Leibniz Institute for
EAST AND SOUTHEAST
EUROPEAN STUDIES

Arbeitsbereich Ökonomie

IOS Working Papers

No. 410 August 2026

Balanced, Yet Still Biased: Correcting Hidden Pairwise Mismatch in Matching Estimators

Richard Bräuer*

*IOS Regensburg

Leibniz Institute for
EAST AND SOUTHEAST
EUROPEAN STUDIES

Landshuter Straße 4
93047 Regensburg

Telefon: (0941) 943 54-10
Telefax: (0941) 943 94-27
E-Mail: info@ios-regensburg.de
Internet: www.leibniz-ios.de

Balanced, Yet Still Biased: Correcting Hidden Pairwise Mismatch in Matching Estimators

Richard Bräuer*[1]

[1]IOS Regensburg

Abstract

Matching estimators exhibit a previously overlooked source of finite sample bias. Even when the assumptions underlying matching hold, the local geometry of the covariate distribution around each observation makes some matching directions more likely than others. This mismatch is ubiquitous and disappears only slowly as the sample size increases. Estimators suffer increased variance and, in most cases where the outcome function is nonlinear, systematic bias. Conventional balance diagnostics generally fail to detect the problem. When they identify it, matching is abandoned unnecessarily. I propose three approaches to mitigate this problem and evaluate them in simulations and empirical applications. A matching procedure that deliberately offsets mismatch across neighboring observations reduces both bias and variance while remaining conceptually and computationally simple.

Keywords: econometrics, matching

Classification: C21

*Corresponding author: Richard Bräuer (IOS), Landshuter Str. 4, 93047 Regensburg, braeuer@ios-regensburg.de. I especially thank Eric Bartelsman (VU Amsterdam) and Steffen Müller (IWH) for their guidance and support. The paper benefited greatly from the comments of Martin Huber, Peter Egger, Tobias Korn, Matthias Mertens and the participants of the IWH DJF and the Fribourg Economics Seminar. I thank Guido Imbens and Yiqing Xu for their provision of data and Luca Li Calzi for his excellent research assistance. [Newest Version]

1 Introduction

Matching and techniques derived from it form an important part of the econometrician's toolbox (King and Nielsen, 2019). Most recently, matching has gained importance as a preprocessing step for a difference-in-differences (DiD) estimation (see Ham and Miratrix, 2024 for an overview). I show that the most common matching procedures produce systematic mismatch that is not detectable by currently employed tests for covariate balance. This mismatch potentially increases both bias and variance of the estimates in finite samples.

The principal contribution of the paper is to demonstrate this problem, show three different strategies to mitigate it, and test the associated new estimators in simulations and real applications where the true treatment effect is known.

The intuition behind the detected bias is that some regions in the covariate space have higher density of observations, so it is more likely that matches will be found there. Any matching estimator will thus gravitate to such regions when matching the surrounding observations. I use simulation exercises to characterize the resulting error: This drift is substantial and vanishes so slowly when increasing the sample size that all practical applications should be considered finite sample estimates for the purposes of this bias. It is often not detected by conventional balance tests, because deviations have opposite signs on the opposite 'sides' of such a higher density region. This behavior leads to biased estimates when the potential outcomes Y_1 and Y_0 are nonlinear functions of $\overrightarrow{X}$ around this denser region of the covariate space, especially if there are interaction effects between variables. I suggest the relationship between univariate mismatch $x_i - x_j$ and x_i as a diagnostic tool and shows that systematic mismatch in this category is widespread.

I develop three different strategies to mitigate this bias: First, one can observe the drift and then alter the distance metric used in matching to correct this tendency. This is the most direct intervention, but computationally expensive. Second, it is possible to observe mismatch and correct in the other direction when matching 'the next closest' observation. This error ac-

cumulation technique breaks the usual independence between the matching of individual observations. The new estimator thus consists of both an order in which observations are matched and a distance metric. It performs best in practice. Third, correctly estimated propensity scores yield a formula for the probability with which the bias inherent in each matching pair enters the estimation. It is possible to ex post reweight pairs and cancel out arbitrary bias, not just from this mechanism. However, this is very sensitive to the propensity score estimate.[1]

Apart from simulations, I apply these estimators to real world applications to gauge their performance in practice. For instance, I apply both the new and conventional estimators to two standard matching applications with public data and treatment randomization. In both applications, conventional estimators have large standard errors, which are substantially reduced after correcting for systemic mismatch. In both cases, the new estimators recover the experimental benchmark faithfully. Mahalanobis matching with error accumulation performs best.

This ties into the large discussion on the small sample properties of matching estimators. Since matching estimators are consistent but potentially biased, finite sample properties of matching estimators are widely studied. More complex techniques have been suggested to improve performance (e.g. Abadie and Imbens (2006, 2011, 2016)). Yet, in practice, simple nearest neighbor matching dominates applied work (King and Nielsen, 2019). Matching with error accumulation combines simplicity with significant performance improvements. King and Nielsen (2019) is particularly close to my paper in spirit, as both papers identify a form of finite-sample mismatch that can be obscured by conventional aggregate diagnostics. The mechanisms, however, are distinct: Their critique concerns propensity-score matching, whereas the problem considered here arises from systematic pairwise mismatch and therefore extends to matching procedures more generally.

Difference-in-differences has become one of the most widely used tools for causal inference in applied economics, accompanied by a rapidly expanding methodological literature addressing treatment timing, violations of parallel

[1] I provide a software package to implement all three estimators.

trends, and inference (Roth et al., 2023). Matching is increasingly combined with difference-in-differences to construct treatment and comparison groups that are more comparable in their observable characteristics and pre-treatment outcomes. This practice has become sufficiently widespread to generate a growing literature specifically concerned with the properties and implementation of matching-based DiD designs (Smith and Todd, 2005; Lee et al., 2025; Lee and Wooldridge, 2026). More recent methodological contributions continue to develop new ways of combining matching or weighting with DiD, including approaches addressing limited overlap and the properties of post-matching estimation (Kim and Lee, 2025). The increasing use of these hybrid designs makes the properties of the underlying matching procedure consequential for a large and growing body of empirical research. The results above identify a previously overlooked source of bias: Even when matching achieves good aggregate balance at the time of treatment, the systematic matching errors at the pair level can give rise to different post-treatment paths and thereby contaminate the difference-in-differences comparison. The application in Section 4.3 illustrates the practical relevance of this problem in a placebo DiD setting: I generate a placebo program for German inventors during the reunification. I document the existence of systematic mismatch using standard estimators and find several spurious effects as well as unbalanced matching in the event study, which are substantially reduced when using the new error accumulation technique. This is related to the extremely skewed distribution of many patenting related variables.

The paper also relates to the literature on ex post evaluation of matching. The systematic mismatch is basically mechanical and as such widely present in applied work. If balance tests detect it, it will occasionally prohibit use of matching, but since the bias is on the pair level and often cancels itself out in aggregate, current balance tests are poorly equipped to deal with it. This is especially true for the popular mean comparisons. However, systematic pairwise mismatch usually exists even if the overall distributions are identical. Some techniques adjacent to matching use these distributional comparisons as an objective function directly (e.g. entropy balancing (Hainmueller, 2012) and covariate balancing propensity score (Imai and Ratkovic, 2014)). They

are in principle vulnerable to a similar critique: Distributional balancing is not in itself enough to guarantee unbiased comparisons if the outcome function is non-linear.

This paper is organized as follows: In section 2, I set up a framework to discuss the causes and size of systemic mismatch due to local asymmetries in the joint distribution of $\overrightarrow{X}$, show that systemic mismatch can be a relevant problem for all practical sample sizes and discuss three estimation strategies to alleviate this problem. Section 3 performs a simulation study to show the performance of the new estimators. Section 4 applies these results to two applications where experimental benchmarks exist to gauge the performance of the new estimators. Section 5 concludes.

2 Theoretical Framework and Notation

2.1 Systemic observation level mismatch in matching

In matching, each observation i is typically matched to a 'similar' comparison observation j (or a group of them). Each matching method constructs j somewhat differently. However, since simple nearest neighbor matching is still the most popular technique (e.g. only 6% of papers in the sample by King and Nielsen (2019) use any more complicated procedure) and conceptually simple, the following discussion will assume it as the baseline. However, the discussion carries over to more complicated estimators.[2] For ease of exposition, I will discuss the case of matching a treated to a control observation. The discussion is exactly mirrored in the reverse case.

Let $\mathcal{M}$ denote the set of realized matched pairs (i, j), where exactly one observation in each pair is treated, and let j denote the control observation matched to treated unit i. Following Abadie and Imbens (2006, 2011), the observed outcome difference within a pair can be decomposed as

$$Y_i - Y_j = \underbrace{Y_i(1) - Y_i(0)}_{\tau_i} + \underbrace{Y_i(0) - Y_j(0)}_{B_{ij}}$$

[2]Figure 1 shows that using more complex estimators does not solve the core problem.

where τ_i denotes the individual treatment effect and B_{ij} captures the mismatch in untreated potential outcomes induced by imperfect covariate matching. $Y_i(0)$ is not observed for treated units, but $Y_j(0)$ serves as a proxy. By assumption, B_{ij} is some function of the covariates of i and j. Abadie and Imbens (2011) propose to estimate an outcome function for untreated observations $m_0(X) = E[Y(0) \mid X]$ and use this to correct the mismatch component of each pair. In this paper, I suggest a different approach: I characterize the expected difference for observation i and adjust the matching routine itself:

Let $\delta_{i,j}(X_i, X_j)$ denote the difference metric used in matching and let $D_i = \delta(X_i, X_j)$ denote the random variable describing the realized matching distance for i induced by the matching procedure. Then, the ex ante expectation for the treatment effect of observation i can be decomposed as follows using the law of iterated expectations:

$$
\begin{aligned}
E(Y_i - Y_j) &= \tau_i + E(B_i) = \tau_i + E[E(B_i \mid D_i)] \\
&= \tau_i + \int_0^\infty \underbrace{E(B_i \mid D_i = \delta)}_{\text{exp. bias at } \delta} \underbrace{N_c \, f_i(\delta)}_{\text{density at } \delta} \underbrace{\left(1 - \int_0^\delta f_i(s) \, \mathrm{d}s\right)^{N_C - 1}}_{\text{p(no closer control)}} \mathrm{d}\delta \\
&= \tau_i + \int_0^\infty E(B_i \mid D_i = \delta) g_i(\delta) \, \mathrm{d}\delta
\end{aligned}
\tag{1}
$$

where the second and third line describe the contribution of every matching distance d to the overall expected bias. It weights the conditional bias if a match is chosen at distance δ with the probability of that event $g(\delta)$. The consistency of the matching estimator can easily be seen from this notation: As $N_c \to \infty$, $g(\delta)$ concentrates its mass at $\delta = 0$ under standard regularity conditions. Trivially, the estimator is unbiased in that case. Note that this holds for every single observation i. However, it is also clear that the matching estimator is not unbiased for positive δ.

The conditional bias at distance δ is itself an expectation over all potential comparison observations located on the hypersphere at distance δ. Hence,

$E(B_i \mid D_i = \delta)$ in eq. 1 is

$$E(B_i \mid D_i = \delta) = \int_{\mathcal{D}_i(\delta)} B(X_i, X_j) f(X_j \mid X_i, D_i = \delta) \, \mathrm{d}X_j \tag{2}$$

which will not generally be 0: First, the density of potential matches $f(X_j \mid X_i, D_i = \delta)$ will not generally be equal at all points on the hypersphere, which means that control observations will in expectation have different covariates than treated observations. Second, even if the density were equal, the bias does not need to be: If the bias is a non-linear function of the covariates, the different points along the hypersphere cannot counterbalance each other and an error will remain. It is the argument of this paper that the first is almost mechanically the case in econometric applications, and thus the small sample properties of matching estimation hinge much more on the linearity of the outcome function than previously understood. The simulations in section 3 and the analysis of select applications will show the ubiquity of this phenomenon.

Equation (1) does not rely on any assumptions regarding the functional form of the outcome process. To connect this framework to the bias correction proposed by Abadie and Imbens (2011), suppose that the untreated outcome function $m_0(X) = E[Y(0) \mid X]$ is continuously differentiable. A first-order Taylor expansion around X_i yields

$$m_0(X_j) = m_0(X_i) + \nabla m_0(X_i)'(X_j - X_i) + R(X_i, X_j),$$

where $R(X_i, X_j)$ contains the higher-order remainder terms. Then, substituting this expression into the conditional expectation gives

$$\begin{aligned} E(B_i \mid D_i = \delta) &= E(m_0(X_i) - m_0(X_j)) \\ &= -\nabla m_0(X_i)' \int_{\mathcal{D}_i(\delta)} (X_j - X_i) f(X_j \mid X_i, D_i = \delta) \, \mathrm{d}X_j \\ &\quad - \int_{\mathcal{D}_i(\delta)} R(X_i, X_j) f(X_j \mid X_i, D_i = \delta) \, \mathrm{d}X_j. \end{aligned} \tag{3}$$

This decomposition separates two conceptually distinct sources of mismatch bias. The first term captures the linear component and depends entirely on the expected displacement $E(X_j - X_i|D_i = \delta) = 0$ of the matched observation relative to the treated observation. The second term reflects nonlinearities in the outcome function through the Taylor remainder. The bias correction of Abadie and Imbens (2011) can be interpreted as estimating the first component by approximating the local gradient of $m_0(X)$. However, the linear component vanishes whenever $E(X_j - X_i|D_i = \delta) = 0$ without the need to estimate any specific outcome function. In addition, reducing the expected displacement $E(X_j - X_i|D_i = \delta) = 0$ also reduces the higher order terms, whereas the first-order correction of Abadie and Imbens (2011) leaves these terms unchanged. Thus, the analysis in the following instead focuses on the geometric properties of the matching distribution that determine the expected displacement itself. The following example will illustrate both the issue and the approach of this paper.

2.2 Illustrative Example

I use a stylized univariate matching problem to build intuition of the for the problem. Examples from simulation studies and real world data are presented in section 2.3 and 4, respectively. Consider a matching exercise with just one variable x_1. $x_1 \sim \mathrm{N}(0, 1)$ in both the treatment and the control group, so there is common support. Nevertheless, any observation with $x_i < 0$ will in expectation be matched to a larger comparison group: There is more probability mass to the right of $X_i < 0$ than to the left, so the chance of matching with an observation with larger x_1 is higher. This is not a special feature of the normal distribution, any unimodal distribution will produce the same logic. Multimodal distributions will create more complex patterns of mismatch, but are also not immune.

To gauge the direction and strength of expected mismatch, define the mass imbalance measure $\eta(x_i, k, \mathbf{D})$, which compares the local probability mass of potential matches immediately to the right and left of observation i along covariate dimension k:

$$\eta(x_i, k, \mathbf{D}) = \frac{M_{ik}^{+}(\mathbf{D}) - M_{ik}^{-}(\mathbf{D})}{M_{ik}^{+}(\mathbf{D}) + M_{ik}^{-}(\mathbf{D})} \tag{4}$$

where

$$M_{ik}^{+}(\mathbf{D}) = \int_{x_i^k}^{x_i^k + d_k} d\big(x, \mathbf{x}_i^{-k}\big)\ dx$$

and

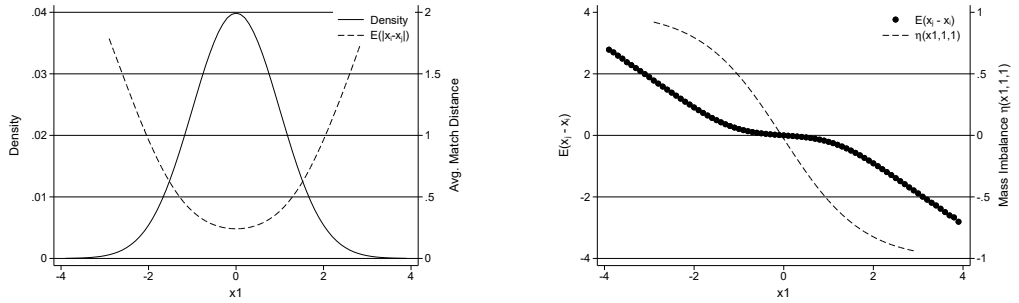
$$M_{ik}^{-}(\mathbf{D}) = \int_{x_i^k - d_k}^{x_i^k} d\big(x, \mathbf{x}_i^{-k}\big)\ dx$$

Here, $d(\mathbf{x})$ denotes the joint density of potential matches over the covariate space. The notation $d(x, \mathbf{x}_i^{-k})$ indicates that the k^{th} covariate varies over the interval of integration while all remaining covariates are fixed at the values of observation i. The term $M_{ik}^{+}(\mathbf{D})$ therefore captures the local density slice immediately to the right of x_i^k, while $M_{ik}^{-}(\mathbf{D})$ captures the corresponding slice to the left. The numerator measures the directional imbalance in local matching opportunities, while the denominator normalizes by total local mass, implying $\eta(x_i, k, \mathbf{D}) \in [-1, 1]$. Positive values indicate that there are more potential matches with higher values, negative values indicate a tenedency to match with smaller x. Note that this is not a complete statistic to recover the direction of expected mismatch, which depends on the joint distribution of all covariates and on the realized matching distance (see section 2.1).

Figure 1 presents the concrete example in visual form. Panel 1a presents the standard normal distribution and the expected match distance (when drawing 4 potential matches from the distribution). Panel 1b reports the mass probability ratio $\eta(x_i, 1, 1)$ for the only variable, using $x_i \mp 1$ and the difference between x_i and x_j. There is systematic mismatch between x_i and x_j: Small x_i get matched to larger x_j and vice versa. The direction of this mismatch is predicted by $\eta(x_i, 1, 1)$. To actually fit the distribution of mismatch, I would need to use $\eta(x_i, 1, b)$ with different calipers b, to account for the different density of matches in different parts of the distribution (see

Figure 1: *Example: Mismatch with an univariate normal distribution*

(a) Normal distribution (b) Expected Mismatch



Notes: Expected mismatch with one standard normal distributed variable ($x1$). Panel 1a shows the normal distribution and the average distance between x_i and its match x_j. Panel 1b shows the probability mass imbalance $\eta(x_i, k, \mathbf{D}) = \frac{M^+_{ik}(\mathbf{D}) - M^-_{ik}(\mathbf{D})}{M^+_{ik}(\mathbf{D}) + M^-_{ik}(\mathbf{D})}$ at point x_i, i.e. the ratio of probability mass to the left and right of the observation. It also shows the actual mismatch after matching, which results from the interplay of $\eta(x_i, k, \mathbf{D})$ and the avg. distance to find a match. Note that there is no violation of common support and that expected mismatch is zero over the entire population, so common balance tests will not flag this issue. *Source*: own simulations.

Panel 1a). However, while in this stylized example, there is a closed form solution of local expected mismatch, I do not consider it to be estimatable in applied cases, where there are multiple variables with potential interactions, sparse observations and unknown distributions.

It is important to note that common covariate balance tests will not flag the above problem: Expected mismatch is systematically negative for large values and systematically positive for small values, but zero in aggregate. Because the resulting aggregate distributions can look highly similar despite this underlying pair-level drift, even more evolved balance tests that examine the whole distribution often fail to detect the mismatch and consider the two samples strictly balanced.

Since the mismatch arises organically from most common distributions, I maintain that the above problem is present in most applied work. Figures 6 and 7 report the mismatch for the two papers reproduced for this study.

Substantial expected mismatch is found in both of them when looking below aggregate properties.

2.3 Simulation Exercise

To understand the potential effect of this mismatch on estimated treatment effects, I perform a simulation exercise: Consider the following simple data generating process (DGP): There are four explanatory variables ($x1$ through $x4$), all drawn independently from the same shifted standard normal distribution (to avoid issues due to sign reversals). They and only they affect probability of treatment and outcome simultaneously. The probability of treatment is a latent variable probit. Apart from the treatment and $x1$ through $x4$, $x1$ and $x2$ also affect the outcome in a non-linear way, but the conditional independen assumption (CIA) is satisfied on $x1$ through $x4$. To simplify, the effects of all x-variables on probability of treatment and outcome are the same. Eq. 5 describes the DGP mathematically.

The following equations describe the DGP:

$$x_j \sim \mathcal{N}(5,1), \qquad j = 1,\ldots,5. \tag{5}$$

$$latent_D = x1 + x2 + x3 + x4 - 20 + N(0,1) \tag{6}$$

$$D = \mathbb{I}(latent_D > 0) \tag{7}$$

$$x5 = (x1)^2 \tag{8}$$

$$x6 = (x1) * (x2) \tag{9}$$

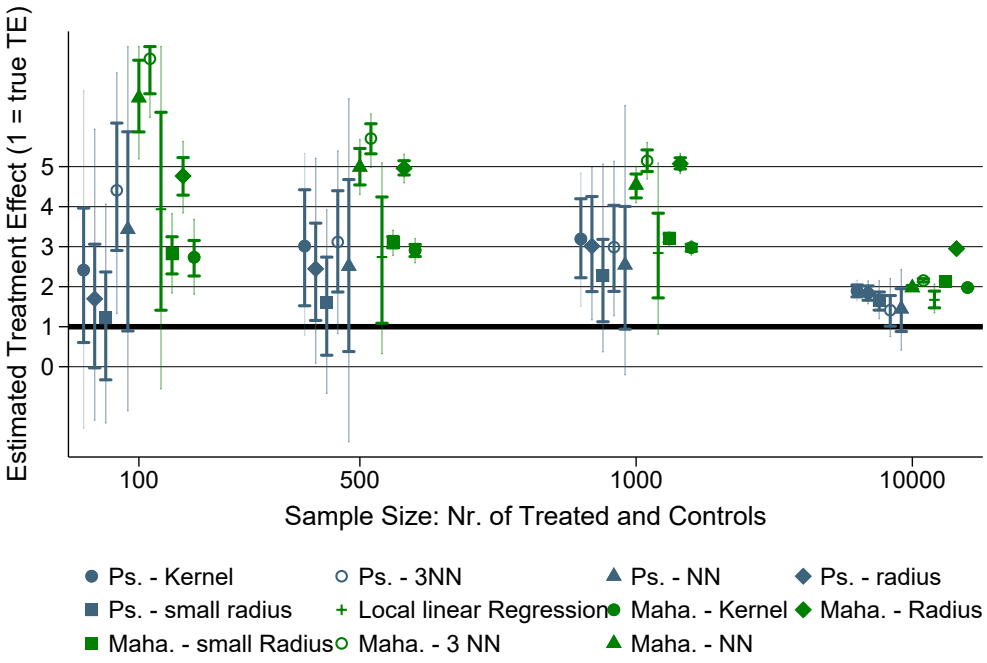
$$Y = 250 * x1 + 250 * x2 + 250 * x3 + 250 * x4 + \tag{10}$$

$$500 * x5 + 500 * x6 + 500 * D + U(0,250) \tag{11}$$

In this setup, $x1$ through $x4$ affect both treatment and probability to treat. They form $\overrightarrow{X}$ for subsequent matching. The researcher does not observe $x5$ and $x6$. For simplicity, they can be predicted perfectly by $\overrightarrow{X}$. However, the researcher cannot use this fact. In any case, conditional independence is satisfied on $\overrightarrow{X}$ alone.

Most of the already existing estimators are strongly biased for all sample sizes. Strict radius and kernel techniques have the smallest biases, but

Figure 2: *Simulation: Performance of Conventional Matching estimators*



Notes: Performance of conventional estimators with the DGP described in section 2.1. Y-axis is rescaled by the true treatment effect (500). While the CIA is fulfilled, all estimators systematically mismatch in finite samples and thus are biased. The bias cannot be eliminated with feasible sample sizes. Radius and kernel matching discard a large (and changing) part of the sample, which leads to the inverted U pattern.
Source: own simulations.

only because they discard a large part of the sample: Mahalanobis matching with a strict caliper (1) discards 60% of the 100 observations sample, 36% of the 500 observations sample, 29% of the 1000 observations sample and 14% of the 10.000 observation sample. At the smallest sample size (100), mostly observations around the mode are actually matched, where the density of observations is highest. Since expected mismatch is zero at the mode, this actually improves the estimate and leads to the paradoxical outcome that bigger sample size leads to worse matches when calipers are very strict. Additionally, discarding observations is not costly, since bias concerns dominate variance and all units have the same treatment effect. However, at this magnitude, it would not be a credible strategy in actual applications.

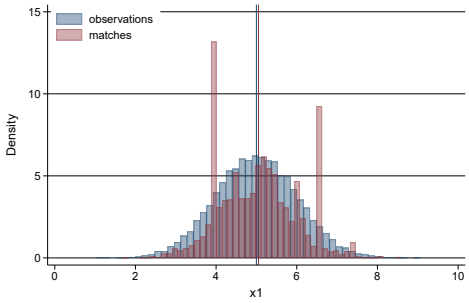
While conventional balancing tests fail to detect this problem, it is visible in the data without using the non-observable hypothetical outcomes: Figure 3 reports diagnostics for propensity score matching (as the most popular estimator) from one of the simulations with 10000 potential matches. Panel 3a reports the entire distribution of $x1$ (note that in the DGP, all x-variables have the same distribution) and the systematic mismatch at the pair level. Clearly, the aggregate distributions and especially means look extremely similar and would likely be acceptable for publication. However, this hides substantial systematic mismatch at the pair level that gives rise to the observed bias.

Some features of the DGP can lead to a strong bias: If the outcome function is linear, the expected mismatch in some areas of the data can be counteracted by other areas with opposite mismatch. In the baseline DGP (eq. 5), the ATE estimate would be virtually unbiased if the outcome function was linear, since the means of the covariate distributions are so close.

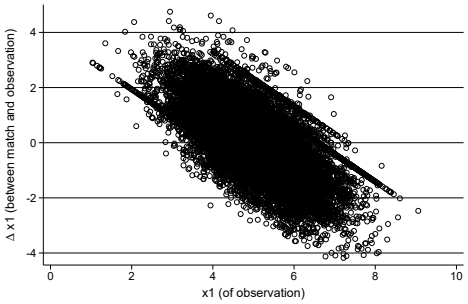
However, this is no longer the case for the ATT: Since the treatment probability increases with all variables, neither the treated nor the untreated observations retain the symmetrical normal distributions of the entire dataset. Thus, there is perceptible 'drift' for matches of the treated and untreated subsamples (which cancel each other out for the ATE). Panels 3c and 3d report the distributions of observations and matches for one ATT estimate and the results from all simulations, respectively. Non-symmetrical distri-

Figure 3: *Diagnosis of Conventional Matching estimators*

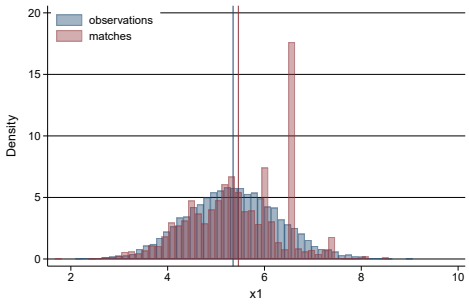
(a) Distribution of x1



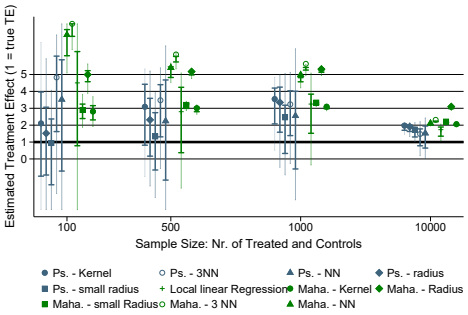
(b) Mismatch (in $x1$; 10000 Obs.)



(c) Distribution of x1 ($D = 1$)



(d) ATT estimates



Notes: Diagnosis of Nearest Neighbor Propensity Score Matching as an example out of the estimators presented in Figure 2. Note that it is by no means the worst performing estimator and that the assumptions underlying propensity score estimation are fulfilled. Panel 3a and 3b report the aggregate distributions and systematic mismatch. Panel 3c again reports aggregate statistics, but only for treated observations. Panel 3d reports the associated ATT estimates. Y-axis is rescaled by the true treatment effect (500).
Source: own simulations.

butions of variables can create aggregate mismatch which will affect results even if the outcome function is linear. Yet, this aggregate mismatch will show up in balance tests, so such cases will lead to unpublished results, not the publication of biased estimates. Note that while the difference might appear small, this is from a sample of 10000 treated and 10000 control observations with only 4 variables. Again, for practically relevant sample sizes, the bias remains substantial.

Panels 4a and 4b report results from a DGP where all $x \sim \mathcal{U}(0, 10)$. Note that bias is much worse: While probability mass is identical for all values, many observations are close to at least one of the boundary of the support, where mismatch is severe. Generally, univariate distributions are not enough to diagnose the issue: Expected bias is a function of the density imbalances along the hypersphere defined by a fixed distance to observation i.

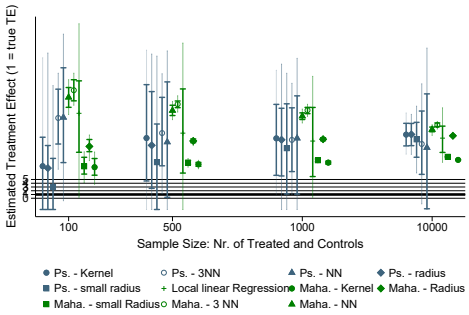
Panels 4c and 4d result from estimation in a DGP without the interaction term in eq. 5. While mismatch is identical to the baseline scenario, bias is substantially lower, since the outcome function is more linear than the baseline. This highlights that while mismatch is diagnosable and the mechanical cause makes it plausible that it exists in a wide variety of applications, the resulting bias is a function of the unknown outcome function. The bias in any application will be a complex interplay between the mechanisms at play: Average matching distance, the non-linearity of the outcome and the shape of the underlying joint distribution of variables all contribute to bias. The following section will discuss three estimation strategies to alleviate this bias and lead to satisfactory performance in finite samples even in the face of strong non-linearity.

2.4 Proposed Estimators

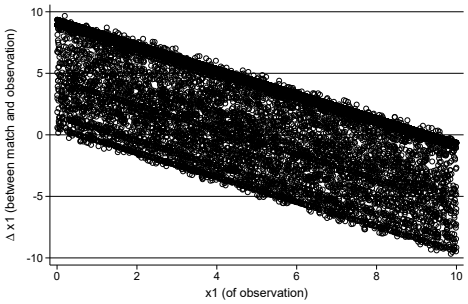
There are three potential avenues to alleviate the systematic mismatch discussed in section 2: First, change the matching criterium to eliminate the bias. Second, remember the mismatch of previous observations and correct when matching the next observations. Third, reweight matches ex post to reduce bias in the treatment effect. I will follow all three ideas in this section.

Figure 4: *Bias with Different DGPs*

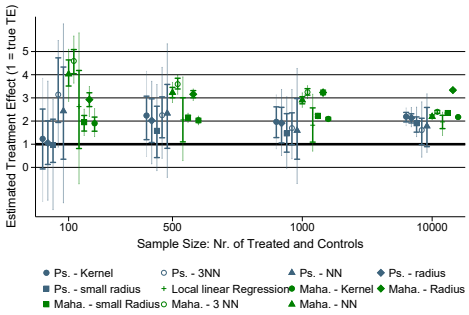
(a) Estimation Results (uniform)



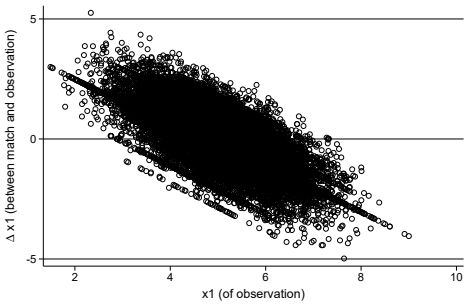
(b) Mismatch (uniform)



(c) Estimation Results (no interaction)



(d) Mismatch (no interaction)



Notes: Result and exemplary mismatch diagnostic for adjusted version of the baseline DGP in Figure 2. Panels 4a and 4b report from a DGP where all x are $\sim \mathcal{U}(0, 10)$. Note that bias is much worse: While probability mass is identical for all values, many observations are close to at least one of the border of the distribution, where mismatch is severe. Panels 4c and 4d refer to a DGP without the interaction term in eq. 5. While mismatch in x-variables is identical to the baseline, bias is much less severe than in the baseline, since the DGP is more linear. *Source*: own simulations.

2.4.1 Ex Ante Reweighted Mahalanobis Matching

The most direct way of addressing the systematic mismatch discussed in Section 2.1 is to modify the matching criterion itself to ensure that positive and negative deviations occur locally and therefore approximately offset each other. In particular, I seek a matching rule that satisfies

$$E(X_i - X_j | \overrightarrow{X} \approx \overrightarrow{X}_i) = 0$$

for every covariate used in matching.

Note that this condition does not imply $E(B_{ij}) = 0$, since the potential outcomes generally depend nonlinearly on the covariates. Nevertheless, balancing the covariate differences themselves ensures that positive and negative deviations occur in the same neighborhood of the covariate space (and enforce that they have mean zero overall, too).

To achieve this, I introduce an asymmetric reweighting of the Mahalanobis distance. For a given observation i, define a vector of variable weights

$$\beta_i = (\beta_{i,1}, \ldots, \beta_{i,K}),$$

a directional variation of which ($\tilde{\beta}_{i,j} = (\tilde{\beta}_{i,j,k}, \ldots, \tilde{\beta}_{i,j,K})$) is used for actual distance calculation. Specifically,

$$\tilde{\beta}_{i,j,k} = \begin{cases} \beta_{i,k} & X_{i,k} < X_{j,k}, \\ 1/\beta_{i,k} & X_{i,k} \geq X_{j,k}. \end{cases}$$

The reweighted Mahalanobis distance is then

$$d_{ij}^{RM} = ((X_i - X_j) \odot \tilde{\beta}_{i,j})' Z ((X_i - X_j) \odot \tilde{\beta}_{i,j})$$

where $\odot$ denotes elementwise multiplication and Z is the inverse covariance matrix used in the standard Mahalanobis metric.

Unlike the conventional Mahalanobis distance, this metric is generally asymmetric ($d_{ij}^{RM} \neq d_{ji}^{RM}$), since i and j will have different weight vectors β_i and β_j and since the difference between variable values will have opposite signs. The asymmetry is intentional. By assigning different penalties to

positive and negative deviations in each covariate, the matching criterion can be adjusted to compensate for systematic mismatch. For a given parameter vector $\beta_i = (\beta_1, \ldots, \beta_K)$, I evaluate the local imbalance around observation i using a kernel-weighted average of covariate differences

$$\Delta_i(\beta) = \frac{\sum_j K_h(d_{ij}^{RM})(X_i - X_j)}{\sum_j K_h(d_{ij}^{RM})},$$

where $K_h(\cdot)$ denotes a Gaussian kernel with bandwidth h. Throughout the simulations, h is chosen as the distance to the second nearest neighbor under the current metric. This ensures that it is as local as possible, while still always containing more than one observation and thus being a continuous (and differentiable) object for optimization. Each β_i^k is then chosen to minimize the local imbalance so that $\Delta_i^k = 0$.

Intuitively, the procedure searches for the directional weighting of covariates that makes the expected local mismatch around observation i as small as possible. After optimization, nearest-neighbour matching is performed using the resulting distance metric.

The procedure is computationally demanding because the optimization is performed separately for each observation. Evaluating a candidate parameter vector requires computing distances to the entire sample and repeatedly estimating the local imbalance criterion. Consequently, computation times are substantially larger than for conventional matching estimators, particularly in large samples. For this reason, simulation results are not reported for the 10000 candidate observations case, where repeated computation is too costly.

The method also relies on a smooth objective function, which this implementation of the kernel aims to provide. Other techniques for estimating the local average of covariates (local linear regressions, nearest neighbor matching, etc.) are in principle possible, but would add another optimization step and further increase computational costs.

In addition, it is unclear how to treat noncontinuous variables in this framework or sparce observations, both of which will lead to non-smooth objective functions and a failure of the routine. In the presented implemen-

tation, such observations are dropped (using the fact that optimization is at the observation level). Observations where the necessary weights rise above 10 or below 0.1 are also flagged as indicative of an inability to find satisfactory matches.

While this method has the intuitive advantage of directly attacking the root of the problem, these practical issues make it difficult to use in actual applications.

2.4.2 Mahalanobis Matching with Locally Accumulating Errors

The previous estimator attempts to eliminate mismatch locally by modifying the distance metric itself. A computationally simpler alternative is to allow mismatch to occur, but to accumulate and offset it across neighboring observations. The objective remains to achieve

$$E(X_i - X_j \mid X \approx X_i) = 0,$$

but without requiring the matching criterion to be optimized separately for every observation.

Let h denote an observation with $X_h \approx X_i$ and $D_h = D_i$. Under local continuity, observations with similar covariates also have similar expected mismatch. Thus, local mismatch can be approximated by averaging mismatch over neighboring observations and local unbiasedness does not require every individual mismatch to be zero. Instead, it is sufficient that mismatch cancels across neighboring observations, i.e.

$$\frac{1}{2}E(X_i - X_{j(i)}) + \frac{1}{2}E(X_h - X_{j(h)}) \approx E(X_i - X_j \mid X \approx X_i) \approx 0.$$

The proposed estimator therefore does not attempt to eliminate the mismatch generated by each individual match. Instead, it carries previously accumulated mismatch forward, encouraging subsequent matches of nearby observations to offset earlier errors. Unlike conventional Mahalanobis matching, the estimator is thus inherently sequential: If $s(1), \ldots, s(N)$ denote the

order in which observations are processed, the accumulated mismatch recursively for each observation is

$$a_{s(1)} = 0$$
$$a_{s(t)} = a_{s(t-1)} + (\overrightarrow{X}_{j(t-1)} - \overrightarrow{X}_{t-1})$$

The accumulated mismatch therefore records the total imbalance generated by all previously matched observations per variable.

For a given observation i, the modified Mahalanobis distance between i and a potential match j is then defined as

$$d_{ij}^{accM} = (\overrightarrow{X}_i - a_i - \overrightarrow{X}_j)' Z (\overrightarrow{X}_i - a_i - \overrightarrow{X}_j),$$

where Z is the inverse covariance matrix used in the standard Mahalanobis metric. Intuitively, if previous matches have generated positive imbalance in a particular covariate, subsequent observations are matched as if their covariate values were lower in that dimension. This increases the likelihood that future matches offset earlier errors. Rather than requiring every individual match to be unbiased, the estimator aims to ensure that mismatch cancels locally within neighborhoods of the covariate space.

Although a_i is defined as a cumulative sum, it is not expected to grow without bound. As accumulated imbalance becomes large in a particular direction, the modified distance metric increasingly favors matches that offset this imbalance. The recursion therefore creates a negative feedback mechanism rather than a random walk.

Since the accumulated mismatch $a_s(t)$ depends on all previous matching decisions, the resulting matching assignments depend on the chosen path through the data. I use the simplest possible implementation in this paper: I draw the first treated observation randomly and then choose the closest unmatched treated observation according to standard Mahalanobis distance to proceed. This procedure is repeated until all treated observations have been processed. The same algorithm is then applied to the untreated observations when estimating the ATE. Consequently, I compare each observation

twice: once to observations with the opposite treatment status to determine its match and once to observations with the same treatment status to determine the next observation to process. This ensures that accumulated mismatch is passed primarily to nearby observations. In this paper, I implement the estimator with replacement of potential matches, but there is no inherent issue with choosing to not replace matches instead.

The estimator is path-dependent: Different starting observations will produce different matching assignments. Additionally, alternative traversal schemes are possible. For example, one could pick start observations according to some outlier criterion (since extreme variable values have more mismatch), start from multiple observations simultaneously, restrict the propagation of accumulated mismatch to observations within a specified caliper, or employ more sophisticated clustering procedures. However, such modifications introduce additional tuning parameters and user discretion, making the estimator more difficult to implement and justify empirically. They also do not seem to be necessary: The simple traversal scheme performs very well in both the simulated data and the actual applications discussed below, while requiring no additional design choices.

The estimator performs well if treated and untreated observations are close to others of their kind, but far apart from each other. This is fairly common in empirical applications. In such cases, previous mismatch will be offset using close observations and local expected mismatch will be close to 0. If this is not the case, it might be beneficial to institute calipers beyond which mismatch does not propagate, to avoid cases where mismatch in one region contaminates the matching process in others. When diagnosing the suitability of the estimator as is, the decay of cumulative mismatch over the traversed distance is a natural diagnostic.

2.4.3 Inverse Match Direction Probability Reweighting

Assuming that the propensity score is estimated precisely, it is possible to remove mismatch bias through inverse probability reweighting: The key observation is that, conditional on the propensity scores, treatment assignment

is random. Before treatment is realized, a given pair of observations could therefore appear in either orientation: observation i treated and j untreated, or vice versa. The mismatch bias enters with opposite signs in these two cases. The proposed weighting scheme assigns each observed pair the inverse of the probability with which its realized orientation occurs. Consequently, orientations that are unlikely receive larger weights, while common orientations receive smaller weights. In expectation, both orientations contribute equally, causing the mismatch bias to cancel, leaving only the average treatment effect.

More formally, the touple (D_i, D_j) defines the 'orientation' of match (i, j). The observed pairwise outcome difference $Y_i - Y_j$ determines the contribution of this match to any matching estimator. If $(D_i, D_j) = (1, 0)$, $Y_i - Y_j = \tau_i + B_{ij}$ and so the bias will contribute positively. Conversely, if $(D_i, D_j) = (0, 1)$, $Y_j - Y_i = \tau_j - B_{ij}$. Then, the expected contribution of every matched pair to the τ estimate $\hat{\tau}$ before treatment status is known is

$$\mathbb{E}[\hat{\tau}] = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \Big[\pi_{ij}^{(1,0)}(\tau_i + B_{ij}) + \pi_{ij}^{(0,1)}(\tau_j - B_{ij})\Big] \tag{12}$$

Equation (12) shows that mismatch bias arises because the two possible orientations of a matched pair generally occur with different probabilities. If each orientation instead contributed equally in expectation (which organically happens if $\pi_{ij}^{(1,0)} = \pi_{ij}^{(0,1)} = 0.5$), the bias terms B_{ij} would cancel. The estimator aims to achieve this more generally by weighting each realized pair by the inverse probability of observing it in its current orientation. This ensures that each observed pair contributes its treatment effect in expectation. Mathematically, I aim to estimate

$$\mathbb{E}[\tau^{\hat{IPW}}] = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \frac{1}{2}\Big[\pi_{ij}^{(1,0)}[\frac{1}{\pi_{ij}^{(1,0)}}(\tau_i + B_{ij})] + \pi_{ij}^{(0,1)}[\frac{1}{\pi_{ij}^{(0,1)}}(\tau_j - B_{ij})]\Big] \tag{13}$$

so that orientation probabilities cancel out and the ex ante expectation for each pair is just $\tau_i + \tau_j$. However, this requires to estimate the orientation

probabilities $\pi_{ij}^{(1,0)}$ and $\pi_{ij}^{(0,1)}$.

The orientation probabilities can be decomposed into the probabilities of several events: For the orientation $(1,0)$, three events must occur simultaneously:

1. observation i is treated
2. observation j is untreated
3. every observation closer to i than j is treated, rendering it ineligible as a match.

The same holds vice versa for orientation $(0,1)$. For the further discussion for a given pair (i,j), define

$$\mathcal{K}(i,j) = \{k : d(X_i, X_k) < d(X_i, X_j)\}$$

as the set of observations closer to i than the realized match j. For j to become the realized match of i, all observations in $\mathcal{K}(i,j)$ must possess the "wrong" treatment status and therefore be inadmissible as matches. Under the standard independent treatment assignment conditional on the propensity scores, this occurs with probability $\prod_{k\in\mathcal{K}(i,j)} e_k$ when i is treated and with probability $\prod_{k\in\mathcal{K}(i,j)}(1-e_k)$ when i is untreated.

Conditional on the covariates and on the event that observations i and j form a matched pair, the probability that this pair is realized in the orientation $(D_i, D_j) = (1,0)$ ($\pi_{ij}^{(1,0)}$) is therefore given by

$$\pi_{ij}^{(1,0)} = \frac{e_i(1-e_j)\prod_{k\in\mathcal{K}(i,j)} e_k}{e_i(1-e_j)\prod_{k\in\mathcal{K}(i,j)} e_k + (1-e_i)e_j\prod_{k\in\mathcal{K}(j,i)}(1-e_k)}$$

where e_i and e_j denote the treatment probabilities of either observation. Similarly, the probability that the same pair appears in the opposite orientation $(D_i, D_j) = (0,1)$ is

$$\pi_{ij}^{(0,1)} = \frac{(1-e_i)e_j\prod_{k\in\mathcal{K}(j,i)}(1-e_k)}{e_i(1-e_j)\prod_{k\in\mathcal{K}(i,j)} e_k + (1-e_i)e_j\prod_{k\in\mathcal{K}(j,i)}(1-e_k)}$$

Replacing the true treatment probabilities in these formulas with their corresponding propensity score estimates yields the feasible estimator of 13. It is essentially the original $\hat{\tau}$ multiplied with an inverse probability weight.

$$\mathbb{E}[\tau^{\hat{IPW}}] = \frac{1}{\sum_m^M \lambda_m} \sum_m^M \lambda_m (Y_{T(p)} - Y_{C(p)})$$

$$\lambda_m = \frac{ps_i(1-ps_j)\prod_{k\in\mathcal{K}(i,j)} ps_k + (1-ps_i)ps_j \prod_{k\in\mathcal{K}(i,j)}(1-ps_k)}{ps_i(1-ps_j)\prod_{k\in\mathcal{K}(j,i)} ps_k} \quad \forall \quad i(m)_{D=1} \tag{14}$$

$$\lambda_m = \frac{ps_i(1-ps_j)\prod_{k\in\mathcal{K}(j,i)} ps_k + (1-ps_i)ps_j \prod_{k\in\mathcal{K}(i,j)}(1-ps_k)}{(1-ps_i)ps_j\prod_{k\in\mathcal{K}(j,i)}(1-ps_k)} \quad \forall \quad i(m)_{D=0}$$

If the propensity score is precisely estimated, the bias term will cancel out in expectation. This estimator thus relaxes the requirements on matching quality (the bias term cancels out in expectation regardless of how large it is), but in exchange, it requires a much better estimate of the propensity score than is usually necessary: In contrast to modern propensity score techniques, it is not doubly robust.

Also, extreme $\lambda(...)$ values will give extreme weights to some observations, which can lead to high variance in the estimation. For this reason, I also run a truncated version of the estimate (where all estimated orientation probabilities below 0.1 and above 0.9 are discarded). In the simulations, this adjustment substantially improves performance, but of course comes at the costs of truncating potentially interesting extreme observations. Nevertheless, these observations are often disregarded in the pre-estimation cleaning of observational data anyway when doing propensity score matching (small λ are the result of multiplying several small propensity scores), so this might not be costly in many applications.

Using $ps_i * (1 - ps_j) * \Pi_k ps_k$ for $p_m(i_{D=1} \odot j_{D=0})$ also assumes the independence of shocks in the propensity score estimation. While this is strictly speaking usually assumed anyway, the assumption carries much more weight here and might have to be examined and tested in practical applications. If

errors are not independent, clustering techniques can recover a more complicated structure, which will lead to more complicated weight formula.

$\Pi_k ps_k$ can be arbitrarily expensive to compute in theory. However, in practice, especially after truncation, there are not many observations k between matched observations and the additional computing time is not impractical, even for large samples.

It is worth noting that this estimator can theoretically recover the treatment effect no matter th mismatch between observations, since the bias term drops out in expectation. However, this crucially relies on a consistent, unbiased and precise estimate of the propensity score. In many empirical applications this trade-off may not be attractive, since poor matching often coincides with poor propensity-score estimation. Nevertheless, the two are conceptually distinct. Treatment assignment may be highly predictable even when (or especially if) treated and control observations remain far apart in covariate space. In such settings, the proposed estimator is expected to provide the greatest gains.

The estimand targeted by the proposed estimator differs subtly from the conventional average treatment effect (ATE), average treatment effect on the treated (ATT) and the average treatment effect on the untreated (ATU). These estimands are defined with respect to the realized treatment assignment in the observed sample. For example, the ATT averages treatment effects over the units that happened to receive treatment. The proposed estimator instead adopts an ex ante perspective, treating the realized treatment assignment as a random variable conditional on the propensity scores. It therefore averages over the two possible treatment orientations of every matched pair rather than conditioning on the orientation that happened to be observed. In this sense, the estimator targets the expected treatment effect associated with the matched pairs themselves, rather than the realized treated or untreated subpopulation. At the same time, the feasible estimator in eq. 14 has to be implemented using only the observed outcome differences. It can therefore also be interpreted as a reweighted version of the conventional matching estimator, where observations whose realized treatment orientation was unlikely receive correspondingly larger weights. Whether this distinction

from the conventional ATE or ATT is empirically important depends on the application.

2.4.4 Inference

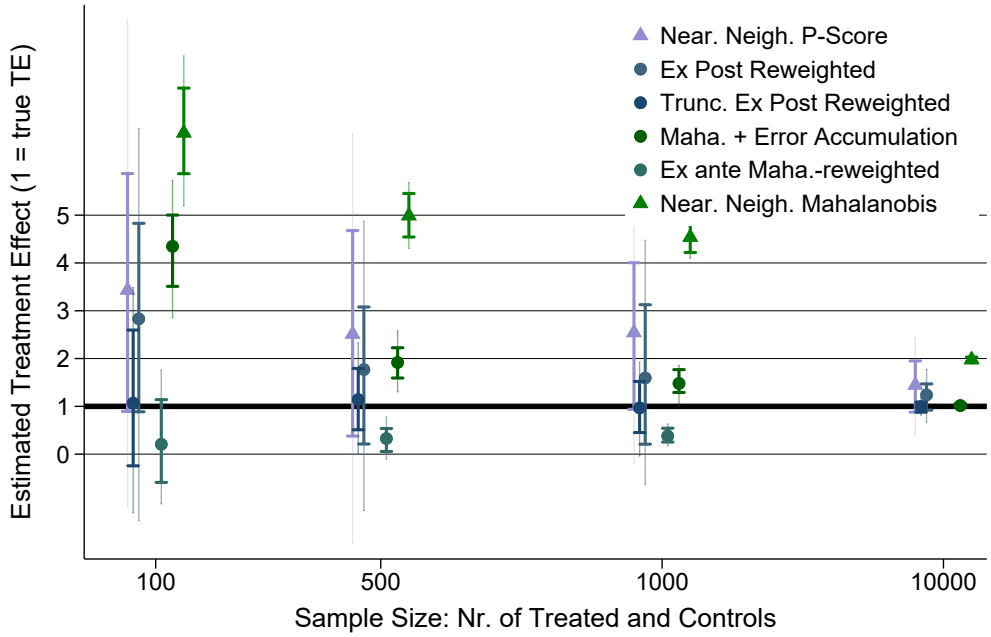
All estimators considered in this paper can be written as the difference between two weighted sums of outcomes, although the algorithms constructing the weights differ across estimators. Since bootstrapping is not generally valid for matching estimators (Abadie and Imbens, 2008) and can be computationally expensive for ex ante reweighted matching especially, inference follows the framework of Abadie and Imbens (2006) as presented in Imbens (2015). Standard errors are constructed by treating each estimator as the difference between two weighted sums while accounting for the fact that individual observations may serve as matches multiple times, implying that the summands are generally not independent. These standard errors are thus constructed conditional on the algorithmic structure, taking it as deterministic. They should be understood as capturing the outcome sampling uncertainty. Throughout the paper, reported standard errors are computed using this approach.

3 Simulation Benchmark

Figure 2 already established that conventional matching estimators can remain severely biased in finite samples despite satisfying the conditions typically invoked to justify matching. In this section, I evaluate the possible corrections discussed in section 2.4 by also applying them to the DGP from 2.3. The aim with all estimations is to recover the ATE, but since the DGP features the same treatment effect for all observations, the various potential estimands are identical ($ATT = ATE = ATU = 500$).

Figure 5 reports the results together with the two most commonly used other matching techniques – Propensity score and Mahalanobis nearest neighbor matching. Three results stand out. First, all conventional matching estimators remain substantially biased despite the fact that the conditional

Figure 5: *Simulation: Performance of Proposed Estimators*



Notes: Performance of conventional and the newly proposed estimators with the DGP described in section 2.1. Y-axis is rescaled by the true treatment effect (500). The new estimators are unstrucated and truncated post reweighting (section 2.4.3), Mahalanobis matching with error accumulation (section 2.4.2) and ex ante reweighted Mahalanobis matching (section 2.4.1; 10.000 sample not reported due to computational costs).
Source: own simulations.

independence assumption is satisfied by construction. Second, all proposed estimators substantially reduce bias despite relying on very different correction mechanisms, which makes me confident that the mismatch process identified in Section 2.1 is indeed the primary source of the finite sample bias and that this bias is quantitatively important (at least in the simulated data). Third, the two computationally simpler approaches – error accumulation and truncated ex post reweighting – perform at least as well as the more demanding ex ante reweighted Mahalanobis estimator.

For the 100 observations sample, only ex post reweighting yields an estimate essentially equal to the true treatment effect, but only when weights

are truncated to discard extreme weights (section 2.4.3). This is not surprising in the sense that this is the most demanding estimator in terms of assumptions. Since these are fulfilled for this DGP, the estimator performs very well. However, already when not truncating the weights, performance declines, since some observations with low propensity scores at the tail of the distributions get very high weights, have noticeable errors (in terms of percentages) in the estimated propensity score and are matched to poorly fitting comparison observations (due to the sparsity of the data in that region of the covariate space). Standard Mahalanobis matching estimates a treatment effect of roughly 700% of the true effect. The new Mahalanobis matching with accumulating errors improves upon it, but is still severely biased. The ex ante reweighted Mahalanobis performs much better estimating a treatment effect at 20% of the true effect.

At 500 matches, both conventional estimators improve their performance, but all 4 new estimators outperform them. Truncated ex post reweighting remains unbiased, while the three others converge further to the true effect, but are still substantially biased. The conventional techniques still estimate effects of 250% or 500%. With 1000 potential matches, the new estimators converge further, while there is little movement in the conventional estimators. Even at 10.000 potential matches, the conventional estimators have not yet converged, whereas the new estimators have.

The improved performance is not bought with increases in variance, in fact all new estimators have less variance than their respective conventional estimator, at least in this simulation. This suggests that the reduction in mismatch improves not only the average quality of the matches but also the stability of the resulting treatment effect estimates.

Taken together, the simulations suggest that the mismatch mechanism identified in Section 2.1 is not merely a theoretical curiosity. In finite samples, it can dominate the behavior of conventional matching estimators and generate biases several times larger than the underlying treatment effect. The proposed adjustments substantially reduce this bias, indicating that the problem can be addressed without abandoning matching-based estimation altogether.

4 Applications

The simulation exercises above demonstrate that systematic mismatch can generate substantial bias even when the conditional independence assumption is satisfied. The next question is whether the same mechanism is quantitatively important in empirical applications. To test this, I study two applications where an actual random experiment was conducted, which gives a natural benchmark against which to test the estimator.

4.1 Application: The National Supported Work Demonstration

To investigate this question, I apply the proposed estimators to the data originally analyzed by LaLonde (1986). The data is based on the National Supported Work Demonstration, a labor market program targeted at individuals with particularly weak labor market attachment. Program participants were randomly assigned either to receive training and job placement assistance or to a control group. As a result, the experimental data provides a credible benchmark estimate of the treatment effect.

The importance of the LaLonde data extends beyond the evaluation of the program itself. LaLonde combined the experimental sample with two nonexperimental comparison groups and asked whether contemporary observational methods could recover the experimental benchmark. His first group of potential controls is from Westat's Matched Current Population Survey–Social Security Administration File, including potential workers (under 55) meeting Westat's criteria (CPS in the following). His second group is from the Panel Study of Income Dynamics and includes household heads under 55 from 1975 to 1978 (PSID in the following). His conclusion was largely negative: The nonexperimental estimators using either group of observations often produced estimates that differed substantially from the experimental result.

Imbens and Xu (2025) revisit the original paper and conclude that with modern estimators, the original effects can often be retrieved. However, they

highlight the importance of understanding treatment assignment mechanism, estimating propensity scores flexibly, verifying overlap and common support, trimming observations with poor overlap, employing modern doubly-robust estimators where possible, and validating identification assumptions through placebo and sensitivity analysis. After substantial trimming of the data and careful examination of outliers, many modern estimators yield something in the vicinity of the actual treatment effect. However, standard errors are very large and results often not significant. When using the PSID data, their trimming changes the population so much that the experimentally confirmed ATT of the surviving observations is close to zero. While modern estimators can recover this parameter, the resulting estimand differs markedly from the original treatment effect of interest.

This application is particularly suited for the present paper: The experimental benchmark provides an objective measure of estimator performance. The problem is challenging for modern estimators due to the sparse data and imperfect control observations. The new estimators discussed in Section 2.1 are designed for situations with finite samples. If the mixed performance of modern estimators is due to systematic mismatch as discussed above, the new estimators should recover the experimental estimate more closely than conventional matching procedures.

I apply my estimators to the untrimmed data to minimize the amount of user discretion. Figure 6 reports mismatch and the results from conventional and new estimators. When using conventional matching in the untrimmed data set, there is substantial systematic mismatch (Panel 6c and 6d). Nearest Neighbor Matching is one of the better performing estimators from Imbens and Xu (2025) and yields an estimate close to the experimental benchmark, but without statistical significance. Error Accumulation in particular estimates a statistically significant effect very close to the experimental baseline.

It is noteworthy that counter to the simulation exercise above, propensity score based methods consistently perform worse than their counterparts (both for new and conventional estimators). In this application, the propensity score might not be estimated correctly (either in terms of specification, finite sample imprecision or omitted variables) leading to finite sample prob-

lems for all estimators based on that metric. The propensity score reweighting technique is especially vulnerable to this, with its reliance not only on propensity score, but on their ratios, leading to extremely imprecise estimations. Its results are unstable, imprecise and substantially different from the ATT in this observation.

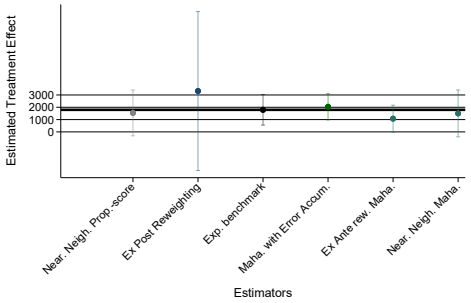
Ex ante reweighting yields results that are not statistically different from the experimental benchmark, but consistently too low. It is worth highlighting that only 4 out of the 10 variables in this application have (quasi-) continuous values and thus are amenable to the optimization used. The implementation used recognizes these issues and reports non-convergence for a substantial share of observations and variables.

Both the error accumulation and the ex ante reweighting technique compare favorably to the conventional estimators reported in Imbens and Xu (2025) with and without trimming, particularly in terms of confidence bands. Importantly, these results are obtained without aggressive trimming of the data or extensive specification search. Given the relatively small sample size and the high variance of the estimators, the primary advantage of the proposed methods in this application appears to lie in precision rather than large reductions in average bias. Most conventional estimators produce confidence intervals spanning from approximately zero to more than twice the experimental treatment effect, making it difficult to determine whether the point estimates are meaningfully biased.

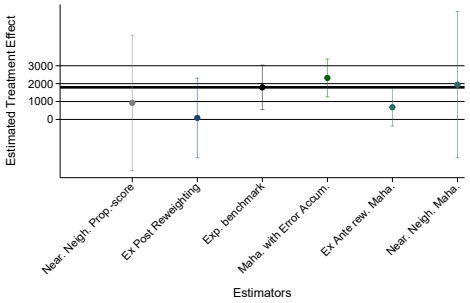
This application further supports the view that aggressive calipers (which are implemented by Imbens and Xu (2025) via trimming) can limit the bias in practice, but this strategy comes with high costs: Dropping observations changes the estimated ATT to one that is of less interest and decreases observations and thus statistical power. Error accumulation in particular can avoid these issues in this application.

Figure 6: *Application: National Supported Work Demonstration*

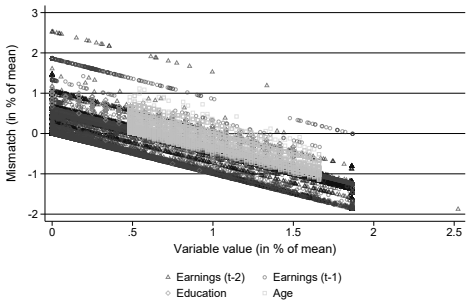
(a) CPS observational data



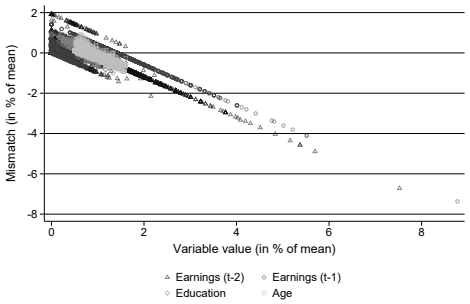
(b) PSID observational data



(c) Prop.-Score Mismatch (CPS)



(d) Prop.-Score Mismatch (PSID)



Notes: Performance of conventional and the newly proposed estimators when evaluating the National Supported Work Demonstration (LaLonde, 1986), using the data from Imbens and Xu (2025). Panels 6a and 6c use control observations from the CPS data and Panel 6b and 6d from PSID. I use the data without trimming. Mahalanobis matching with error accumulation (section 2.4.2) recovers the ATT with much tighter confidence bands. Ex ante reweighted Mahalanobis matching (section 2.4.1) is statistically indistinguishable from the benchmark and has lower variance, but estimates a substantially smaller effect.[a] Truncated ex post reweighting (section 2.4.3) performs worst. Simple Nearest Neighbor matching on Mahalanobis distance also recovers the ATT, though the confidence band is very large. Unadjusted propensity score matching (as an example of conventional estimators) produces large mismatch below the aggregate statistics (Panel 6c and 6d).
Source: Data from Imbens and Xu (2025).

[a]However, note that ex ante reweighted Mahalanobis targets an ATT where observations have different weights and is thus strictly speaking not completely comparable (section 2.4.1).

4.2 Application: Subsidized Home Ownership in Oklahoma

Next, I apply the estimators to Grinstein-Weiss et al. (2013). They study the long-run effects of Individual Development Accounts (IDAs), a savings subsidy program targeted at low-income households. In a randomized experiment conducted in Tulsa, Oklahoma, treatment-group participants received matched savings accounts that subsidized home down payments at a 2:1 rate. Earlier work found that the program substantially increased home ownership rates during the active subsidy period (7%), particularly among renters Mills et al. (2008).

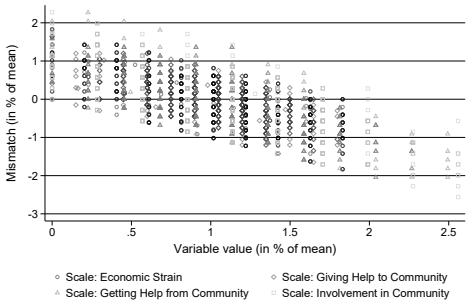
Grinstein-Weiss et al. (2013) examined whether these gains persisted after the program ended. They find that the initially large treatment effects on home ownership largely disappeared over time. Control-group households caught up rapidly after the program ended, implying that the program mainly shifted the timing of home purchases rather than permanently increasing home ownership. By 2009, the estimated effects on home ownership rates, years of ownership, home equity, and foreclosure outcomes were economically small and statistically insignificant.

For the purposes of this paper, the application is useful for two reasons. First, the data features substantial heterogeneity and relatively sparse support. Second, the existence of a randomized benchmark makes it possible to assess whether the proposed estimators reduce mismatch relative to conventional matching procedures and whether this translates into improved finite sample performance. Due to data sparsity, no baseline estimators are statistically significantly different from the 7% difference reported in Mills et al. (2008).

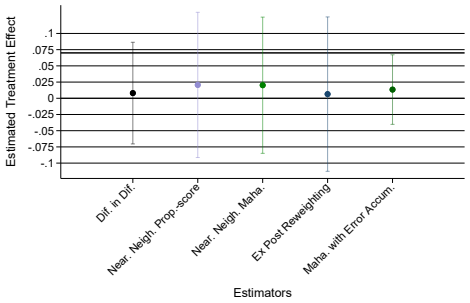
Figure 7 reports the results from my examination. Panel 7a shows that there is systematic mismatch between treatment and control observations in all variables with more than ten values (the study design relies heavily on dummies and categorial variables): Small variable values are matched to larger counterparts and vice versa. This pattern conforms to the trends established in the simulation study. Panel 7b reports the estimation results

Figure 7: *Application: Long Term Effect of IDA on Home Ownership*

(a) Systematic Mismatch



(b) Estimation Results



Notes: Reproduction of Grinstein-Weiss et al. (2013). Panel 7a demonstrates the mismatch in the 4 variables (out of 33) with more than 10 distinct values: The research design relies heavily on dummies and categorization. Panel 7b reports the results from conventional and the newly proposed estimators. The research design evaluates the IDA program of home ownership subsidies in Tulsa 6 years after the end of the program. While the program was running, home ownership was 7% higher in the treatment group (black line).
Source: Data from Grinstein-Weiss et al. (2013).

with conventional matching estimators, ex post reweighting and Mahalanobis matching with error accumulation. Both new estimators are closer to the experimental DiD estimate, they also have lower variance. As in section 4.1, nearest neighbor Mahalanobis matching with accumulating errors is the best performer both in bias and variance. It is also the only estimator that has substantially more statistical power than the simple experimental DiD.

The vector of covariates in this study contains 29 dummy and categorical variables with fewer than 10 distinct values. This makes ex ante reweighted Mahalanobis Matching unsuitable, since the estimator was designed for continuous covariates to produce a differentiable objective. For example, a binary gender indicator cannot be reweighted continuously to eliminate expected mismatch: Females can only be mismatched to males and vice versa. When the estimator is applied nonetheless, all observations are flagged as failed convergence cases. In this sense, the estimator correctly identifies the setting as unsuitable for its underlying optimization procedure. However,

dummy variables are widespread in matching, so this is a shortcoming of this technique in practice.

4.3 Application: Placebo Inventor Program after German Reunification

To also test matching in a difference-in-differences setting, I construct a placebo treatment in a panel of German inventors around reunification. Patent and inventor data provide a frequent setting for matching in a difference-in-differences framework because they contain large numbers of potential controls and rich information on pre-treatment inventor characteristics (for recent applications, see e.g. Domnik et al. (2026); Prato (2025); Bräuer (2024)).

The German state faced substantial challenges in extending a market economy to the previously communist East Germany. Even today, productivity and innovation outcomes in East Germany are substantially below the West German average. A detailed discussion of the historical context, the communist system and its economic impact can be found in Akcigit et al. (2026), a description of the East German patent system and the movement of inventors in Fritsch et al. (2025); Hünermund and Hipp (2025). However, I refrain from discussing the institutional background in detail, since the placebo program is self-constructed.

I use the patent data provided by the EPO (PATSTAT) to study the effect of a hypothetical inventor program in 1992, aimed at increasing innovation in both parts of the country and technology diffusion between them. The program targets inventors who were active both before and after reunification, active in 1991 or 1992 and who were located in either part of Germany both in 1989 (before reunification) and in 1992.

Among eligible inventors, treatment is assigned probabilistically, but more productive and more mobile inventors are more likely to be treated: Treatment propensity is increasing in cumulative patenting in German patent offices (as a measure of the inventor's prior visibility) and in the inventor's cumulative number of previous moves (as a measure of ex ante mobility).

Since the treatment is artificially generated, the true treatment effect is zero by construction.

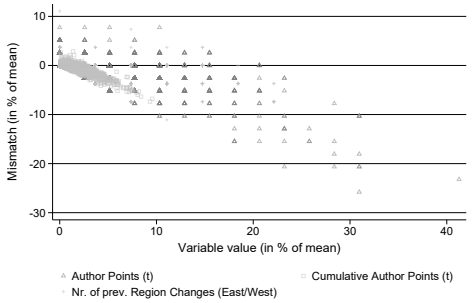
The raw patenting data published by the EPO does not contain readily usable inventor identifiers, complete localization on the East-West Germany level or broad technology indicators appropriate for the study. These are constructed ex post. Most importantly, inventor IDs are disambiguated starting from name strings using coauthor, firm, location and technology connections between inventors with similar names. Locations are taken from inventor and firm address information and imputed where unavailable. After this process, there are roughly 10.000 inventors in East Germany and 60.000 in the West around reunification. Of these, 6.000 end up in the event study either in the treatment or the control group – most inventors only patent once and are thus not eligible. The entire process is described in Bräuer (2024) in detail, but not elaborated here for brevity of exposition: Since the placebo treatment is assigned within the constructed data, the exercise does not require that the underlying data-construction procedures recover the true treatment status of individual inventors.

The matching procedure uses information available prior to treatment. The matching variables comprise cumulative patenting in German and worldwide patent offices, current and lagged annual patenting in German and worldwide patent offices, and cumulative previous mobility. Thus, the treatment determining variables are included among the matching variables.[3] To ensure comparability along particularly important discrete dimensions, matching is performed exactly within cells defined by the inventor's location in 1989, location in 1992, and technological community. The latter is based on the technological content of the inventor's patent portfolio. There are six broad technological communities in the final sample, four predominantly industrial communities, a life-sciences community, and an ICT community. The resulting design therefore combines a plausible source of treatment selection with substantial observable heterogeneity in inventor productivity, career history, technology, and location, while retaining a known zero treatment effect against which the matching estimators can be evaluated.

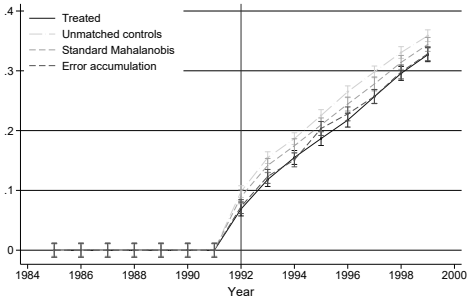
[3]Throughout the application, each inventor receives $\frac{1}{n}$ of a patent with n coauthors.

Figure 8: *Application: Placebo Inventor Program*

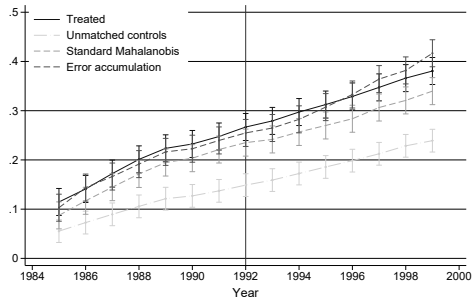
(a) Systematic Mismatch in Matching Variables



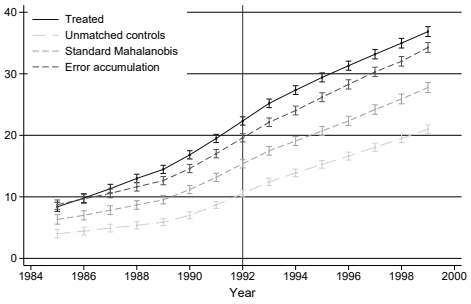
(b) Exit



(c) Nr. of Moves



(d) Patent Output



Notes: Event study of a placebo program for recently patenting inventors in unified Germany in 1992. Any systematic post-1992 deviation from zero is an estimation/design artifact, not an economic effect.
Source: Data from EPO & DPMA.

Figure 8 shows the resulting event study (targeting the ATT) across three different outcomes: cumulative inventor exit (Panel 8b), cumulative number of inventor moves (Panel 8c) and cumulative patents (Panel 8d). As expected, there is substantial systematic mismatch within the matching variables in the pre-treatment periods (Panel 8a). I again find that small values are matched with too large controls and vice versa. The remaining panels report event study estimates based on three estimation strategies: Just using the untreated inventors as controls, using a standard Mahalanobis distance nearest neighbor matching estimator and using the error accumulation technique discussed in section 2.4.2. Conventional matching substantially improves upon using the unmatched controls.

Figure 8 also illustrates that the consequences of the systematic mismatch depend on the outcome function and therefore differ across outcomes: Exit and Nr. of moves appeared to be well matched ex ante, with no pre-trends to speak of. However, treated and control observations diverge post placebo for both variables: There must be a nonlinear relationship between patent productivity and both exit and movement probability. In contrast, patent output remains visibly unbalanced even in aggregate. This is due to the extremely skewed distribution of patent output amongst the potential control inventors: They have 11 patents on average, but with a variance of 103, a Skewness of 3 and a Kurtosis of 21. The top 65 inventors are responsible for 10% of the output in the sample of potential controls. More importantly, 90% of the treated inventors have more controls with less output than with more, explaining the substantial drift. This is a feature of the highly skewed Poisson distribution that approximately describes patenting output in general. As a result of that imbalance, the event study graph for patent output exhibits a clear pre-trend, which should caution the researcher against a causal interpretation of the post-treatment estimates. Error accumulation cannot eliminate the pre-trend entirely, but shrinks it substantially. This illustrates how accounting for accumulated matching error can substantially improve the credibility of a matching-based DiD design, although the remaining pre-trend suggests that further improvements to the estimator would be required in this application, for example by exploiting the time dimension of the data

or explicitly accounting for the unusual distributions of patenting outcome variables.

5 Conclusion

This paper has discussed a structural problem of contemporary matching estimators. I have shown that under reasonable assumptions, $\overrightarrow{X}$ matching estimators have systematic mismatch in finite samples. The underlying mechanism is that some regions in the covariate space have higher densities, so it is more likely that matches can be found there. The estimator will thus gravitate to such regions when matching the surrounding observations. Using simulation exercises, I demonstrate that this drift is substantial, not detected by conventional balance tests and vanishes so slowly when increasing the sample size that all practical applications should be considered finite sample estimates for the purposes of this bias.

This mismatch tendency biases the estimator if the potential outcomes Y_1 and Y_0 are a nonlinear function of $\overrightarrow{X}$ or if there are interactions between variables. This mechanism can also lead to persistent aggregate mismatch when joint distributions of $\overrightarrow{X}$ are degenerate in such a way that 'drift' in matching is always in one direction. However, conventional balance tests will flag such cases and thus likely prevent publication of the results.

I develop three different strategies on how to mitigate or even eliminate this bias: Either, one can observe the drift and then alter the distance metric used in matching to correct this tendency. This is the most direct intervention, but computationally expensive. Alternatively, it is possible to observe mismatch and correct in the other direction when matching 'the next closest' observation. This error accumulation technique breaks the usual independence between the matching of individual observations. The estimator thus consists of both an order in which observations are matched and a distance metric. It performs best in practice. Last, correctly estimated propensity scores yield a formula for the probability with which the bias inherent in each matching pair enters the estimation. It is possible to ex post reweight pairs and theoretically cancel out arbitrary bias, not just from the mecha-

nism studied here. However, this is very sensitive to the propensity score estimate.

I apply both the new and conventional estimators to three applications with public data and treatment randomization, one of which is a difference-in-differences design. These studies yield a natural benchmark, since the true effect is known. In all cases, the new estimators can offer substantial improvements, with error accumulation with Mahalanobis matching performing best. In small applications, conventional estimators have large standard errors, which are substantially reduced after correcting for systemic mismatch. In two cases, the error accumulation technique recovers the experimental benchmark faithfully. In the event study, it substantially improves on conventional matching but cannot fully eliminate the dynamic bias with the given sample size. Among the three new estimation strategies, it is the most robust in practice as it does not rely on either precisely estimated propensity scores or on continuous optimization techniques. Since it is also the conceptually simplest estimator, I recommend its use for future matching exercises.

6 Bibliography

G. King, R. Nielsen, Why Propensity Scores Should Not Be Used for Matching, Political Analysis 27 (2019) 435–454.

D. W. Ham, L. Miratrix, Benefits and costs of matching prior to a Difference in Differences analysis when parallel trends does not hold, 2024. ArXiv:2205.08644 [stat.ME].

A. Abadie, G. Imbens, Large Sample Properties of Matching Estimators for Average Treatment Effects, Econometrica 74 (2006) 235–267. Publisher: Econometric Society.

A. Abadie, G. W. Imbens, Bias-Corrected Matching Estimators for Average Treatment Effects, Journal of Business & Economic Statistics 29 (2011) 1–11. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/jbes.2009.07333.

A. Abadie, G. W. Imbens, Matching on the Estimated Propensity Score, Econometrica 84 (2016) 781–807. Publisher: Econometric Society.

J. Roth, P. H. C. Sant'Anna, A. Bilinski, J. Poe, What's trending in difference-in-differences? A synthesis of the recent econometrics literature, Journal of Econometrics 235 (2023) 2218–2244.

J. A. Smith, P. E. Todd, Does matching overcome LaLonde's critique of nonexperimental estimators?, Journal of Econometrics 125 (2005) 305–353.

K. Lee, Y. Jeong, S. Han, S. Joo, J. Park, K. Qi, Difference-in-Differences with matching methods in leadership studies: A review and practical guide, The Leadership Quarterly 36 (2025) 101813.

S. J. Lee, J. M. Wooldridge, A Simple Transformation Approach to Difference-in-Differences Estimation for Panel Data, 2026.

B. Kim, M.-j. Lee, Overlap-weighted difference-in-differences: A simple way to overcome poor propensity score overlap, Economics Letters 250 (2025). Publisher: Elsevier.

J. Hainmueller, Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies, Political Analysis 20 (2012) 25–46.

K. Imai, M. Ratkovic, Covariate Balancing Propensity Score, Journal of the Royal Statistical Society Series B: Statistical Methodology 76 (2014) 243–263.

A. Abadie, G. W. Imbens, On the Failure of the Bootstrap for Matching Estimators, Econometrica 76 (2008) 1537–1557. Publisher: [Wiley, Econometric Society].

G. W. Imbens, Matching Methods in Practice: Three Examples, The Journal of Human Resources 50 (2015) 373–419. Publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System].

R. J. LaLonde, Evaluating the Econometric Evaluations of Training Programs with Experimental Data, The American Economic Review 76 (1986) 604–620. Publisher: American Economic Association.

G. W. Imbens, Y. Xu, Comparing Experimental and Nonexperimental Methods: What Lessons Have We Learned Four Decades after LaLonde (1986)?, Journal of Economic Perspectives 39 (2025) 173–202.

M. Grinstein-Weiss, M. Sherraden, W. G. Gale, W. M. Rohe, M. Schreiner, C. Key, Long-Term Impacts of Individual Development Accounts on Homeownership among Baseline Renters: Follow-Up Evidence from a Randomized Experiment, American Economic Journal: Economic Policy 5 (2013) 122–145.

G. Mills, W. G. Gale, R. Patterson, G. V. Engelhardt, M. D. Eriksen, E. Apostolov, Effects of individual development accounts on asset purchases and saving behavior: Evidence from a controlled experiment, Journal of Public Economics 92 (2008) 1509–1530.

C. Domnik, A. Esparza, M. Prato, S. Weik, Business Drain: Aggregate Effects of Startup Reallocation, 2026.

M. Prato, The Global Race for Talent: Brain Drain, Knowledge Transfer, and Growth, The Quarterly Journal of Economics 140 (2025) 165–238.

R. Bräuer, The Aggregate Effects of Incumbent Firms Preventing Disruptive Innovation, 2024.

U. Akcigit, R. Bräuer, A. Markevich, J. Mirande, A. Zherdeva, Battle of Ideologies: Firm Dynamics and Productivity in Planned Versus Market Economies, 2026.

M. Fritsch, M. Greve, M. Wyrwich, The long-lasting impact of historical shocks on innovation activity: The case of German separation and reunification, Technovation 148 (2025) 103318.

P. Hünermund, A. Hipp, Inventor Mobility After the Fall of the Berlin Wall, 2025. ArXiv:2409.01861 [econ.GN].

R. Bräuer, Matching on the Global Inventor Firm Labor Market, 2024.