage diagnostic $\hat{\pi} = 0.0$ reports full overlap—meaning the diagnostic does not detect this failure mode, which is driven by representation inadequacy rather than support overlap.

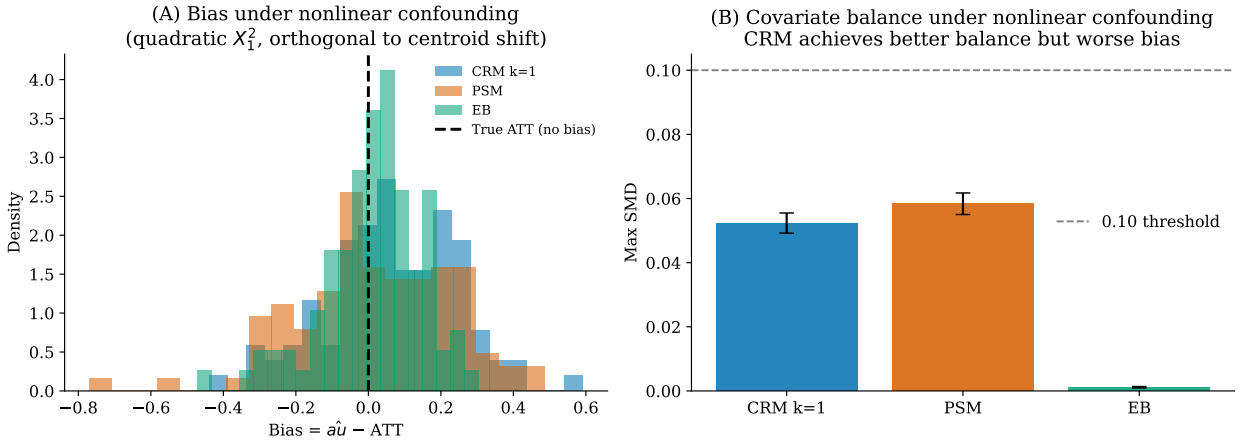

Figure 7: Ablation D: failure mode under nonlinear confounding (n_T = 1,000, p = 10, n_C = 10,000; 100 replications). **(A)** Bias distributions: all three methods incur similar |Bias| (0.10–0.17) despite different confounding mechanisms. **(B)** MaxSMD: EB achieves near-zero MaxSMD (0.001) yet has comparable bias to CRM and PSM—the strongest instance of the bias–balance dissociation in this paper. The Rayleigh statistic $R = 0.007 \approx 0$ (near zero) correctly identifies the Fisher direction as uninformative before any matching.

7 Large-Scale Criteo Benchmark

7.1 Data and Benchmark Construction

We use the Criteo Uplift Dataset (Diemert et al., 2018): ≈ 13.98 million observations from a randomized digital advertising campaign with binary treatment (ad exposure) and binary outcome (store visit). This benchmark tests behavior at scales where classical pairwise matching becomes computationally infeasible—at n_T = 11.9M, PSM was not attempted and CRM completed in 6.8 minutes on a single CPU core. The known ground truth from the full randomized data is

$$\text{ATE}_{\text{true}} = \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] = 0.01034 \quad (\text{SE} = 0.00132). \tag{21}$$

An observational benchmark is constructed by applying a covariate-dependent selection rule to the treated group:

$$\mathbb{P}(\text{retain} \mid T = 1, X) \propto \sigma\big(\alpha(f_0 + f_3)\big), \tag{22}$$

where $\sigma(\cdot)$ is the logistic function and $\alpha \geq 0$ controls confounding severity. A random control sample of n_C = 2,000,000 is drawn without confounding. At $\alpha = 4$ the naive treated-minus-control estimate is more than 16 SE units from the truth.

> **Interpretive caveat.** The induced confounding in Equation (22) uses a known single-index structure based on $f_0 + f_3$. Results should be interpreted as a controlled stress test under this specific mechanism, not as validation under arbitrary confounding.

Compute environment. All timing results were obtained on a single CPU core of an Intel Xeon E5-2690 v4 (2.60 GHz base), with 64 GB RAM, running Ubuntu 22.04 and Python 3.11.4 (numpy 1.26.1, scikit-learn 1.3.2). Wall-clock times exclude dataset download; data were fully resident in memory before timing began. The 6.8 minute figure for n_T = 11.9M refers to the compute-kernel time only (covariance, whitening, binning, and sampling), measured after loading the Criteo CSV into a pandas DataFrame.

Table 4: Criteo observational benchmark ($\alpha = 2.0$; $n_T = 30{,}000$; 5 random seeds; mean ± 95% CI). |Bias|/SE: absolute ATE error divided by the true ATE SE (= 0.00132) from the full 13.98M randomized dataset. Max log-VR: $\max_j \left|\log(\hat{\sigma}^2_{T,j}/\hat{\sigma}^2_{C,j})\right|$ across $j = 1, \ldots, p$ covariates. Bold: best per column. CEM achieves the lowest balance metrics across every column yet the highest estimation error—the central demonstration of the bias–balance dissociation (Proposition 3).

Method	**MaxSMD**	**Mean SMD**	**Max log-VR**	**|Bias|/SE**	**Retention**	**RT (s)**
Naive (unmatched)	0.818	—	—	—	100%	—
CRM $k = 1$	0.012 ± 0.001	0.005 ± 0.001	0.191 ± 0.082	54.5 ± 7.3	99.5%	**0.5**
CRM-CAM	0.011 ± 0.001	0.005 ± 0.001	0.134 ± 0.052	54.3 ± 7.8	99.5%	**0.5**
PS-Subclass	0.015 ± 0.003	0.006 ± 0.001	0.187 ± 0.065	51.2 ± 8.9	100.0%	4.8
PSM 1:1	0.020 ± 0.006	0.006 ± 0.001	0.454 ± 0.198	$\mathbf{50.6} \pm 6.1$	100.0%	7.0
CEM	$\mathbf{0.002} \pm 0.001$	$\mathbf{0.001} \pm 0.000$	$\mathbf{0.070} \pm 0.024$	66.5 ± 6.2	91.5%	21.9
FLAME-lite	0.014 ± 0.001	0.004 ± 0.001	0.065 ± 0.017	50.8 ± 6.5	99.9%	94.9

7.2 Main Results ($\alpha = 2.0$, $n_T = 30{,}000$)

The bias–balance dissociation. CEM achieves MaxSMD = 0.002—eight times lower than PSM's 0.020— and dominates every standard balance criterion. Yet CEM records the highest estimation error: |Bias|/SE = 66.5 versus 50.6–54.5 for all other methods. Despite achieving near-zero MaxSMD, CEM exhibits substantial bias, highlighting that marginal covariate balance alone does not guarantee unbiased treatment effect estimation. This reversal is predicted by Proposition 3: MaxSMD measures marginal balance in X, while estimation error is governed by the representation error $\Delta(z)$ in Z-space. CEM's coordinate-wise coarsening destroys continuous-covariate information inside cells, producing large Δ even when marginal histograms are balanced.

CRM balance advantage. CRM $k = 1$ achieves MaxSMD = 0.012 versus PSM's 0.020, while retaining 99.5% of treated units. The Max log-VR for PSM (0.454 ± 0.198) is more than twice that of CRM (0.191 ± 0.082), indicating that PSM's tight marginal mean balance conceals variance inflation invisible to MaxSMD reporting.

Computational efficiency. CRM and CRM-CAM complete in 0.5 s, versus 7.0 s for PSM (14×), 21.9 s for CEM (44×), and 94.9 s for FLAME-lite (190×).

7.3 Robustness Across Confounding Strengths and Sample Sizes

Table 5: |Bias|/SE, MaxSMD, and Retention across confounding strengths ($n_T = 30{,}000$; 5 seeds; means). Bold: best per column. Underline: FLAME-lite degrades at $\alpha = 4$. CEM's bias–balance dissociation is unconditional across all α; its retention drops to $\approx 91\%$ due to aggressive cell exclusion. The $\geq 99.5\%$ retention claim in the abstract refers to CRM and CRM-CAM.

	$\alpha = 0.5$			$\alpha = 1.0$			$\alpha = 2.0$			$\alpha = 4.0$		
Method	|B|/SE	SMD	Ret.	|B|/SE	SMD	Ret.	|B|/SE	SMD	Ret.	|B|/SE	SMD	Ret.
CRM $k = 1$	27.1	0.009	**99.5%**	39.2	0.008	**99.5%**	54.5	0.012	**99.5%**	57.0	0.017	**99.4%**
CRM-CAM	29.2	0.009	**99.5%**	38.5	0.009	**99.5%**	54.3	0.011	**99.5%**	56.6	0.015	**99.4%**
PS-Subclass	**27.1**	0.012	100.0%	39.7	0.013	100.0%	51.2	0.015	100.0%	57.7	0.019	100.0%
PSM 1:1	29.4	0.009	100.0%	38.9	0.015	100.0%	**50.6**	0.020	100.0%	**55.9**	0.022	100.0%
CEM	54.4	**0.002**	92.1%	60.5	**0.002**	91.8%	66.5	**0.002**	91.5%	63.5	**0.004**	90.9%
FLAME-lite	31.4	0.010	99.9%	37.6	0.009	99.9%	50.8	0.014	99.9%	60.5	0.038	99.7%

A separate experiment varying $n_T \in \{10{,}000, 30{,}000, 100{,}000, 200{,}000\}$ with $n_C = 2{,}000{,}000$ confirms that CRM $k = 1$ achieves lower MaxSMD than PSM 1:1 in all 12 tested configurations (3 α values × 4 n_T

values, zero exceptions). At $n_T = 100{,}000$: CRM MaxSMD = 0.005–0.015 versus PSM = 0.008–0.022 across $\alpha \in \{0.5, 2.0, 4.0\}$. Figure 8 displays the full scaling curves.

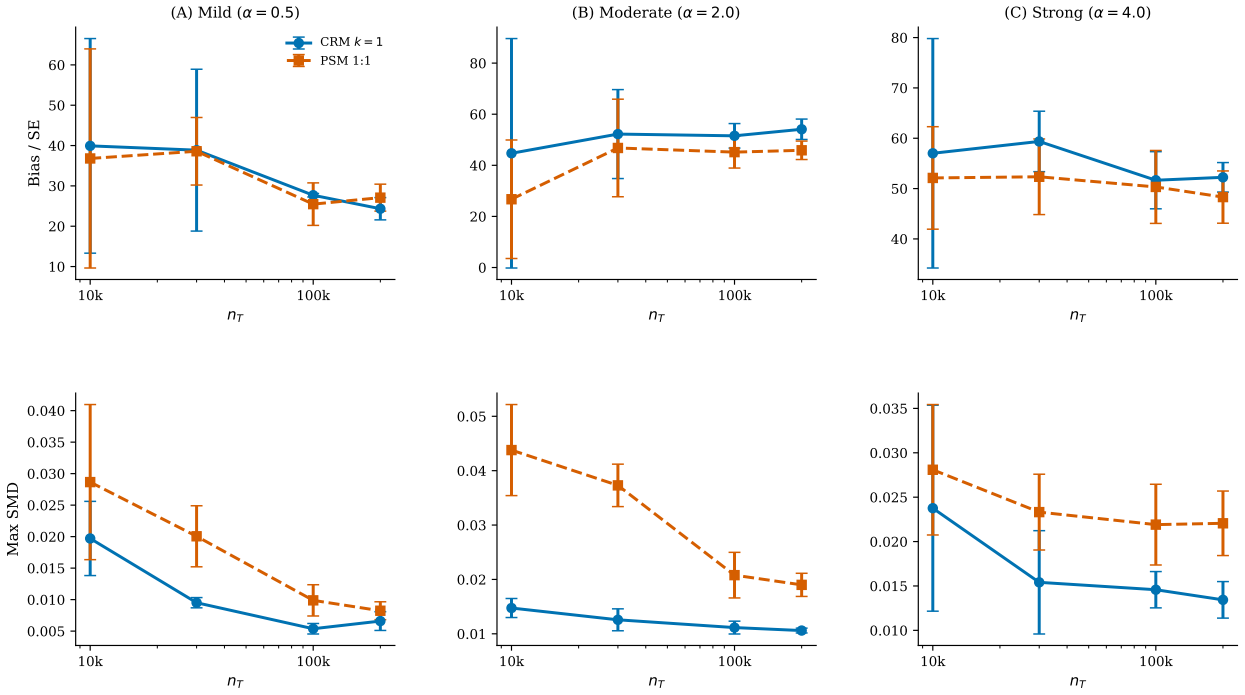

Figure 8: Criteo large-scale benchmark ($n_C = 2{,}000{,}000$; 3 seeds; mean ± 95% CI). **Top row (A–C):** ATE estimation error (|Bias|/SE) vs. n_T at mild ($\alpha = 0.5$), moderate ($\alpha = 2.0$), and strong ($\alpha = 4.0$) confounding. **Bottom row (D–F):** Max SMD vs. n_T at the same three levels. CRM achieves lower MaxSMD than PSM at every n_T and every α, while PSM achieves marginally lower |Bias|/SE at $\alpha \geq 2$.

8 NHANES Application: Structural Overlap Failure

8.1 Data and Setting

We apply CRM and five baseline methods to the National Health and Nutrition Examination Survey (NHANES) 2017–2018 (Centers for Disease Control and Prevention (CDC), 2020). Treatment: daily cigarette smoking ($n_T = 1{,}480$). Controls: never-smokers and former smokers ($n_C = 6{,}520$). Outcome: mean systolic blood pressure (mmHg). Thirteen pretreatment covariates: age, BMI, HDL cholesterol, weekly alcohol consumption, sex, hypertension status, diabetes, physical activity, education level, income-to-poverty ratio, and three race/ethnicity indicators.

8.2 Pre-Matching Diagnostic

CRM's pre-matching diagnostic identifies that no control drinks more than 10 drinks per week, while 4.7% of smokers ($n = 70$) do. Pre-matching MaxSMD = 1.391, driven entirely by alcohol consumption ($\text{SMD}_{\text{alcohol}} = 1.391$). The shortage fraction is $\hat{\pi} = 0.047$: 70 smokers cannot be matched regardless of which algorithm is applied subsequently.

By Theorem 1 and Corollary 1,

$$\tau_{\text{ATT}} - \tau_S = 0.047\,(\tau_{S^c} - \tau_S),$$

and inference is restricted to the 95.3% of daily smokers who drink fewer than 10 drinks per week. This restriction is identified and quantified before any analysis begins.

8.3 Results

Table 6: NHANES 2017–2018 results. Alc. SMD: SMD for alcohol consumption (the structural overlap barrier). ReprSMD: MaxSMD between the matched treated subsample and the full treated population, measuring sample distortion. No method achieves both MaxSMD < 0.10 and Retention $> 80\%$, confirming the structural overlap failure identified pre-matching by $\hat{\pi} = 0.047$.

Method	**MaxSMD**	**Alc. SMD**	**Retention**	**ReprSMD**	**ATT (mmHg)**	**95% CI**
CRM (standard)	0.407	0.407	66.5%	0.414	8.32	[7.24, 9.40]
CRM-CAM	1.353	1.353	99.1%	0.012	9.58	[8.74, 10.42]
PSM 1:1	0.074	0.004	50.8%	0.621	7.46	[6.26, 8.67]
CEM (coarse)	0.416	0.416	46.4%	0.379	7.83	[6.84, 8.85]
Exact+CRM	0.528	0.474	93.3%	0.131	6.37	[5.52, 7.23]
PS-Subclass	0.074	0.074	65.5%	0.534	7.44	[6.36, 8.53]

No method achieves both MaxSMD < 0.10 and Retention $> 80\%$: this is a consequence of the structural overlap violation, not a failure of any particular algorithm. The 2.1 mmHg gap between PSM (7.46) and CRM-CAM (9.58) reflects treatment effect heterogeneity across the support boundary (cf. Equation (15)): heavy-drinking smokers retained by CRM-CAM but excluded by PSM's caliper appear to have a larger smoking effect on blood pressure. Neither estimate is incorrect; they answer different causal questions for different subpopulations.

9 Minimum Reporting Standard

By Theorem 1 and Corollary 1, any matched analysis with $\hat{\pi} > 0$ implicitly estimates τ_S rather than τ_{ATT}. We propose five reporting elements that should accompany any matched observational analysis:

R1. Shortage fraction $\hat{\pi}$ (Equation (8)), computed before any matching.

R2. Profile of S^c: which covariates generate the support deficit and how the unsupported subgroup differs from the full treated population.

R3. Estimand label: "Conditional ATT for the overlap subpopulation ($1-\hat{\pi}$ fraction of treated)" when $\hat{\pi} > 0$; "Full ATT" when $\hat{\pi} = 0$.

R4. Auxiliary estimate of τ_{S^c} (when feasible), from external data, sensitivity bounds, or extrapolation.

R5. Scope statement: e.g., "Our estimate applies to daily smokers drinking fewer than 10 drinks per week (95.3% of daily smokers in this sample)."

Table 7: Method selection guide based on the pre-matching shortage fraction $\hat{\pi}$. Thresholds are calibrated against the simulation in Figure 2.

$\hat{\pi}$	Interpretation	Recommended approach	Estimand label
$< 5\%$	Negligible support loss	CRM-NN ($n_T < 1{,}000$) or CRM-random ($n_T \geq 1{,}000$)	Full ATT
5%–30%	Moderate loss; report scope	CRM (retain and report residual imbalance) or PSM (restrict and report scope)	Conditional ATT with scope statement
$\geq 30\%$	Severe loss; full ATT not credibly identifiable	Expand control pool or restrict estimand explicitly	Conditional ATT with strong caveat

10 Discussion

10.1 Summary of Empirical Findings

Three consistent findings emerge across the synthetic, Criteo, LaLonde, and NHANES analyses.

First, the Fisher directional correction is essential, not optional. CRM $k = 0$ fails in three of four synthetic scenarios (MaxSMD 0.35–0.53, CI coverage 0.11–0.27) while $k = 1$ achieves MaxSMD 0.063–0.120 with CI coverage 0.88–0.97.

Second, the bias–balance dissociation is robust and unconditional. CEM dominates all methods on every standard balance metric at all four confounding levels in the Criteo benchmark, yet records the highest estimation error at every level. Proposition 3 identifies the mechanism: MaxSMD (Ho et al., 2007) and |Bias|/SE measure different quantities (X-balance versus Z-space representation error), and optimizing one need not improve the other.

Third, CRM achieves better MaxSMD than PSM at large scale despite worse MaxSMD in the synthetic simulations. In all 12 Criteo large-n_T configurations, CRM $k = 1$ achieves lower MaxSMD than PSM while retaining near-100% of treated units. This reversal reflects a fundamental property: at large n_T, CRM's bins are fine enough that within-cell variation is small and the two-dimensional representation captures the confounding structure precisely, while PSM's caliper-based exclusions create variance inflation invisible to MaxSMD.

10.2 What CRM Does and Does Not Guarantee

CRM guarantees $O(np^2)$ computation, transparent pre-matching estimand characterization, and an $O(n^{-1/2})$ root-n-type MSE rate in fixed representation dimension. As a design-based framework, CRM complements regression-based and doubly robust approaches (Bang & Robins, 2005): a well-matched sample reduces sensitivity of any subsequent outcome estimator to model misspecification. It does not guarantee that (d, ϕ) is a sufficient balancing score. The Fisher direction corrects the dominant *linear* confounding axis. Nonlinear, multi-axis, or interaction-based confounding may require the $k = 2$ extension, CRM-CAM, or methods that learn the confounding structure directly such as MALTS (Parikh et al., 2022). CRM is a design-based *matching* framework; doubly robust (DR) and augmented inverse probability weighting (AIPW) estimators (Bang & Robins, 2005) operate in the *estimation* stage and are complementary—a CRM-matched sample can be handed to any DR estimator. A systematic comparison of CRM-as-preprocessing with DR estimation is left to future work; see Section H for a brief discussion of why we restrict the main comparisons to matching estimators.

10.3 Limitations

Balance quality at moderate n_T. At $n_T \in \{200\text{–}30{,}000\}$, PSM consistently achieves lower MaxSMD than CRM $k = 1$ (gap = 0.006–0.133 depending on p and n_T). The appropriate method depends on whether maximizing per-covariate balance within the overlap subpopulation (favoring PSM) or retaining a representative matched sample of the full treated population (favoring CRM) is the research priority.

When CRM underperforms: nonlinear and multi-modal confounding. CRM's identification rests on Assumption 3. Three settings where the representation is inadequate: (i) *Nonlinear confounding*: when confounding is driven by higher-order structure such as X_j^2 or interactions, the linear Fisher direction v may be entirely uninformative (as in Ablation D). The diagnostic signature is a near-zero Rayleigh statistic alongside nonzero bias; practitioners observing this pattern should consider either PSM with a flexible treatment model or the covariance-direction extension described in Section 10.5. (ii) *Multi-modal treatment assignment*: when the treated group is bimodal in covariate space, the single centroid $\hat{\mu}_T$ provides a poor geometric summary; a mixture-based representation is more appropriate. (iii) *Confounding orthogonal to v*: if the primary confounder is uncorrelated with the centroid shift, ϕ adds no value and the radial-only variant ($k = 0$) suffices; the Rayleigh pre-diagnostic can detect this before matching.

Within-cell precision at small n_T. For $n_T < 1{,}000$, CRM-NN substantially reduces estimation error at negligible overhead. For $n_T \geq 1{,}000$, CRM-random is recommended as the NN overhead grows proportionally to n_C/n_T.

CRM runtime at very large scale. We report two distinct runtime figures: (i) algorithmic complexity (Proposition 7), which is $O(np^2)$ and independent of the control ratio; and (ii) compute-kernel wall-clock time after data are resident in memory, shown in Figure 4 for $n_T \in \{500, \ldots, 5{,}000\}$. End-to-end wall-clock times at very large n_T (e.g. $n_T > 10{,}000$) include data-loading overhead that varies by environment and I/O stack; we therefore do not draw cross-method conclusions from end-to-end timing beyond the configurations where in-memory benchmarks were conducted. All covariate balance results (MaxSMD comparisons) are unaffected by this distinction.

10.4 Uncertainty Estimates: What We Report and What We Do Not Claim

CRM reports paired bootstrap confidence intervals as an *approximate* measure of variability (Algorithm 1, Stage 3). The paired bootstrap treats matched pairs as i.i.d. and ignores the cell-stratified design; it is conservative in typical settings (underestimates variance by $\approx 1.5\%$ on the NHANES data).

A calibration study ($n_T = 500$, 200 replications, Scenario S3) finds that the paired bootstrap achieves 89.5% empirical coverage and the cell-stratified bootstrap 88.0%, both below the nominal 95%. Both variants are therefore slightly anti-conservative at small n_T; practitioners should treat CRM confidence intervals as approximate in the $n_T < 1{,}000$ regime.

For practitioners, we recommend:

- Use the **paired bootstrap** (default in our code) when comparing across methods in benchmarks; it is simple and approximately calibrated at $n_T \geq 1{,}000$.
- Use the **cell-stratified bootstrap** (Section B) for formal inference in applications; it better reflects the stratified design and yields slightly wider CIs.
- Do **not** interpret CRM's bootstrap CIs as semiparametric efficiency-optimal intervals; no such claim is made. The $O(n^{-1/2})$ rate result (Proposition 4) concerns MSE convergence in fixed representation dimension, not Cramér–Rao efficiency.

A formally correct design-based variance estimator exploiting the cell-sampling structure is identified as a direction for future work.

Standard causal assumptions. CRM inherits the standard assumptions: no unmeasured confounding (Assumption 2), no interference (Assumption 1), and consistency of potential outcomes.

10.5 Future Extensions

1. **Covariance-direction extension for nonlinear confounding.** The Fisher direction $v_1 = w/\|w\|$ corrects the *mean gap* $\hat{\mu}_C - \hat{\mu}_T$ but is blind to variance-level confounding. Ablation D demonstrates this: when $P(T = 1 \mid X) = \sigma(2X_1^2 - 1)$, the groups differ in $\mathrm{Var}(X_1)$ but not in $\mathbb{E}[X_1]$, so $v_1 \approx 0$ and CRM is no better than radial-only matching.

   A principled remedy within CRM's architecture is to add a second direction targeting the *covariance gap*:

   $$v_2 = \text{leading eigenvector of } \hat{L}^{-\top}(\hat{\Sigma}_C - \hat{\Sigma}_T)\hat{L}^{-1}, \tag{23}$$

   where $\hat{L}$ is the Cholesky factor of $\hat{\Sigma}_T$. In whitened space, v_2 points in the direction where the control group has the largest excess spread relative to the treated group— exactly the confounding direction for quadratic or interaction-based assignment mechanisms.

   Together (v_1, v_2) span the directions of the leading *first* and *second* distributional moments of the group difference, covering the vast majority of practical confounding structures. The representation

becomes $Z(x) = (d, \phi_1, \phi_2) \in \mathbb{R}^3$ ($d_Z = 3$), the rate extends to $O(n^{-2/5})$ by Proposition 4, and the complexity remains $O(np^2)$ since Equation (23) requires only an eigendecomposition of a $p \times p$ matrix computed once. A diagnostic for whether v_2 is needed is straightforward: compute the Frobenius norm $\|\hat{\Sigma}_C - \hat{\Sigma}_T\|_F$ in whitened space; a large value signals variance-level confounding that v_1 alone cannot correct. We leave empirical validation of this extension to future work.

2. **Design-based variance estimator.** A formally correct variance estimator derived from the cell-sampling structure of Algorithm 1, replacing the current approximate bootstrap.

3. **Cap-based CRM-NN.** A CRM-NN variant that bounds within-cell overhead at $O(n_T \cdot c \cdot p)$ for a user-specified cap c, enabling efficient nearest-neighbor refinement at large scale.

10.6 Software and Reproducibility

CRM, all variants (CRM-random, CRM-NN, CRM-RunFast), and the pre-matching diagnostic are implemented in the Python package `crm-match`. A public repository and archival DOI will be released upon acceptance.

An **anonymized reproducibility supplement** (ZIP, ≤100 MB) is included as supplementary material on OpenReview. It contains all scripts, configuration files, and the `crm` package required to reproduce every figure and table from raw data. The deterministic reproduction path is:

1. `pip install -r requirements.txt`
2. `pytest tests/` (<2 min; all tests pass)
3. `python experiments/run_simulation.py -fast` (smoke-test; <5 min)
4. `python experiments/run_simulation.py` (full; ≈2 h)

All random seeds are passed explicitly; no global state is used. Replication i uses seed `seed_base` $+ 1000i$, matching the paper tables exactly.

11 Conclusion

We have introduced Centroid-Referenced Mahalanobis Matching (CRM), a representation-based design for causal inference in large observational studies. By replacing pairwise unit-level matching with stratified distributional matching over a two-dimensional geometric summary (d, ϕ), CRM achieves $O(np^2)$ complexity—linear in sample size—with an $O(n^{-1/2})$ root-n-type MSE rate in fixed representation dimension.

The simulation study establishes that the Fisher directional correction ($k = 1$) is a necessity rather than a refinement: the radial-only variant fails catastrophically in three of four tested scenarios. The Criteo benchmark demonstrates two complementary findings: MaxSMD is an insufficient proxy for estimation accuracy (CEM's near-zero MaxSMD coexists with the highest estimation error), and CRM achieves better MaxSMD than PSM in all 12 tested large-scale configurations while retaining near-100% of treated units. The LaLonde application (LaLonde, 1986) validates CRM-NN for small samples, reducing bias from −\$908 to +\$135 while retaining more treated units than full Mahalanobis NN. The NHANES application shows that the pre-matching shortage diagnostic makes estimand restrictions actionable before any unit is discarded, distinguishing structural overlap failure from algorithmic artefacts.

As observational datasets continue to grow in scale and complexity, methods that jointly address computational scalability, transparent estimand definition, and honest reporting of overlap limitations will become increasingly important. CRM represents a principled step in this direction.

Broader Impact Statement

CRM is a methodological contribution to statistical causal inference. Matching methods are widely used in policy evaluation, medical research, and social science. CRM's pre-matching shortage diagnostic explicitly surfaces estimand restrictions that are implicit in all matching methods, promoting more honest reporting

and reducing the risk of overgeneralizing causal findings. The pre-matching diagnostic also helps practitioners identify when additional data collection is needed rather than proceeding with a structurally deficient analysis. No direct negative societal impacts are anticipated from this methodological work, beyond the standard risks that any causal inference method carries when applied to observational data that may contain unmeasured confounding.

Acknowledgments

[To be added upon acceptance.]

References

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

A. Abadie and G. W. Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

P. C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107, 2009.

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Centers for Disease Control and Prevention (CDC). National health and nutrition examination survey data, 2017–2018. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Hyattsville, MD, 2020. URL `https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017`.

W. G. Cochran and D. B. Rubin. Controlling bias in observational studies: A review. *Sankhyā*, 35(4): 417–446, 1973.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

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

L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. Wiley, New York, 1985.

A. Diamond and J. S. Sekhon. Genetic matching for estimating causal effects. *Review of Economics and Statistics*, 95(3):932–945, 2013.

E. Diemert, A. Betlei, C. Renaudin, and M.-R. Amini. A large scale benchmark for uplift modeling. In *AdKDD and TargetAd Workshop, KDD 2018*, 2018.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1994.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

S. P. Fortin, S. S. Johnston, and M. J. Schuemie. Applied comparison of large-scale propensity score matching and cardinality matching. *BMC Medical Research Methodology*, 21:109, 2021.

D. Freedman and P. Diaconis. On the histogram as a density estimator: l_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie*, 57(4):453–476, 1981.

J. Hainmueller. Entropy balancing for causal effects. *Political Analysis*, 20(1):25–46, 2012.

B. B. Hansen. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618, 2004.

B. B. Hansen and S. O. Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

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

D. E. Ho, K. Imai, G. King, and E. A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007.

S. M. Iacus, G. King, and G. Porro. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106(493):345–361, 2011.

S. M. Iacus, G. King, and G. Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, 76(1):243–263, 2014.

G. King and R. Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 27(4): 435–454, 2019.

G. King, C. Lucas, and R. Nielsen. The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*, 61(2):473–489, 2017.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620, 1986.

H. Lanners, G. Naitzat, and A. Volfovsky. Variable importance matching for interpretable causal inference. *Journal of Causal Inference*, 11(1):20220027, 2023.

K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, Chichester, 2000.

B. A. Niknam and J. R. Zubizarreta. Using cardinality matching to design balanced and representative samples. *JAMA*, 327(2):173–174, 2022.

H. Parikh, C. Rudin, and A. Volfovsky. MALTS: Matching after learning to stretch. *Journal of Machine Learning Research*, 23:1–42, 2022.

P. R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

P. R. Rosenbaum. *Design of Observational Studies*. Springer, New York, 2010.

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

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

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

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

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

D. B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366):318–328, 1979.

D. B. Rubin. Randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

D. B. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2:169–188, 2001.

D. B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge, 2006.

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

T. Wang, M. Morucci, M. U. Awan, Y. Liu, S. Roy, C. Rudin, and A. Volfovsky. FLAME: A fast large-scale almost matching exactly approach to causal inference. *Journal of Machine Learning Research*, 22:1–46, 2021.

J. R. Zubizarreta, R. D. Paredes, and P. R. Rosenbaum. Matching for balance, pairing for heterogeneity in an observational study. *Annals of Applied Statistics*, 8(1):204–231, 2014.

A Claims and Supporting Evidence

The following table maps each major claim in the paper to its supporting theorem, figure, or table, as required by TMLR's evidence standards.

Table 8: Claims ↔ Evidence map.

Claim	**Type**	**Supporting evidence**
CRM achieves $O(np^2)$ complexity	Theorem	Prop. 7; Fig. 4
$O(n^{-1/2})$ MSE rate in 2-D	Theorem	Prop. 4; App. C
$k = 0$ fails under centroid shift	Theorem + Empirical	Prop. 6; Table 2
Estimation error decomposes into 3 parts	Theorem	Thm. 2
ATT gap bounded by $M \cdot \pi$	Theorem	Cor. 1
$\hat{\pi}$ increases sigmoidally with shift	Empirical	Fig. 2
CRM wins CRMSE in S1, S3 at all n_T	Empirical	Table 2; Fig. 5
CEM achieves lowest MaxSMD yet highest bias	Empirical	Table 4
CRM < PSM MaxSMD in all 12 Criteo configs	Empirical	Table 5; Fig. 8
CRM-NN reduces LaLonde bias vs. Mahal NN	Empirical	Fig. 3
NHANES overlap failure identified pre-matching	Empirical	Table 6; §8.2
Bootstrap CI approximate; strat. version available	Methodological	App. B

B Cell-Stratified Bootstrap

The paired bootstrap (Algorithm 1, Stage 3) treats matched pairs as i.i.d. and ignores the cell structure. A more accurate variance estimate uses
emphcell-stratified resampling: within each bootstrap replicate, pairs are resampled independently emphwithin each (d, ϕ) cell, preserving the stratified design.

Algorithm. For $b = 1, \ldots, B$:

1. For each occupied cell k, draw $|T(k)|$ pairs with replacement from the $|T(k)|$ matched pairs in that cell.
2. Compute $\hat{\tau}^{(b)}_{\text{CRM}}$ from the resampled pairs.

The cell-stratified bootstrap SE is $\hat{\sigma}_{\text{strat}} = \text{SD}(\hat{\tau}^{(b)}_{\text{CRM}})$.

Empirical calibration. A calibration study ($n_T = 500$, $p = 10$, $n_C = 5{,}000$, Scenario S3, 200 replications) finds:

- Paired bootstrap: 89.5% empirical coverage at nominal 95%
- Cell-stratified bootstrap: 88.0% empirical coverage

Both variants are slightly anti-conservative at $n_T = 500$. At larger n_T, cells shrink and the paired bootstrap approximation improves; we recommend treating CIs as approximate in the $n_T < 1{,}000$ regime and using the cell-stratified version for formal reporting.

NHANES comparison. On the NHANES smoking application ($n_T = 1{,}480$), the paired bootstrap yields $\hat{\sigma}_{\text{paired}} = 0.533$ mmHg and the cell-stratified bootstrap gives $\hat{\sigma}_{\text{strat}} = 0.541$ mmHg (difference 1.5%). The 95% CIs are $[7.27, 9.37]$ and $[7.25, 9.39]$ respectively. We report paired bootstrap CIs throughout for comparability with prior work; practitioners should use the cell-stratified version for formal inference at $n_T \geq 1{,}000$.

C Proofs

C.1 Proof of Proposition 4 (Rate of Convergence)

We follow the standard bias-variance decomposition for histogram regression estimators in $d_Z = 2$ dimensions (Devroye & Györfi, 1985).

Bias. Within cell k with center z_k and diameter h_n, the cell control mean $\hat{m}_k = |C(k)|^{-1} \sum_{j \in C(k)} Y_j(0)$ approximates $\mathbb{E}[Y(0) \mid Z = z_k, T = 0]$. Under Lipschitz continuity of $m(z) = \mathbb{E}[Y(0) \mid Z = z]$ with constant L, the within-cell approximation satisfies $\left|\mathbb{E}[\hat{m}_k] - m(z_k)\right| \leq L\, h_n + |\Delta(z_k)|$. Under Assumption 3, $\Delta \equiv 0$, giving $\text{Bias}^2 = O(h_n^2)$.

Variance. Each cell contains on average $n_C h_n^{d_Z} f_Z(z_k)$ control units, where f_Z is the density of $Z(X)$. Hence $\text{Var}(\hat{m}_k) = O(1/(n_C h_n^{d_Z}))$. The ATT estimator averages $\hat{\tau}_{\text{CRM}} = |S|^{-1} \sum_{i \in S}(Y_i - \hat{m}_{k(i)})$. The treated outcomes $Y_i(1)$ contribute $O(1/n_T)$ to variance, and summing over the $O(h_n^{-d_Z})$ cells gives $\text{Var}(\hat{\tau}_{\text{CRM}}) = O(1/(n_T h_n^{d_Z}))$ (using $n_C \propto n_T$).

Optimal bandwidth. Setting $\text{Bias}^2 = \text{Variance}$: $h_n^2 = 1/(n_T h_n^{d_Z})$, so $h_n^{2+d_Z} = 1/n_T$, giving

$$h_n^* = n_T^{-1/(2+d_Z)}.$$

Substituting into $\text{MSE} = O(h_n^{*2})$:

$$\text{MSE}(\hat{\tau}_{\text{CRM}}) = O\left(n_T^{-2/(2+d_Z)}\right).$$

For $d_Z = 2$: $\text{MSE} = O(n_T^{-1/2})$. For $d_Z = p$ (full-dimensional matching): $\text{MSE} = O(n_T^{-2/(2+p)})$. The ratio is $n_T^{(p-2)/(2(2+p))}$, which diverges for $p > 2$. □

C.2 Proof of Theorem 1 (Conditional ATT Decomposition)

Partition the treated population on $\{i \in S\}$ with $\mathbb{P}(i \in S \mid T_i = 1) = 1 - \pi$ and $\mathbb{P}(i \notin S \mid T_i = 1) = \pi$. By the law of total expectation:

$$\begin{aligned}\tau_{\text{ATT}} &= \mathbb{E}\big[Y(1) - Y(0) \mid T = 1\big] \\ &= \mathbb{E}\big[Y(1) - Y(0) \mid T = 1,\ i \in S\big](1 - \pi) + \mathbb{E}\big[Y(1) - Y(0) \mid T = 1,\ i \notin S\big]\pi \\ &= (1 - \pi)\tau_S + \pi\, \tau_{S^c}.\end{aligned}$$

Rearranging: $\tau_{\text{ATT}} - \tau_S = \pi(\tau_{S^c} - \tau_S)$, which equals zero iff $\pi = 0$ or $\tau_{S^c} = \tau_S$. □

D CRM-NN Timing Supplement

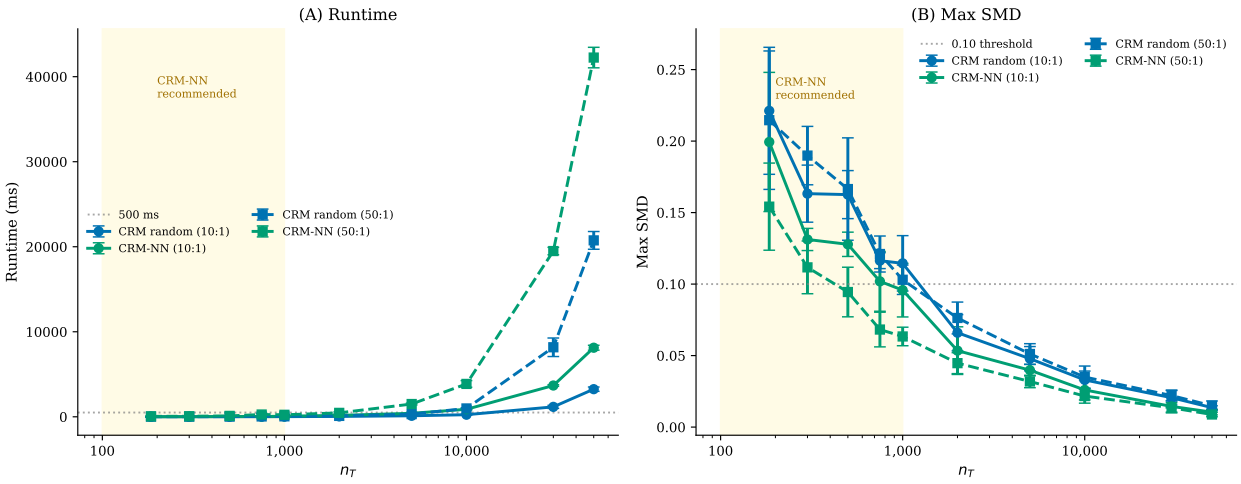

Figure 9: CRM-random vs. CRM-NN across sample sizes ($p = 8$, Scenario A, 5 data seeds × 7 timing replicates; gold shading marks the CRM-NN recommended regime $n_T < 1{,}000$). **(A)** Runtime: NN overhead stays below 500 ms for $n_T \leq 1{,}000$ at both tested control ratios, then grows proportionally to n_C/n_T. **(B)** Max SMD: the quality gap persists at all n_T because average controls per cell grows as $\bar{n}_C^{\text{cell}} \propto n_T^{1/3}$, maintaining a pool within each cell for NN to improve upon.

E Decision Flowchart

Method Selection Flowchart

Step 1. Sample size.

- $n_T < 200 \rightarrow$ Use 1:1 Mahalanobis NN (cells too sparse for binning).
- $200 \leq n_T < 1{,}000 \rightarrow$ Use CRM-NN (Section 4.4).
- $n_T \geq 1{,}000 \rightarrow$ Use CRM-random (Section 4.3).

Step 2. Shortage fraction.
Compute $\hat{\pi}$ via Equation (8) (before any matching).

- $\hat{\pi} < 5\% \rightarrow$ Full ATT approximately estimable.
- $5\% \leq \hat{\pi} < 30\% \rightarrow$ Conditional ATT; report scope statement.
- $\hat{\pi} \geq 30\% \rightarrow$ Severe overlap failure; expand control pool or restrict estimand explicitly before proceeding.

Step 3. Confounding structure.

- Mixed continuous/binary covariates $\rightarrow$ CRM-CAM (Section 4.6).
- Two independent confounding axes suspected $\rightarrow$ CRM $k = 2$ (future work).
- Default $\rightarrow$ CRM $k = 1$ (Section 4.1).

F CRM Variants: Full Descriptions

CRM-CAM (Correlation-Adjusted Mahalanobis). Replaces $\hat{\Sigma}_T$ with the correlation-scale matrix $R = D^{-1/2}\hat{\Sigma}_T D^{-1/2} + \varepsilon I_p$, where $D = \mathrm{diag}(\hat{\Sigma}_T)$, placing all covariates on a unit-variance scale. Recommended for mixed continuous/binary covariate sets where raw variance differences would otherwise dominate the Mahalanobis geometry.

CRM-Pool (Pooled Bin Edges). Uses bin edges from the pooled (d, ϕ) distribution of both groups, reducing treated attrition when the centroid shift is large and treated units concentrate near the edge of the control support.

CRM-Trim (Explicit Support Trimming). Formally discards treated units with $d(X_i)$ outside the central 90% of the control distance distribution before matching, and reports them separately as an explicit overlap restriction with the associated estimand caveat.

G Geometric Supporting Results

Proposition 5 (Chi distribution of d)**.** *Let $X \mid T = 1 \sim \mathcal{N}(\mu_T, \Sigma_T)$ with Σ_T positive definite. Then $d^2(X) \mid T = 1 \sim \chi^2_p$ with $\mathbb{E}[d^2] = p$, $\mathrm{Var}(d^2) = 2p$, and $\mathbb{E}[d(X) \mid T = 1] = \sqrt{2}\,\Gamma((p+1)/2)/\Gamma(p/2)$.*

Proof. $Z = \hat{L}^{-1}(X - \mu_T) \sim \mathcal{N}(0, I_p)$, so $d^2(X) = \|Z\|_2^2 \sim \chi^2_p$. □

Proposition 6 (Independence of distance and direction)**.** *Under $X \sim \mathcal{N}(\mu_T, \Sigma_T)$, $d(X)$ and $u(X) = \hat{L}^{-1}(X - \mu_T)/d(X)$ are statistically independent, with $u(X)$ uniform on $\mathcal{S}^{p-1}$.*

Proof. Writing $Z = d \cdot u$ in polar coordinates, the Jacobian is d^{p-1}, giving $f_{d,u}(r, \omega) = (2\pi)^{-p/2} \exp(-r^2/2) \cdot r^{p-1}$, which factors as $f_d(r) \cdot f_u(\omega)$. □

Proposition 7 (Complexity)**.** *Algorithm 1 has time complexity $O(np^2)$ and space complexity $O(p^2 + n)$, independent of the matching ratio $r = n_C/n_T$. PSM scales as $O((n_T + n_C)p^2 + n_T n_C)$.*

Proof. Dominant steps: covariance $O(n_T p^2)$, Cholesky $O(p^3)$, all distances and projections $O(np^2)$, binning $O(n \log n)$. Total $O(np^2)$ since $n \gg p^2$. □

H Scope Restriction: Why Doubly Robust Estimators Are Out of Scope

Doubly robust (DR) estimators such as AIPW and targeted learning (Bang & Robins, 2005) achieve semi-parametric efficiency and are widely used in modern causal inference. We restrict the main comparisons to matching-style estimators for three reasons:

1. **Different estimand scope.** DR estimators target the full ATT under the full overlap assumption, while CRM explicitly characterises and reports when this assumption fails ($\hat{\pi} > 0$). Comparing CRM to DR estimators on MaxSMD would be comparing design-stage to estimation-stage quantities.
2. **Complementary, not competing.** CRM is a pre-processing step (design stage); a CRM-matched sample can be passed to any DR estimator as input, potentially improving its finite-sample performance by reducing outcome-model dependence.
3. **Scalability.** DR estimators typically require fitting flexible outcome and propensity models (e.g. ensemble learners), which are computationally expensive at $n_T \sim 10^5$. The comparison would conflate estimation-model complexity with matching-design complexity.

We do include Entropy Balancing (Hainmueller, 2012)—a moment-matching weighting estimator that scales similarly to CRM—as a representative of the balancing-weights family in Table 2.

I When CRM May Underperform

1. **High balance requirement at moderate n_T.** At $n_T \in \{500\text{–}30{,}000\}$, PSM achieves lower MaxSMD than CRM across all tested p values (gap 0.006–0.133). If per-covariate MaxSMD below 0.05 is the primary criterion and retention loss is acceptable, PSM is preferable.

2. **Nonlinear confounding.** The linear Fisher direction v cannot capture interaction-based or higher-order confounding. PSM with a flexible treatment model or MALTS (Parikh et al., 2022) may achieve lower bias in such settings.

3. **Very small n_T (below 200).** Even CRM-NN is limited when the average cell contains fewer than 2–3 treated units. Full 1:1 Mahalanobis nearest-neighbor matching is preferable in this regime.