Biometrika (2023), **110**, 3, *pp.* 631–644
https://doi.org/10.1093/biomet/asac066
Advance Access publication 30 November 2022

On the statistical role of inexact matching in observational studies

By KEVIN GUO and DOMINIK ROTHENHÄUSLER

Department of Statistics, Stanford University,
390 Jane Stanford Way, Stanford, California 94305, U.S.A.
kxguo@stanford.edu   rdominik@stanford.edu

Summary

In observational causal inference, exact covariate matching plays two statistical roles: (i) it effectively controls for bias due to measured confounding; (ii) it justifies assumption-free inference based on randomization tests. In this paper we show that inexact covariate matching does not always play these same roles. We find that inexact matching often leaves behind statistically meaningful bias, and that this bias renders standard randomization tests asymptotically invalid. We therefore recommend additional model-based covariate adjustment after inexact matching. In the framework of local misspecification, we prove that matching makes subsequent parametric analyses less sensitive to model selection or misspecification. We argue that gaining such robustness is the primary statistical role of inexact matching.

Some key words: Matching; Observational causal inference; Randomization test.

1. Introduction

1.1. *Motivation*

We consider the problem of using a large observational dataset $\{(X_i, Y_i, Z_i)\}_{i\leqslant n}$ to test whether a binary treatment $Z_i \in \{0, 1\}$ has any causal effect on an outcome $Y_i \in \mathbb{R}$. The vector $X_i \in \mathbb{R}^d$ contains covariates whose confounding effects must be controlled away. Throughout this article, we assume that all relevant confounders are contained in X_i.

For over 70 years statisticians have been tackling such problems using matching methods. These methods control for the effects of the covariates X_i by pairing each treated observation i with a similar untreated observation $m(i)$. Conceptually, this pairing process is often seen as extracting an approximate randomized experiment from an observational dataset (Rubin, 2007; Rosenbaum, 2010; King & Nielsen, 2019).

When all treated observations are matched exactly, i.e., $X_i = X_{m(i)}$, a matched observational study reconstructs a randomized experiment in a statistically precise sense: conditional on the matches, the treatment distribution among matched units is the same as the treatment distribution in a paired experiment (Rosenbaum, 2002, § 3.2.3). Using this connection, p-values and confidence intervals can be computed for exactly matched observational studies using the same design-based randomization tests originally developed for experiments. These tests exploit only the randomness in Z and thus are valid without any assumptions on the X-Y relationship.

However, it is often not possible to find an exact match for every treated observation. For example, no treated units will be exactly matched when covariates are continuously distributed.

In the presence of inexact matches, the precise statistical connection between matched-pairs studies and randomized experiments breaks down. Unlike in a paired experiment, the treatment distribution in an inexactly matched observational study is neither uniform within pairs (Hansen, 2009) nor independent across pairs (Pashley et al., 2021, § 5). As a result, standard randomization tests based on uniformity and independence lose their finite-sample validity. Indeed, Shah & Peters (2020) have shown that no nontrivial test can have assumption-free validity when continuous covariates are present.

This paper examines the statistical role that matching plays when not all units can be matched exactly. We consider two main possibilities.

(A) Perhaps matching discrepancies typically become negligible in large samples, so that standard randomization tests remain asymptotically valid despite not being finite-sample exact. If so, the statistical role of inexact matching would be the same as that of exact matching: controlling overt bias and providing a basis for nonparametric inference.

(B) Alternatively, matching discrepancies could remain statistically meaningful even in large samples and render standard randomization tests invalid. If so, additional model-based adjustment after matching would be necessary to obtain valid inference. In this case, Ho et al. (2007) have argued that the statistical role of matching is to provide a pre-processing step that makes subsequent model-based inferences less sensitive to model selection or misspecification.

1.2. *Outline and overview of results*

In the first part of this article, we investigate possibility (A) by studying the large-sample properties of randomization tests in matched observational studies. Our formal results are developed for the optimal Mahalanobis matching scheme of Rosenbaum (1989). However, much of the intuition extends to other matching schemes.

We find that conventional randomization tests are not generally valid in large samples, even under strong smoothness assumptions. In fact, their Type I error may be dramatically inflated even when the true outcome model is linear, only a handful of covariates are present, and conventional balance tests are passed. The main issue is that covariate matching does not eliminate bias at a fast rate. A secondary issue is that randomization tests may underestimate the sampling variance of commonly used test statistics. Previously, Abadie & Imbens (2006), Sävje (2021) and others have reported on this bias, although the variance issue seems to be a new finding.

Therefore, we caution against applying standard randomization tests after matching inexactly. Although the idea that matching approximates a randomized experiment is a useful conceptual tool (Rubin, 2007, 2008), the analogy is often not precise enough to form the basis for inference.

In the second part of this article, we argue that (B) provides a more compelling justification for inexact matching. We prove that in an appropriately matched dataset, p-values based on linear regression remain approximately valid, even if the linear model is locally misspecified. Moreover, after matching, all sufficiently accurate model specifications will yield nearly identical t-statistics. These results give formal support to claims made by Rubin (1973, 1979), Ho et al. (2007) and others. However, our analysis yields additional insights. In particular, we find that it is generally necessary to use matching with replacement, rather than pair matching, to achieve the full extent of robustness attainable by matching.

Based on these results, we recommend model-based adjustment and inference after matching inexactly. Conceptually, this mode of inference makes transparent that structural assumptions, such as approximate linearity, are still required for reliable inference after inexact matching. It also cleanly separates the randomness used for study design, X_i and Z_i, from the randomness used in outcome analysis, Y_i. The Bayesian approach advocated by Rubin (1991) also has these conceptual advantages, but the present article focuses on frequentist inference.

In summary, our findings suggest a rethinking of the role of inexact matching in observational studies. On its own, covariate matching may not remove enough bias to justify the use of assumption-free randomization tests. Moreover, inexact matching may lead randomization tests to underestimate the sampling variability of common test statistics. However, matching does play an important role in the design stage, by making downstream parametric analyses more robust to model selection or misspecification.

1.3. *Setting*

The setting of this article is the Neyman–Rubin causal model with an infinite superpopulation. We assume that units $\{(X_i, Y_i(0), Y_i(1), Z_i)\}$ are independent samples from a common distribution P and that only (X_i, Y_i, Z_i) is observed, where $Y_i = Y_i(Z_i)$. The problem of interest is to use the observed data to test Fisher's sharp null hypothesis,

$$H_0: \ Y_i(0) = Y_i(1) \text{ with probability 1 under } P. \tag{1}$$

All our results extend to testing a constant treatment effect, $H_\tau : Y_i(0) + \tau = Y_i(1)$. However, they do not extend to the weak null hypothesis $E\{Y_i(1) - Y_i(0)\} = 0$.

Throughout, we assume that the underlying population satisfies a few conditions.

Assumption 1. The distribution P satisfies the following conditions.

(i) Unconfoundedness: $\{Y(0), Y(1)\} \perp\!\!\!\perp Z \mid X$.
(ii) Overlap: $P(Z = 0 \mid X) \geqslant \delta > 0$.
(iii) More controls than treated: $0 < P(Z = 1) < 0.5$.
(iv) Moments: $\|X\|$ and Y have more than four moments.
(v) Nonsingularity: $\text{var}(X \mid Z = 1) \succ 0$ and $\text{var}(Y \mid X, Z) > 0$.

Unconfoundedness and overlap are standard identifying assumptions. Meanwhile, the condition $0 < P(Z = 1) < 0.5$ ensures that treated observations exist, and that it is eventually possible to find an untreated match for each treated observation. The last two conditions are needed for various technical reasons, for instance to ensure that the Mahalanobis distance exists.

The analysis in this paper is asymptotic, and we assume that the sample size n becomes large as the dimension d stays fixed. In fact, following the advice of Rubin (1980), we recommend thinking of d as a fairly small number, say 8 or less.

The key asymptotic concept studied in this paper is the asymptotic validity of p-values.

Definition 1 (Asymptotic validity). *A sequence of p-values $\hat{p}_n$ is said to be asymptotically valid at $P \in H_0$ if (2) holds for every $\alpha \in (0, 1)$ under independent sampling from P:*

$$\limsup_{n\to\infty} \operatorname{pr}(\hat{p}_n < \alpha) \leqslant \alpha. \tag{2}$$

Remark 1 (*Alternative sampling models*). The independent sampling model used in this paper differs from several alternatives in the matching literature. One is the design-only framework which models Z_i as random, but treats both the matching and the unit characteristics $\{X_i, Y_i(0), Y_i(1)\}$ as fixed (Rosenbaum, 2002). While this simplifies many issues, it precludes analysing the typical size of matching discrepancies. It also assumes away the complex dependence between the treatments Z_i and the matching $\mathcal{M}$, which may be practically relevant (Pimentel, 2022). Another alternative assumes that the number of untreated observations N_0 grows much faster than the number of treated observations N_1. For example, Abadie & Imbens (2012) and Ferman (2021) assumed that $N_0 \gg N_1^{d/2}$. This scaling is favourable for matching, but the sample size requirement is stringent even for $N_1 = 100$ and $d = 5$. We find that the standard sampling regime gives better approximations in problems where N_0 is only a constant multiple of N_1.

2. Large-sample properties of paired randomization tests

2.1. *Optimal matching and Fisher's randomization test*

In this section, we present our findings on the large-sample properties of standard randomization tests in inexactly matched observational studies.

The pair-matching procedure we study is the optimal Mahalanobis matching scheme of Rosenbaum (1989). This matching scheme pairs each treated observation i with a unique untreated observation $m(i)$ in a way that minimizes the total Mahalanobis distance across pairs,

$$\sum_{Z_i=1} \{(X_i - X_{m(i)})^{\mathrm{T}}\, \hat{\Sigma}^{-1} (X_i - X_{m(i)})\}^{1/2}. \tag{3}$$

Ties may be broken arbitrarily. In (3), $\hat{\Sigma}$ denotes the sample covariance matrix of X, and we arbitrarily set $\hat{\Sigma}^{-1} = I_{d\times d}$ when $\hat{\Sigma}$ is singular. We also denote the set of matched units by $\mathcal{M} = \{i : Z_i = 1 \text{ or } i = m(j) \text{ for some treated unit } j\}$.

The randomization test we study is the paired Fisher randomization test. This test computes a p-value for the null hypothesis (1) as follows. First, the user computes a test statistic $\hat{\tau} \equiv \hat{\tau}(\{(X_i, Y_i, Z_i)\}_{i\in\mathcal{M}})$ on the matched data. Two widely used test statistics are the difference-of-means statistic (4) and the regression-adjusted statistic (5):

$$\hat{\tau}^{\mathrm{DM}} = \frac{1}{N_1} \sum_{Z_i=1} \{Y_i - Y_{m(i)}\}, \tag{4}$$

$$\hat{\tau}^{\mathrm{REG}} = \arg\min_{\tau\in\mathbb{R}} \min_{(\gamma,\beta)\in\mathbb{R}^{1+d}} \sum_{i\in\mathcal{M}} (Y_i - \gamma - \tau Z_i - \beta^{\mathrm{T}} X_i)^2. \tag{5}$$

Then, conditional on the original data $\mathcal{D}_n = \{(X_i, Y_i, Z_i)\}_{i \leqslant n}$, the user defines pseudo-assignments $\{Z_i^*\}_{i \in \mathcal{M}}$ by randomly permuting the true assignments Z_i across matched pairs. Finally, the p-value is defined as $\hat{p} = \text{pr}(|\hat{\tau}_*| \geqslant |\hat{\tau}| \mid \mathcal{D}_n)$, where $\hat{\tau}_* \equiv \hat{\tau}(\{(X_i, Y_i, Z_i^*)\}_{i \in \mathcal{M}})$ is the test statistic evaluated using the pseudo-assignments instead of the true ones. When there are more treated than control units or no treated units, we arbitrarily set $\hat{p} = 1$ since $\mathcal{M}$ is undefined.

The distribution of $\hat{\tau}_*$ given $\mathcal{D}$ is called the randomization distribution of $\hat{\tau}_*$. In practice, this distribution will be approximated using randomly sampled permutations. Since the approximation error can be made arbitrarily small by sampling a large number of permutations, we will consider the idealized case where randomization probabilities are computed exactly.

2.2. *The paired Fisher randomization test is not generally valid*

In this subsection, we give theoretical and numerical examples showing that the paired Fisher randomization test may fail to control asymptotic Type I error even in problems with smooth propensity scores and outcome models. We also explain what goes wrong in each example. All formal claims are proved in the Supplementary Material.

Throughout, we use the following notation: $e(x) = P(Z = 1 \mid X = x)$ is the propensity score, $\hat{p}^{\text{DM}}$ is the randomization p-value based on the difference-of-means statistic (4), and $\hat{p}^{\text{REG}}$ is the randomization p-value based on the regression-adjusted statistic (5).

Example 1 (*One covariate*). Our first example is based on the analysis in Sävje (2021). Suppose that $P \in H_0$ satisfies Assumption 1 and that

$$\begin{aligned} X &\sim \text{Un}(0,1), \\ Z \mid X &\sim \text{Ber}(\theta_0 + \theta_1 X), \\ Y \mid X, Z &\sim N(\beta_0 + \beta_1 X, \sigma^2). \end{aligned}$$

If $\beta_1 \neq 0$ and $P\{e(X) \geqslant 0.5\} > 0$, then $\text{pr}(\hat{p}^{\text{DM}} < \alpha) \to 1$ for every $\alpha \in (0,1)$. In other words, if overt bias is present and any units in the population have propensity score greater than 0.5, then the paired Fisher randomization test will almost always make a false discovery. In this example, the same conclusion would hold for optimal propensity score matching.

Example 2 (*Two covariates*). A multivariate analogue of Example 1 can be constructed using the method of Rubin (1976). Let $D = \{x \in \mathbb{R}^2 : \|x\|_2 \leqslant 1\}$ be the unit disc in the plane. Let $P \in H_0$ be any distribution satisfying Assumption 1 and suppose that

$$\begin{aligned} X &\sim \text{Un}(D), \\ Z \mid X &\sim \text{Ber}\{0.35(1 + \theta^{\mathrm{T}} X)\}, \\ Y \mid X, Z &\sim N(\theta^{\mathrm{T}} X, \sigma^2) \end{aligned}$$

for some $\theta \in D$. A typical large sample from this distribution will have nearly twice as many untreated units as treated units. However, some of those units will have propensity scores greater than 0.5. As a result, we still have $\text{pr}(\hat{p}^{\text{DM}} < \alpha) \to 1$ for every $\alpha \in (0,1)$. The same conclusion holds under optimal propensity score matching, nearest-neighbour matching or any other maximal pair-matching scheme.

In both of these examples, the paired Fisher randomization test fails because the test statistic $\hat{\tau}^{\rm DM}$ is asymptotically biased, but the randomization distribution does not replicate this bias. The bias is mainly driven by the presence of units with propensity scores greater than 0.5, because any pair-matching scheme must eventually run out of close matches for these units. After all, treated units outnumber untreated units in regions of covariate space with $e(x) > 0.5$. See Sävje (2021) for a careful discussion of this issue.

In large samples, a careful analyst may detect this bias and attempt to remove it by using regression adjustment. For example, Carpenter (1977) stated that 'if the residual bias after matching is unacceptably large it may be removed by analysis of covariance'. However, the following example shows that this may not be enough to rescue the randomization p-value.

Example 3 (*Regression-adjusted test statistics*). Let P satisfy the requirements of Example 1, including the condition $P\{e(X) \geqslant 0.5\} > 0$. Consider the paired Fisher randomization test based on the regression-adjusted test statistic $\hat{\tau}^{\rm REG}$. Standard least-squares theory tells us that this test statistic is exactly unbiased in finite samples. Nevertheless, we have

$$\limsup_{n\to\infty} \text{pr}(\hat{p}^{\rm REG} < \alpha) > \alpha$$

for every $\alpha \in (0, 1)$. Thus, even the paired Fisher randomization test based on a correctly specified regression model does not control asymptotic Type I error, when units with propensity scores greater than 0.5 are present.

The issue here is more subtle and is caused by a variance mismatch. Since matching fails to balance covariates when $P\{e(X) \geqslant 0.5\} > 0$, the sample correlation between X_i and Z_i in $\mathcal{M}$ does not vanish in large samples. However, X_i and Z_i^* are uncorrelated in the randomization distribution. Correlation harms precision in least-squares regression, so this mismatch leads the randomization variance of $\hat{\tau}_*^{\rm REG}$ to underestimate the sampling variance of $\hat{\tau}^{\rm REG}$.

Rather than using regression-adjusted test statistics, some authors have recommended using only pair matching in populations where all units have propensity scores of less than 0.5. For example, Yu et al. (2020) stated, with the notation edited to match ours, 'In concept in large samples, pair matching is feasible if $\{1 - e(x)\}/e(x) > 1$ for all x.'

In such populations, it is eventually possible to find an arbitrarily close match for every treated unit. As a result, $\hat{\tau}^{\rm DM}$ will be asymptotically unbiased and consistent. However, asymptotic unbiasedness is not enough to justify randomization tests. Valid inference requires biases to be so small that 'they are buried in estimated standard errors' (Rubin, 2022). Since the standard errors of the randomization distribution tend to zero at rate $n^{-1/2}$ (Bai et al., 2022), valid inference requires the bias to decay at a rate faster than $n^{-1/2}$. This is stated formally in the following proposition.

Proposition 1 (Bias requirement). *Suppose that $P \in H_0$ satisfies Assumption 1 and $P\{e(X) < 0.5\} = 1$. Then the randomization p-value $\hat{p}^{\rm DM}$ is asymptotically valid if and only if $E[\hat{\tau}^{\rm DM} \mid \{(X_i, Z_i)\}_{i\leqslant n}] = o_P(n^{-1/2})$.*

When the linear model $E(Y \mid X, Z) = \gamma + \beta^{\rm T} X$ holds, Proposition 1 requires that optimal matching achieve very fine covariate balance in the direction of β:

$$\frac{1}{N_1} \sum_{Z_i=1} \{X_i - X_{m(i)}\}^{\rm T} \beta = o_P(n^{-1/2}). \qquad (6)$$

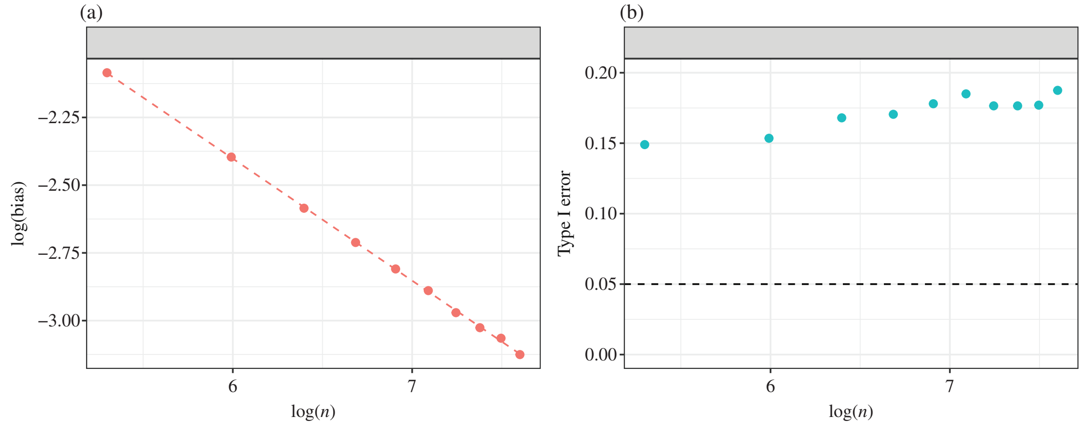

Fig. 1. (a) Average bias of the difference-of-means statistic in Example 4 at various sample sizes, plotted on a log-log scale; the slope of the best-fit line is approximately −0.45, suggesting that the bias does not satisfy the $o(n^{-1/2})$ decay rate required by Proposition 1. (b) Type I error of paired randomization tests of nominal level 5% based on the difference-of-means statistic. All matches were computed in R (R Development Core Team, 2023) using the optmatch package (Hansen & Klopfer, 2006); randomization p-values were approximated using 1000 randomly sampled permutations.

Unless the dimension d is very small, Theorem 2(ii) of Abadie & Imbens (2006) suggests that such strict balance is hard to achieve. This is illustrated by the following numerical example.

Example 4 (*Slow bias decay*). For various sample sizes n between 200 and 2000, we sampled data from the linear/logistic model

$$\begin{aligned} X &\sim \mathrm{Un}([-1,1]^4), \\ Z \mid X &\sim \mathrm{Ber}[1/\{1+\exp(1.1-X_1)\}], \\ Y \mid X, Z &\sim N(3X_1, 1). \end{aligned}$$

This was repeated 2000 times per sample size. In each simulation, we recorded the bias of $\hat{\tau}^{\mathrm{DM}}$ and the paired randomization test p-value. The results are shown in Fig. 1. Even though this distribution has $P\{e(X) < 0.475\} = 1$, the paired randomization test performs poorly because of failure of the bias condition (6). In fact, the Type I error of tests of nominal level 5% appears to increase with the sample size.

Remark 2 (*Balance tests*). The balance condition (6) will not hold even in a completely randomized experiment, so the asymptotic validity of $\hat{p}^{\mathrm{DM}}$ cannot be certified by any balance test with a completely randomized reference distribution. This includes the two-sample t-test and all the examples in Rosenbaum (2010, Ch. 10). Indeed, in each of our simulations in Example 4, we also performed a balance check of nominal level 10% using Hotelling's T^2-test. Imbalance was not detected even a single time. A balance test based on a paired experiment reference distribution would be powerful enough to detect cases where $\mathrm{pr}(\hat{p}^{\mathrm{DM}} < \alpha) \to 1$ (Hansen, 2009), though not powerful enough to certify $\mathrm{pr}(\hat{p}^{\mathrm{DM}} < \alpha) \to \alpha$. See Imai et al. (2007), Austin (2008), Hansen (2008) and Branson (2021) for further discussion of balance tests.

Remark 3 (*Calipers*). Some of the poor behaviour in these examples could be avoided or mitigated by using calipers on the propensity score or the raw covariates (Hansen, 2009).

However, the correct scaling of the caliper is a delicate issue which goes beyond the scope of this paper. In the simulations of Pimentel (2022), propensity calipers did help to reduce the false positive rate of the paired Fisher randomization test. On the other hand, the simulations of Schafer & Kang (2009) used propensity calipers, but still found $\hat{\tau}^{\mathrm{DM}}$ to be severely biased. Kim et al. (2021) studied randomization tests based on finely calipered stratifications of the covariate space. Their results suggest that in moderate dimensions, obtaining validity via covariate calipers may require discarding the vast majority of treated observations.

2.3. *Sufficient conditions for validity*

The examples in the previous subsection show that stringent sampling assumptions are required for the paired Fisher randomization test to be asymptotically valid. For completeness, in this subsection we give two sets of sufficient conditions that make this work.

First, we consider the test based on the difference-of-means statistic. To control the bias that ruins the validity in Examples 1 and 2, we must assume that no units have propensity scores greater than 0.5. However, Example 4 shows that this is not enough and we also need to restrict attention to very low-dimensional problems.

Proposition 2 (Sufficient conditions for $\hat{p}^{\mathrm{DM}}$). *Suppose that $P \in H_0$ satisfies Assumption 1 and the following conditions.*

(i) *No large propensity scores: $P\{e(X) < 0.5 - \eta\} = 1$ for some $\eta > 0$.*
(ii) *Smooth outcome model: the map $x \mapsto E(Y \mid X = x)$ is Lipschitz-continuous.*
(iii) *One continuous covariate: X_i is scalar-valued and has a continuous, positive density supported on a compact interval.*

Then the p-value $\hat{p}^{\mathrm{DM}}$ based on the difference-of-means statistic (4) is asymptotically valid.

We conjecture that asymptotic validity continues to hold with up to three continuous covariates. However, proving this is beyond our current abilities. In dimension four, Example 4 suggests that asymptotic validity will no longer hold.

Next, we consider the test based on the regression-adjusted test statistic. If we assume the model is correctly specified, then the bias of $\hat{\tau}^{\mathrm{REG}}$ is controlled, even if more than three continuous covariates are present. Meanwhile, the variance mismatch in Example 3 can be ruled out by assuming that no units have propensity scores greater than 0.5.

Proposition 3 (Sufficient conditions for $\hat{p}^{\mathrm{REG}}$). *Suppose that $P \in H_0$ satisfies Assumption 1 and the following conditions.*

(i) *No large propensity scores: $P\{e(X) < 0.5 - \eta\} = 1$ for some $\eta > 0$.*
(ii) *Correctly specified outcome model: $E(Y \mid X, Z) = \gamma + \beta^{\mathrm{T}} X$ for some $(\gamma, \beta) \in \mathbb{R}^{1+d}$.*

Then the p-value $\hat{p}^{\mathrm{REG}}$ based on the regression-adjusted test statistic (5) is asymptotically valid.

By appropriately modifying our proofs, the same conclusions can be extended to other correctly specified parametric regression models, such as logistic regression. However, under assumption (ii) in Proposition 3, it is not necessary to use randomization inference for hypothesis testing. Model-based sandwich standard errors would work just as well, if not better, since they remain valid even when the propensity score condition (i) fails. Meanwhile, Example 3 shows that Fisher's randomization test may be invalid when large propensity scores are present.

Remark 4 (*Randomization tests versus randomization inference*). Although the results in this section provide some justification for randomization tests, the justifications are not truly design-based. The key principle of design-based randomization inference is to base probability statements on the conditional randomness in Z_i given everything else. However, as Pashley et al. (2021) and Pimentel (2022) point out, the conditional distribution of $(Z_i)_{i\in\mathcal{M}}$ given $\mathcal{M}$ and $\{(X_i, Y_i(0), Y_i(1)\}_{i\leqslant n}$ is highly intractable unless all matches are exact. To get around this issue, the proofs of Propositions 2 and 3 actually condition on treatments and use the outcome as the source of randomness. In other words, the justification has nothing to do with design.

3. An alternative role for matching

3.1. *Combining matching with parametric outcome modelling*

In this section, we recommend an alternative framework for inference after matching. Specifically, we suggest viewing matching as a pre-processing step for a conventional statistical analysis based on parametric outcome models. This type of post-matching analysis has been recommended by Ho et al. (2007) and Stuart (2010).

The leading example we have in mind is an analysis that fits the linear regression model (7) in the matched sample, interrogates the linearity assumption using specification tests or diagnostic plots, and reports inferential summaries for the coefficient $\hat{\tau}^{\mathrm{REG}}$ based on heteroskedasticity-consistent robust standard errors:

$$(\hat{\tau}^{\mathrm{REG}}, \hat{\gamma}, \hat{\beta}) = \underset{(\tau,\gamma,\beta)}{\arg\min} \sum_{i\in\mathcal{M}} (Y_i - \gamma - \tau Z_i - \beta^{\mathrm{T}} X_i)^2. \tag{7}$$

Let us give some motivation for this approach. It is well known that accurate outcome modelling improves a matched analysis by cleaning up the residual imbalances that remain after matching (Rubin, 1973, 1979). It may be less clear what role matching plays in improving an outcome analysis that already makes parametric assumptions. For example, Hill (2002) asks: 'If the response surfaces are linear why wouldn't standard regression work just as well for covariance adjustment, even perhaps more efficiently than [matching methods]?'

Our main contribution in this respect is to prove that matching improves parametric outcome analysis by reducing sensitivity to model selection and misspecification. Indeed, we regard this as the primary statistical role of matching. Prior empirical work has made the same point, using a combination of simulations and informal arguments. However, our formal analysis yields additional insights. For example, we show that more robustness is gained from matching with replacement than from optimal pair matching.

3.2. *The local misspecification framework*

To study the role of misspecification, we consider a class of nonlinear models defined through small perturbations of a baseline linear model.

Let $P \in H_0$ be some distribution satisfying the linear outcome model $E(Y \mid X, Z) = \gamma + \beta^{\mathrm{T}} X$. For any bounded nonlinear function $g : \mathbb{R}^d \to \mathbb{R}^k$, let $P_{h,g}$ be the distribution of the vector $\{X, Y(0) + h^{\mathrm{T}} g(X), Y(1) + h^{\mathrm{T}} g(X), Z\}$ when $\{X, Y(0), Y(1), Z\} \sim P$. Since we have simply added the same nonlinearity to both potential outcomes, Fisher's sharp null hypothesis (1) continues to hold under $P_{h,g}$. However, the outcome model now contains a nonlinear term:

$$E_{h,g}(Y \mid X, Z) = \gamma + \beta^{\mathrm{T}} X + h^{\mathrm{T}} g(X).$$

A sequence of models $\{P_{h_n,g}\}_{n \geqslant 1}$ is said to be locally misspecified if the coefficient $h \equiv h_n$ tends to zero with the sample size at rate $n^{-1/2}$. This scaling is meant to model problems where nonlinearities are large enough to affect inference, but not so large that they can easily be caught. In smooth models, no specification test can consistently detect the nonlinearity in a locally misspecified sequence (Leeb & Pötscher, 2006, Lemma A.1).

We say that a sequence of p-values is robust to local misspecification near $P \in H_0$ if it remains asymptotically valid, even when the linear model is locally misspecified. A more formal definition is the following.

Definition 2 (Locally robust p-values). *A sequence of p-values $\hat{p}_n$ is said to be robust to local misspecification near $P \in H_0$ if* (8) *holds for every $\alpha \in (0, 1)$, every radius $C < \infty$ and every bounded nonlinear function g:*

$$\limsup_{n \to \infty} \sup_{\|h\| \leqslant C n^{-1/2}} P^n_{h_n,g}(\hat{p}_n < \alpha) \leqslant \alpha \tag{8}$$

Outside of exceptional cases, p-values based on parametric outcome models that fit to the full unmatched sample are not robust to local misspecification. In contrast, p-values based on best-performing semiparametric methods (Robins et al., 2008; van der Laan & Rose, 2011) typically achieve guarantees far stronger than (8). We will see that parametric tests gain some of the robustness of semiparametric methods when the data are first pre-processed using matching.

3.3. *Matching protects against local model misspecification*

The first main result of this section says that when $P\{e(X) < 0.5\} = 1$, model-based p-values computed after optimal Mahalanobis matching remain valid, even if the model is locally misspecified.

Theorem 1 (Matching confers local robustness). *Suppose that $P \in H_0$ satisfies Assumption 1, $P\{e(X) < 0.5\} = 1$ and the linear outcome model $E(Y \mid X, Z) = \gamma + \beta^{\mathrm{T}} X$. Let $\hat{p}^{\mathrm{HC}}$ be the one- or two-sided robust standard error p-value for testing the coefficient $\hat{\tau}^{\mathrm{REG}}$ in the regression* (7)*. Then $\hat{p}^{\mathrm{HC}}$ is robust to local misspecification near P in the sense of Definition 2.*

The intuitive explanation for this robustness is that regression after matching combines two complementary methods of bias reduction. The first is the nearly correctly specified outcome model, which eliminates most of the bias and gets us within an $O(n^{-1/2})$ neighbourhood of the correct answer. From there, the nonparametric consistency of matching kicks in to handle the residual nonlinearity. This is conceptually similar to the doubly robust estimator of Robins et al. (1994), which combines an outcome model and a propensity model to gain robustness and efficiency. However, regression after optimal Mahalanobis matching does not produce a consistent estimate of the propensity score, so inferences based on Theorem 1 are not semiparametrically efficient. See Lin et al. (2021) for related discussion.

Unfortunately, the conclusion of Theorem 1 does not extend to populations where some units have propensity scores above 0.5. Pair matching may still improve robustness in such problems, but it will not protect against all directions of local misspecification. The reason is that pair matching runs out of controls in some parts of the covariate space, costing one of the bias reduction methods used in Theorem 1.

Our next result shows that this problem can be avoided by matching with replacement:

$$m_r(i) \in \underset{j:\, Z_j=0}{\arg\min} (X_i - X_j)^{\mathrm{T}}\, \hat{\Sigma}^{-1}(X_i - X_j). \tag{9}$$

To account for the fact that the same control unit may be matched more than once, we also replace the ordinary least-squares regression (7) by a weighted least-squares regression with multiplicity-counting weights.

THEOREM 2 (REPLACING CONTROLS HELPS). *Suppose that $P \in H_0$ satisfies Assumption 1 and the linear outcome model. Let $\mathcal{M}_r$ be the set of units matched by the scheme (9). Let $W_i = 1$ if observation i is treated, and otherwise set $W_i = \sum_{j=1}^{n} Z_j\, \mathbb{I}\{m_r(j) = i\}$. Let $\hat{p}^{\mathrm{HC}}$ be the one- or two-sided robust standard error p-value for testing the coefficient $\hat{\tau}^{\mathrm{REG}}$ in the weighted regression (10):*

$$(\hat{\tau}^{\mathrm{REG}}, \hat{\gamma}, \hat{\beta}) = \underset{(\tau,\gamma,\beta)}{\arg\min} \sum_{i \in \mathcal{M}_r} W_i (Y_i - \gamma - \tau Z_i - \beta^{\mathrm{T}} X_i)^2. \tag{10}$$

Then $\hat{p}^{\mathrm{HC}}$ is robust to local misspecification in the sense of Definition 2.

The reason matching with replacement helps is that it ensures no region of the covariate space will run out of untreated units. Therefore, the bias-correction opportunity from matching is present even when $P\{e(X) < 0.5\} \neq 1$. Based on this result, we generally recommend matching with replacement over pair matching, unless there is good reason to believe that no units have propensity scores greater than 0.5.

The combination of matching with replacement and weighted linear regression was previously implemented by Dehejia & Wahba (2002). In the R programming language, matching with replacement is implemented by default in the `Matching` package due to Sekhon (2011).

3.4. *Matching reduces model dependence*

Finally, we show that matching makes parametric analyses less sensitive to the exact model specification. This provides rigorous support for the main claim in Ho et al. (2007).

Let $P_{h_n,g}$ be a locally misspecified sequence centred around a baseline linear model $P \in H_0$. Thus, when the sample size is n, the true regression model takes the form

$$E_{h_n,g}(Y \mid X, Z) = \gamma + \beta^{\mathrm{T}} X + h_n^{\mathrm{T}} g(X) \tag{11}$$

for some sequence $h_n = O(n^{-1/2})$. We further assume that $\mathrm{var}[\{X, g(X)\} \mid Z = 1] \succ 0$, so that g is genuinely nonlinear in places with treated observations.

Consider the following three different modelling strategies that might be used to analyse the matched data.

(i) Baseline: the first procedure fits a regression model that controls for X linearly; in R and S formula notation, this procedure fits the model $Y \sim 1 + Z + X$.

(ii) Saturated: the second procedure fits a model that correctly includes the nonlinearity in (11), $Y \sim 1 + Z + X + g(X)$.

(iii) Model selector: the final procedure fits the saturated model, drops insignificant components of g and then makes inferences as if the chosen model were prespecified. We make no assumptions on what significance tests are used in the model-pruning step.

After model specification, each procedure produces a p-value based on the heteroskedasticity-consistent standard errors of White (1980) in their chosen models. We denote these by $\hat{p}^{\mathrm{HC1}}, \hat{p}^{\mathrm{HC2}}$ and $\hat{p}^{\mathrm{HC3}}$, respectively.

In the full unmatched dataset, we would expect only the p-value based on the saturated model to perform well. After all, the baseline model is misspecified and the model selector's p-value is rendered irregular by model selection. However, the story is entirely different in the matched sample.

Theorem 3 (Matching reduces model dependence). *Consider the setting above. If unweighted regressions based on pair matching are used, assume that P satisfies the conditions of Theorem 1. If weighted regressions based on matching with replacement are used, assume only that P satisfies the conditions of Theorem 2. Let $\phi^{(k)} = \mathbb{I}\{\hat{p}^{\mathrm{HC}k} < \alpha\}$ denote the accept/reject decision based on $\hat{p}^{\mathrm{HC}k}$. Then*

$$\lim_{n\to\infty} P^n_{h_n,g}(\phi^{(1)} = \phi^{(2)} = \phi^{(3)}) = 1.$$

In other words, all three models yield the same conclusion with high probability.

This phenomenon may be understood as follows. Under the assumptions of Theorem 3, matching is able to balance any function of X, i.e., $\sum_{Z_i=1}\{g(X_i) - g(X_{m(i)})\}/N_1 \to 0$. This makes Z and $g(X)$ approximately orthogonal in the matched sample. From standard least-squares theory we know that the inclusion or exclusion of a nearly orthogonal predictor has very little impact on the other regression coefficients, explaining the similarity of the three p-values. However, this argument would not work for the full data, since there is no reason to expect Z and $g(X)$ to be nearly orthogonal before matching.

Acknowledgement

We are grateful to the referees and associate editor for valuable comments and suggestions. We thank John Cherian, Peng Ding, Colin Fogarty, Isaac Gibbs, Samir Khan, Sam Pimentel, Fredrik Sävje and seminar participants at various conferences for feedback. We especially thank Sky Cao for help with the proof of Lemma B.4.

Supplementary material

The Supplementary Material includes proofs of all the examples, propositions and theorems.

References

Abadie, A. & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–67.

Abadie, A. & Imbens, G. W. (2012). A martingale representation for matching estimators. *J. Am. Statist. Assoc.* **107**, 833–43.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statist. Med.* **27**, 2037–49.

Bai, Y., Romano, J. P. & Shaikh, A. M. (2022). Inference in experiments with matched pairs. *J. Am. Statist. Assoc.* **117**, 1726–37.

Branson, Z. (2021). Randomization tests to assess covariate balance when designing and analyzing matched datasets. *Observ. Stud.* **7**, 1–36.

Carpenter, R. G. (1977). Matching when covariables are normally distributed. *Biometrika* **64**, 299–307.

Dehejia, R. H. & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Statist.* **84**, 151–61.

Ferman, B. (2021). Matching estimators with few treated and many control observations. *J. Economet.* **225**, 295–307.

Hansen, B. B. (2008). The essential role of balance tests in propensity-matched observational studies: Comments on 'A critical appraisal of propensity-score matching in the medial literature between 1996 and 2003' by Peter Austin. *Statist. Med.* **27**, 2050–4.

Hansen, B. B. (2009). Propensity score matching to recover latent experiments: Diagnostics and asymptotics. Tech. Rep. 486, University of Michigan.

Hansen, B. B. & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comp. Graph. Statist.* **15**, 609–27. https://dept.stat.lsa.umich.edu/ bbh/hansen-psm-da-2009-06-08.pdf

Hill, J. (2002). Comment on 'Covariance adjustment in randomized experiments and observational studies' by Paul R. Rosenbaum. *Statist. Sci.* **17**, 304–27.

Ho, D. E., Imai, K., King, G. & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15**, 199–236.

Imai, K., King, G. & Stuart, E. (2007). Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Statist. Soc.* A **171**, 481–502.

Kim, I., Neykov, M., Balakrishnan, S. & Wasserman, L. (2021). Local permutation tests for conditional independence. *arXiv:* 2112.11666v2.

King, G. & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Polit. Anal.* **27**, 435–54.

Leeb, H. & Pötscher, B. M. (2006). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Economet. Theory* **22**, 69–97.

Lin, Z., Ding, P. & Han, F. (2021). Estimation based on nearest neighbor matching: From density ratio to average treatment effect. *arXiv:* 2112.13506.

Pashley, N. E., Basse, G. W. & Miratrix, L. W. (2021). Conditional as-if analyses in randomized experiments. *J. Causal Infer.* **9**, 264–84.

Pimentel, S. D. (2022). Covariate-adaptive randomization inference in matched designs. *arXiv:* 2207.05019.

R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Robins, J., Li, L., Tchetgen, E. & van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, D. Nolan & T. Speed, eds. Beachwood, Ohio: Institute of Mathematical Statistics, pp. 335–421.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression-coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *J. Am. Statist. Assoc.* **84**, 1024–32.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17**, 286–327.

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

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29**, 185–203.

Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. *Biometrics* **32**, 121–32.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Am. Statist. Assoc.* **74**, 318–28.

Rubin, D. B. (1980). Bias reduction using Mahalanobis-metric matching. *Biometrics* **36**, 293–8.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* **47**, 1213–34.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statist. Med.* **26**, 20–36.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Statist.* **2**, 808–40.

Rubin, D. B. (2022). Interview with Don Rubin. *Observ. Stud.* **8**, 77–94.

Sävje, F. (2021). On the inconsistency of matching without replacement. *Biometrika* **109**, 551–8.

Schafer, J. & Kang, J. (2009). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychol. Meth.* **13**, 279–313.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *J. Statist. Software* **42**, 1–52.

Shah, R. D. & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48**, 1514–38.

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

van der Laan, M. J. & Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–38.

Yu, R., Silber, J. H. & Rosenbaum, P. R. (2020). Rejoinder: Matching methods for observational studies derived from large administrative databases. *Statist. Sci.* **35**, 371–4.

[*Received on* 14 *April* 2022*. Editorial decision on* 21 *November* 2022]