Statistica Sinica **33** (2023), 1789-1808
doi:https://doi.org/10.5705/ss.202020.0343

LARGE SAMPLE PROPERTIES OF MATCHING FOR BALANCE

Yixin Wang and José R. Zubizarreta

University of Michigan and Harvard University

Abstract: Matching methods are widely used for causal inference in observational studies. Of these methods, nearest neighbor matching is arguably the most popular. However, nearest neighbor matching does not, in general, yield an average treatment effect estimator that is consistent at the $\sqrt{n}$ rate. Are matching methods not $\sqrt{n}$-consistent in general? In this paper, we examine a recent class of matching methods that use integer programming to directly target aggregate covariate balance, in addition to finding close neighbor matches. We show that under suitable conditions, these methods can yield simple estimators that are $\sqrt{n}$-consistent and asymptotically optimal.

Key words and phrases: Causal inference, integer programming, matching methods, observational studies, propensity score.

1. Introduction

In observational studies, matching methods are widely used for causal inference. The appeal of matching methods lies in the transparency of their covariate adjustments. These adjustments are an interpolation based on the available data, rather than an extrapolation based on a potentially misspecified model (Rubin (1973); Rosenbaum (1989); Abadie and Imbens (2006)). The structure of the data after matching is also simple (often, a self-weighted sample), making statistical inferences and sensitivity analyses straightforward (Rosenbaum (2002, 2010, 2017)). Matching methods are commonly used under the identification assumptions of strong ignorability (Rosenbaum and Rubin (1983)) or selection on observables (Imbens and Wooldridge (2009)), but are also used under the different assumptions required by instrumental variables (e.g., Baiocchi et al. (2010)) and discontinuity designs (e.g., Keele, Titiunik and Zubizarreta (2015)).

Although there is extensive literature on matching methods, large-sample characterizations of matching estimators have centered around nearest neighbor matching only (Abadie and Imbens (2006, 2011)). In its simplest form, this

Corresponding author: José R. Zubizarreta, Departments of Health Care Policy, Biostatistics, and Statistics, Harvard University, Cambridge, MA 02138-2901, USA. E-mail: zubizarreta@hcp.med.harvard.edu.

algorithm matches each treated unit to the closest available control in terms of a covariate distance (e.g., the Mahalanobis distance; Rubin (1973)). In an important paper, Abadie and Imbens (2006) showed that the resulting difference-in-means estimator is not, in general, $\sqrt{n}$-consistent for the average treatment effect when matching with replacement. This estimator contains a bias that decreases at a rate inversely proportional to the number of covariates used for matching. As a result, its convergence can be very slow when matching on many covariates.

Different variants of nearest neighbor matching have been proposed to address this issue. In one variant, Abadie and Imbens (2011) proposed a class of bias-corrected matching estimators, where the missing potential outcomes are imputed using a regression model. This imputation corrects the bias of classical nearest neighbor matching. In another variant, Abadie and Imbens (2016) formalized matching on the estimated propensity score, which reduces the matching space into a single dimension. All of these variants achieve $\sqrt{n}$-consistency. However, in these cases, the faster convergence rate depends on specifying either the treatment or the outcome model correctly, or restricting the outcome model to be Lipschitz continuous on the covariates.

Here, we study a recent class of optimization-based matching methods that directly target aggregate covariate balance, and do not explicitly model the treatment or the outcome (Zubizarreta (2012); Diamond and Sekhon (2013); Nikolaev et al. (2013); Zubizarreta, Paredes and Rosenbaum (2014)). These methods formulate the matching exercise as an integer programming problem. For instance, cardinality matching (Zubizarreta, Paredes and Rosenbaum (2014)) optimizes the number of matched treated and control units, subject to constraints that approximately balance the empirical distributions of the covariates. We show that, under suitable conditions, the resulting difference-in-means treatment effect estimator is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient. These results show that matching for aggregate covariate balance can be asymptotically optimal when nearest neighbor matching with replacement is not. To the best of our knowledge, this is the first work to show that a matching estimator can be semiparametrically efficient under suitable conditions.

To perform this asymptotic analysis of matching for balance, we establish a connection between matching and weighting. We view matching as a form of weighting for covariate balance that weights both the treatment and the control units, and encodes an assignment between them. Examples of weighting methods for covariate balance include Hainmueller (2012), Imai and Ratkovic (2014), Zubizarreta (2015), Chan, Yam and Zhang (2016), Fan et al. (2016),

Zhao and Percival (2017), Athey, Imbens and Wager (2018), Hirshberg and Wager (2021), Zhao (2019), and Wang and Zubizarreta (2020). This connection between matching and weighting enables us to analyze matching for balance using asymptotic techniques developed for weighting.

Despite its connection to weighting, matching for balance retains some essential features of nearest neighbor matching and other optimal covariate distance matching algorithms (e.g., Hansen (2004)). In a similar way to distance matching algorithms, matching for balance can also focus on forming close unit matches, in addition to achieving aggregate covariate balance. In fact, matching for balance can be followed by re-matching for homogeneity to not only preserve aggregate covariate balance in the matched sample, but also to minimize the total covariate distances between its matched units (see Zubizarreta, Paredes and Rosenbaum (2014) for details). As discussed by Rosenbaum (2005) and Visconti and Zubizarreta (2018), re-matching for homogeneity can improve the efficiency and sensitivity of certain matching estimators to unobserved covariates. The main finding of this study is that matching for balance (along with re-matching for homogeneity) can improve the large-sample properties of the classical difference-in-means estimator in causal inference, achieving asymptotic optimality under suitable conditions.

The remainder of this paper is organized as follows. In Section 2, we describe the identification assumptions, matching methods, and the matching estimator. In Section 3, we present and discuss our main results. In Section 4, we evaluate the empirical performance of the estimator. Section 5 concludes the paper. All proofs are provided in the Supplementary Material.

2. Matching for Aggregate Covariate Balance

In this section, we describe the causal estimation problem, and introduce a class of matching methods that target aggregate covariate balance. We use the potential outcomes framework for causal inference (Neyman (1923, 1990); Rubin (1974)). With binary treatments, this framework posits that each unit, indexed by $i = 1, \ldots, N$, has a pair of potential outcomes $\{Y_i(0), Y_i(1)\}$, where $Y_i(1)$ is realized if unit i is assigned to treatment ($Z_i = 1$), and $Y_i(0)$ is realized if the unit is assigned to control ($Z_i = 0$). Thus, for each unit i, we observe either $Y_i(0)$ or $Y_i(1)$, and the observed outcome is expressed as $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. In our setting, the units $i = 1, \ldots, N$ are a random sample from a population of interest and, thus, the potential outcomes are viewed as random variables.

Denote X_i as the vector of observed covariates of unit i. These covariates can

be continuous or discrete. Given these covariates, we assume strong ignorability of the treatment assignment (Rosenbaum and Rubin (1983): $Z_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid X_i$, and $0 < \Pr(Z_i = 1 \mid X_i) < 1$. As implied by our notation, we also require the stable unit treatment value assumption (SUTVA; Rubin (1980)).

The goal is to estimate the average treatment effect (ATE), $\mu = \mathbb{E}[Y_i(1) - Y_i(0)]$. We choose this goal for notational convenience only. For example, our arguments for consistent and efficient estimation of the ATE can be extended directly to the ATT, $\mu_t = \mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = 1]$.

We study matching methods that directly balance the empirical distributions of the observed covariates. Examples of these methods are provided by Zubizarreta (2012), Diamond and Sekhon (2013), and Nikolaev et al. (2013); other related examples include Fogarty et al. (2016), Pimentel et al. (2015), and Kallus (2020). At a high level, these methods aim to balance the covariates, or certain transformations of them, that span a function space (see Wang and Zubizarreta (2020) for a discussion). We call these matching methods *matching for balance*. Extending the formulation in Zubizarreta, Paredes and Rosenbaum (2014), we study the following matching method:

$$\text{max.} \quad M \tag{2.1}$$

$$\text{s.t.} \quad m_{ij} \in \{0,1\}, \qquad i,j = 1,\ldots,n, \tag{2.2}$$

$$\sum_{j=1}^{n}(1 - Z_j)m_{ij} = M, \qquad \forall i \in \{i : Z_i = 1\}, \tag{2.3}$$

$$\sum_{i=1}^{n} Z_i m_{ij} = M, \qquad \forall i \in \{j : Z_j = 0\}, \tag{2.4}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i Z_j m_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n}(1 - Z_i)(1 - Z_j)m_{ij} = 0, \tag{2.5}$$

$$\left| \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{Z_i(1 - Z_j)m_{ij}\{B_k(X_i) - B_k(X_j)\}}{\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i(1 - Z_j)m_{ij}} \right| < \delta_k, , k \in [K] \tag{2.6}$$

$$\left| \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{(1 - Z_i)Z_j m_{ij}\{B_k(X_i) - B_k(X_j)\}}{\sum_{i=1}^{n}\sum_{j=1}^{n}(1 - Z_i)Z_j m_{ij}} \right| < \delta_k, k \in [K], \tag{2.7}$$

where m_{ij} is a binary decision variable that indicates whether unit i is matched to unit j (2.2). Equations (2.3) and (2.4) require each treated unit be matched to M control units, and each control unit be matched to M treated units, respectively. Equation (2.5) ensures that no treated unit is matched to another treated unit,

and that no control unit is matched to another control unit; in other words, only matches between different treatment groups are allowed. Finally, Equations (2.6) and (2.7) ensure that the covariate distributions of the matched treated and control units are balanced. In these constraints, the functions $B_k(\cdot)$ are suitable transformations of the covariates. Each of them maps the multivariate covariate vector X_i into a suitable summary scalar. For example, they can be polynomials or wavelets. Thus, Equations (2.6) and (2.7) constrain the imbalances in these basis functions in the matched sample up to a level δ_k. The constant δ_k is a tuning parameter chosen by the investigator. Zhao (2019) and Wang and Zubizarreta (2020) describe algorithms to automatically select the tuning parameter δ_k in covariate balance optimization problems such as those in (2.1)–(2.7).

On the whole, the optimization problem (2.1)–(2.7) finds the largest matched sample with replacement that is balanced according to the conditions specified in (2.6) and (2.7). An interesting feature of this approach is that it accomplishes the task of matching with replacement without predefining the $1 : M$ matching ratio, instead optimizing M from the data at hand, subject to the aggregate covariate balance constraints. We may posit additional constraints in order to match without replacement: $\sum_{i=1}^{n} Z_i m_{ij} \leq 1, \forall j \in \{j : Z_j = 0\}$. In the asymptotic analyses below, we focus on matching with replacement, but these analyses can be extended to matching without replacement.

Of course, the above optimization problem of matching for balance may be infeasible. For example, this is the case under practical violations of the positivity assumptionor, more specifically, if there is limited overlap in the covariate distributions, as characterized by the functions $B_k(\cdot)$. In this case, one can either base the covariate adjustments on a model that goes beyond the support of the data, or discard some units and possibly change the target of inference (Crump et al. (2009)). In this regard, the infeasibility certificate of the matching for balance problem provides valuable information to characterize the data at hand. In the Supplementary Material, we provide sufficient conditions that guarantee the existence of a solution to the matching for balance optimization problem.

In order to estimate the ATE with matching for balance, we use the following simple difference-in-means estimator:

$$\hat{\mu} := \frac{1}{n}\left[\sum_{i=1}^{n} Z_i \left\{ Y_i - \frac{\sum_{j=1}^{n}(1-Z_j)m_{ij}Y_j}{\sum_{j=1}^{n}(1-Z_j)m_{ij}} \right\} + \sum_{i=1}^{n}(1-Z_i)\left(\frac{\sum_{j=1}^{n} Z_j m_{ij} Y_j}{\sum_{j=1}^{n} Z_j m_{ij}} - Y_i \right)\right]. \tag{2.8}$$

This estimator computes the average difference between each unit and its matches. For example, the first term of (2.8) is the difference between the outcome of each treated unit Y_i and the mean outcome of the units it is matched to, $\{Y_j : m_{ij} = 1, Z_j = 0\}$. Analogously, the second term is the difference between the outcome of each control unit Y_i and the mean outcome of its matches, $\{Y_j : m_{ij} = 1, Z_j = 1\}$.

Using (2.3) and (2.4), we can rewrite this difference-in-means estimator as

$$\hat{\mu} = \frac{1}{n}\left[\left\{\sum_{i=1}^{n} Z_i Y_i + \sum_{j=1}^{n} \frac{\sum_{i=1}^{n}(1-Z_i)m_{ij}}{M} Z_j Y_j\right\} - \left\{\sum_{i=1}^{n}(1-Z_i)Y_i + \sum_{j=1}^{n} \frac{\sum_{i=1}^{n} Z_i m_{ij}}{M}(1-Z_j)Y_j\right\}\right]. \tag{2.9}$$

This form implies that each unit j receives weight $\{1 + \sum_{i=1}^{n}(1-Z_i)m_{ij}/M\}$ if it is treated, and weight $\{1 + \sum_{i=1}^{n} Z_i m_{ij}/M\}$ if it is a control. Using (2.3) and (2.4), we can also rewrite the covariate balance constraints in (2.6) as

$$\frac{1}{\sum_{i=1}^{n} Z_i}\left|\sum_{i=1}^{n} Z_i B_k(X_i) - \sum_{j=1}^{n} \frac{\sum_{i=1}^{n} Z_i m_{ij}}{M}(1-Z_j)B_k(X_j)\right| < \delta_k,$$

$$\frac{1}{\sum_{i=1}^{n}(1-Z_i)}\left|\sum_{i=1}^{n}(1-Z_i)B_k(X_i) - \sum_{j=1}^{n} \frac{\sum_{i=1}^{n}(1-Z_i)m_{ij}}{M} Z_j B_k(X_j)\right| < \delta_k.$$

We observe that the weights of the units in both the constraints and the ATE estimator are functions of the frequencies to which they are matched, namely,

$$w_T(X_j) = \frac{\sum_{i=1}^{n}(1-Z_i)m_{ij}}{M} \quad \text{if } X_j \text{ is treated,} \tag{2.10}$$

$$w_C(X_j) = \frac{\sum_{i=1}^{n} Z_i m_{ij}}{M} \quad \text{if } X_j \text{ is control.} \tag{2.11}$$

Note that the numerator and denominator of the weights $w_T(X_j)$ and $w_C(X_j)$ must be integers, owing to (2.2) and (2.3). This restricts the values that the weights can take. Aside from this constraint, the integer program for matching resembles the convex optimization problem in covariate balancing weights (Zhao (2019); Wang and Zubizarreta (2020)).

This connection between matching and weighting allows us to establish the asymptotic optimality of matching for balance. In the following section, we show that under suitable conditions, the above difference-in-means ATE estimator is

$\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient. As mentioned, we focus on the ATE for simplicity of exposition. These consistency and asymptotic normality results readily extend to the ATT, because the integer programming problem for the ATT is analogous to that of the ATE. The difference is that we match control units to each treated unit, but not treated units to controls. Note that calculating the asymptotic variance of the ATT estimator is more nuanced. In particular, the semiparametric efficiency bound depends on whether we know the true model for the propensity score.

3. Asymptotic Properties of Matching for Balance

In this section, we show that under standard assumptions, matching for balance is asymptotically optimal: the resulting ATE estimator is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient.

We start by describing the assumptions required. We posit three sets of conditions on the basis functions $B(x) = (B_1(x), \ldots, B_K(x))^\top$, the propensity score function $\pi(x) = \Pr(Z_i = 1 \,|\, X_i = x)$, and the mean potential outcome functions $\mathbb{E}[Y_i(z) \,|\, X_i = x]$ for $z \in \{0, 1\}$.

Assumption 1. *Assume the following conditions hold on the basis functions $B(x) = (B_1(x), \ldots, B_K(x))^\top$. There exist constants $C_0, C_1, C_2 > 0$ such that:*

1. $\sup_{x \in \mathcal{X}} ||B(x)||_2 \leq C_0 K^{1/2}$, *where $\mathcal{X}$ is the domain of the covariates X, which is compact.*
2. $||\mathbb{E}[B(X_i)^\top B(X_i)]||_2 \leq C_1$.
3. $\lambda_{\min}\{\mathbb{E}[B(X_i)B(X_i)^\top]\} > C_2$, *where $\lambda_{\min}\{\mathbb{E}[B(X_i)B(X_i)^\top]\}$ denotes the smallest eigenvalue of the matrix $\mathbb{E}[B(X_i)B(X_i)^\top]$.*

Assumptions 1.1–1.3 are standard regularity conditions on the basis functions. They restrict their magnitude by the norm of the length-K basis function vector. These conditions are common in nonparametric sieve estimation (see Assumption 4.1.6 of Fan et al. (2016) and Assumption 2(ii) of Newey (1997)). They are satisfied by many classes of basis functions, including the regression spline, trigonometric polynomial, and wavelet bases (Newey (1997); Horowitz and Mammen (2004); Chen (2007); Belloni et al. (2015); Fan et al. (2016)).

Assumption 2. *Assume the following conditions hold on the propensity score function $\pi(x) = \Pr(Z_i = 1 \,|\, X_i = x)$:*

1. *There exists a constant $C_3 > 0$ such that $C_3 < \pi(x) < 1 - C_3$.*

2. *There exist vectors* $(\lambda_{1T}^*)_{K\times 1}, (\lambda_{1C}^*)_{K\times 1}$ *such that the true propensity score function* $\pi(\cdot)$ *satisfies* $\sup_{x\in\mathcal{X}} |1/\pi(x) - B(x)^\top \lambda_{1T}^*| = O(K^{-r_\pi})$ *and* $\sup_{x\in\mathcal{X}} |1/\{1-\pi(x)\} - B(x)^\top \lambda_{1C}^*| = O(K^{-r_\pi})$, *where* $r_\pi > 1$.

3. *There exists a set* $\mathcal{M}$ *such that the propensity score function satisfies* $1/\pi(x)$ $\mathcal{M}$ *and* $1/\{1-\pi(x)\} \in \mathcal{M}$. *Moreover, the set* $\mathcal{M}$ *is a set of smooth functions such that* $\log n_{[]}\{\varepsilon, \mathcal{M}, L_2(\Pr)\} \leq C_4(1/\varepsilon)^{1/k_1}$, *where* $n_{[]}\{\varepsilon, \mathcal{M}, L_2(\Pr)\}$ *denotes the covering number of* $\mathcal{M}$ *by* ε*-brackets,* $L_2(\Pr)$ *is the norm defined as* $||m_1(\cdot) - m_2(\cdot)||_{L_2(\Pr)} = \mathbb{E}[m_1(X_i) - m_2(X_i)]^2$, C_4 *is a positive constant, and* $k_1 > 1/2$.

Assumption 2.1 requires overlap between the treatment and the control populations. This is part of the identification assumption described in Section 2. Assumption 2.2 is a smoothness condition on the inverse propensity score function. It requires that the inverse propensity score be uniformly approximable by the basis functions $B(x) = (B_1(x), \ldots, B_K(x))^\top$. For example, when we choose the basis functions to be multidimensional splines or power series, this assumption is satisfied for $r_\pi = s/d$, where s is the number of continuous derivatives of $1/\pi(x)$, and d is the dimension of x, for x with a compact domain $\mathcal{X}$ (Newey (1997)). Assumption 2.3 constrains the complexity of the function class to which the inverse propensity score function belongs. This assumption is satisfied, for example, by the Hölder class with smoothness parameter s defined on a bounded convex subset of $\mathbb{R}^d$, with $s/d > 1$ (van Der Vaart and Wellner (1996); Fan et al. (2016)). This is a key assumption that enables the use of empirical process techniques when establishing consistency and asymptotic normality.

Assumption 3. *Assume the following conditions hold on the mean potential outcome functions* $Y_z(x) \stackrel{\Delta}{=} \mathbb{E}[Y_i(z) \mid X_i = x]$ *for* $z \in \{0, 1\}$*:*

1. $\mathbb{E}|Y_i - Y_0(X_i)| < \infty$ *and* $\mathbb{E}|Y_i - Y_1(X_i)| < \infty$.

2. $|\mu| < \infty$, *where* $\mu = \mathbb{E}[Y_i(1) - Y_i(0)]$ *is the true average treatment effect.*

3. *There exist* $r_y > 1/2$, $(\lambda_{2C}^*)_{K\times 1}$, *and* $(\lambda_{2T}^*)_{K\times 1}$ *such that* $\sup_{x\in\mathcal{X}} |Y_0(x) - B(x)^\top \lambda_{2C}^*| = O(K^{-r_y})$ *and* $\sup_{x\in\mathcal{X}} |Y_1(x) - B(x)^\top \lambda_{2T}^*| = O(K^{-r_y})$.

4. *The potential outcome functions satisfy* $Y_0(\cdot) \in \mathcal{H}$ *and* $Y_1(\cdot) \in \mathcal{H}$, *where* $\mathcal{H}$ *is a set of smooth functions satisfying* $\log n_{[]}\{\varepsilon, \mathcal{H}, L_2(\Pr)\} \leq C_5(1/\varepsilon)^{1/k_2}$, C_5 *is a positive constant, and* $k_2 > 1/2$. *As in Assumption* 2.3, $n_{[]}\{\varepsilon, \mathcal{H}, L_2(\Pr)\}$ *denotes the covering number of* $\mathcal{H}$ *by* ε*-brackets, and* $L_2(\Pr)$ *is the norm* $||m_1(\cdot) - m_2(\cdot)||_{L_2(\Pr)} = \mathbb{E}[m_1(X_i) - m_2(X_i)]^2$.

Assumptions 3.1 and 3.2 are regularity conditions on the mean potential outcomes. Assumptions 3.3 and 3.4 are analogous to Assumptions 2.2 and 2.3; they constrain the smoothness of the mean potential outcome functions and the complexity of the function class to which they belong. Under strong ignorability, one may get a rough sense of the function approximation quality (Assumption 2.3) by evaluating the prediction error of a fitted outcome model on a holdout sample. This model explains the observed outcomes in terms of the K basis functions of the observed covariates plus the treatment assignment indicator. The approximation is likely to be good if the prediction error is small on the holdout data. That said, such an empirical evaluation provides only a rough sense of the quality of the approximation, because we have only finite samples. Note that while no specific modeling assumptions are required for the inverse propensity score function and the potential outcome functions, Assumptions 2.2 and 2.3 do require that both have the same form of smoothness, namely, that they can all be well approximated by the same set of basis functions.

Assumption 4. *Assume the following conditions on the matching for balance problem:*

1. $K = o(n^{1/2})$.
2. $||\delta||_2 = O_p[K^{1/2}\{(\log K)/n + K^{-r_\pi}\}]$, *where* $\delta = (\delta_1, \ldots, \delta_K)$.
3. $n^{1/\{2(r_\pi + r_y - 0.5)\}} = o(K)$, *where* r_π *and* r_y *are the smoothness parameters defined in Assumptions* 2.2 *and* 3.3.

Assumption 4.1 quantifies the rate at which the number of basis functions we balance can grow with the number of units. Assumption 4.2 limits the extent to which there can be imbalances in the basis functions. Despite these imbalances, we show that the optimal large-sample properties of the matching estimator are maintained. Assumption 4.3 characterizes the growth rates of K and n with respect to the uniform approximation rates r_π and r_y.

Now, we state the main result of this paper.

Theorem 1. *Under Assumptions* 1, 4, 3, 2, *the ATE estimator*

$$\hat{\mu} := \frac{1}{n}\left[\sum_{i=1}^{n} Z_i \left\{ Y_i - \frac{\sum_{j=1}^{m}(1 - Z_j) m_{ij} Y_j}{\sum_{j=1}^{m}(1 - Z_j) m_{ij}} \right\} + \sum_{i=1}^{n}(1 - Z_i)\left\{ \frac{\sum_{j=1}^{m} Z_j m_{ij} Y_j}{\sum_{j=1}^{m} Z_j m_{ij}} - Y_i \right\}\right]$$

is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient:

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, V_{opt}),$$

where V_{opt} is equal to the semiparametric efficiency bound

$$V_{opt} = \mathbb{E}\left[\frac{Var[Y_i(1) \mid X_i]}{\pi(X_i)} + \frac{Var[Y_i(0) \mid X_i]}{1-\pi(X_i)} + \{\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i] - \mu\}^2\right],$$

and $\pi(X_i)$ is the propensity score of unit i. If, in addition, $r_y > 1$ holds, then the estimator

$$\begin{aligned}\hat{V}_K = \sum_{i=1}^{n} \Bigg[& Z_i w(X_i) Y_i - \frac{\sum_{i=1}^n Z_i w(X_i) Y_i}{\sum_{i=1}^n Z_i} \\ & - (1 - Z_i) w(X_i) Y_i + \frac{\sum_{i=1}^n (1 - Z_i) w(X_i) Y_i}{\sum_{i=1}^n (1 - Z_i)} \\ & - \hat{Y}_T(X_i) \times \left\{ Z_i w(X_i) - \frac{1}{\sum_{i=1}^n Z_i} \right\} \\ & + \hat{Y}_C(X_i) \times \left\{ (1 - Z_i) w(X_i) - \frac{1}{\sum_{i=1}^n (1 - Z_i)} \right\} \Bigg]^2\end{aligned}$$

is a consistent estimator of the asymptotic variance V_{opt}, where

$$\hat{Y}_T(X_i) = B(X_i)^\top \left\{ \frac{\sum_{i=1}^n Z_i w(X_i) B(X_i)^\top B(X_i)}{\sum_{i=1}^n Z_i} \right\}^{-1} \cdot \left\{ \frac{\sum_{i=1}^n Z_i w(X_i) B(X_i) Y_i}{\sum_{i=1}^n Z_i} \right\}$$

$$\begin{aligned}\hat{Y}_C(X_i) = B(X_i)^\top & \left\{ \frac{\sum_{i=1}^n (1 - Z_i) w(X_i) B(X_i)^\top B(X_i)}{\sum_{i=1}^n (1 - Z_i)} \right\}^{-1} \\ & \cdot \left\{ \frac{\sum_{i=1}^n (1 - Z_i) w(X_i) B(X_i) Y_i}{\sum_{i=1}^n (1 - Z_i)} \right\}.\end{aligned}$$

Proof. The proof uses empirical process techniques to analyze the ATE estimators, as in Fan et al. (2016) (see also Wang and Zubizarreta (2020)). The key challenge in this proof lies in the need to characterize the entire class of matching solutions in matching for balance. More specifically, the optimization objective of matching for balance does not involve the matching solution m_{ij} directly, so it does not correspond to a unique matching solution. Hence, we need to study the ATE estimates resulting from all possible matching solutions. In contrast, the balancing weights (Wang and Zubizarreta (2020)) and the covariate balancing propensity score (Fan et al. (2016)) both work with optimization objectives that involve all the weights; these problems also assume a unique weighting solution

with infinite samples.

The proof starts by showing that the implied weights of matching for balance (2.10) approximate the true inverse propensity score function $\pi(x)^{-1}$. Moreover, this approximation is consistent, owing to the balancing constraints (Equations (2.6)–(2.7)). The rest of the proof involves a decomposition of $\hat{\mu} - \mu$ into seven components, where six of them converge to zero in probability, and the other is asymptotically normal and semiparametrically efficient. Each of the first six components can be controlled by the bracketing number of the function classes of the inverse propensity score and the outcome functions. Assumption 2.3 and 3.4 provide this control. The full proof is provided in Section A of the Supplementary Material.

An intuitive explanation of Theorem 1 relies on two observations. The first observation is that the ATE is an estimand derived from the entire population, rather than from individual units. The asymptotic optimality of our ATE estimator depends primarily on whether the covariate distribution of the treated units is close in aggregate to that of the control units. For this type of estimator, how the individual units are matched to each other plays a secondary role. More specifically, the ATE estimator depends only on the number of times each treated (or control) unit is matched. Thus, an aggregate covariate balance is sufficient for the asymptotic optimality of the matching estimators for the ATE. Matching for balance precisely targets this aggregate covariate balance. Equations (2.6) and (2.7) ensure that the covariate distributions after matching are balanced in aggregate for the treated and control units.

The second observation is the connection between matching for balance and covariate balancing weights (Hainmueller (2012); Imai and Ratkovic (2014); Zubizarreta (2015); Chan, Yam and Zhang (2016); Fan et al. (2016); Zhao and Percival (2017); Zhao (2019); Wang and Zubizarreta (2020)). Both methods formulate the estimation problem as a mathematical program under similar covariate balancing constraints to those in (2.6) and (2.7). Covariate balancing weights have been shown to be asymptotically optimal. Thus, if matching for balance admits a solution, its implied weights, as in (2.10), are as good as the covariate balancing weights. For this reason, under the conditions required by Assumptions 1, 2, 3, 4, matching for balance can also be asymptotically optimal for the difference-in-means estimators for the ATE. When these assumptions do not hold, the nearest neighbor matching estimator for the ATE (which is a competing semiparametric estimator) does not similarly achieve the semiparametric efficiency bound.

We conclude this section with a discussion of Theorem 1 and its assumptions. Unlike other matching methods that assume a correct propensity or outcome

model, Theorem 1 studies matching for balance that posits explicit conditions on covariate balance. Such conditions are practically appealing because covariate balance is typically what is checked in practice. Other regularity conditions and smoothness conditions are standard in nonparametric sieve estimation.

Under these conditions, Theorem 1 establishes the asymptotic optimality of the simple difference-in-means estimator for the ATE, after matching for aggregate covariate balance. In practice, these conditions indicate the following: (i) using basis functions, such as power series or wavelets (Assumption 1); (ii) considering settings with a smooth inverse propensity score and potential outcome functions, with more continuous derivatives than the number of covariates (Assumptions 2 and 3); and (iii) restricting the number of balancing basis functions K and the approximate balance tolerance δ to scale appropriately with the number of samples (roughly, $K = O(n^{1/2-\epsilon})$ and $\delta = O[K^{1/2}\{(\log K)/n + K^{-r_\pi}\}]$, where $\epsilon > 0$ is a small number (Assumption 4). Although Assumptions 4 of Theorem 1 states that one may balance up to $K = O(n^{1/2-\epsilon})$ basis functions as the sample size goes to infinity, in practice, one should proceed with caution in any given finite sample. The matching for balance optimization problem may not admit a solution. Even if a solution exists, the finite-sample performance of the resulting estimator may not be ideal. As Robins and Ritov (1997) suggested, nonparametric estimators may suffer from the curse of dimensionality. A nonparametric estimator with good finite-sample performance may not exist without knowledge of the true propensity score.

Note that Abadie and Imbens (2011) also devise a matching estimator that is consistent at the $\sqrt{n}$-rate, but matching for balance achieves the $\sqrt{n}$-rate in a different way. Abadie and Imbens (2011) correct the bias in nearest neighbor matching by positing a consistent regression model for the mean potential outcome function. In contrast, matching for balance avoids this bias by directly balancing the observed covariates in aggregate. Balancing covariates in aggregate has been shown to be equivalent to nonparametric estimation of the inverse propensity score and mean potential outcome functions (Fan et al. (2016); Zhao and Percival (2017); Hirshberg and Wager (2021); Zhao (2019); Wang and Zubizarreta (2020)). This nonparametric approach relieves us from positing a model for the mean potential outcome function that needs to be specified correctly. Although Theorem 1 requires certain conditions on both the propensity score and the potential outcome functions, it shows that the matching for balance estimator can achieve semiparametric efficiency beyond $\sqrt{n}$-consistency.

Finally, Abadie and Imbens (2012) provide a martingale representation of a widespread nearest neighbor matching estimator and derive its asymptotic distri-

bution. They decompose the estimator into a martingale term and a conditional bias term. Both their and our analysis require that the conditional bias term vanish in order to achieve asymptotic consistency. Specifically, Abadie and Imbens (2012) posit regularity conditions under which the conditional bias term converges in probability to zero. Theorem 1 uses the covariate balance conditions in (2.6) and (2.7) to ensure that the conditional bias term vanishes.

4. Simulation Study

Here, we illustrate the empirical performance of matching for balance. Our simulation study is based on a real data set about the importance of market access for economic development (Redding and Sturm (2008)). The covariates in this data set are non-Gaussian, and cannot be characterized by their first two moments. We focus on a setting in which both the propensity score and the outcome are nonlinear functions of the covariates, and study the MSE and coverage probabilities of matching for balance.

To generate the data, we take the actual covariate values from Redding and Sturm (2008) and simulate the treatment and outcome values as follows. To simulate the treatment assignment indicator Z_i, we first fit a logistic regression model to the original indicator in the data set. Specifically, we fit the model $\Pr(Z_i = 1 \,|\, X_i) = \text{sigmoid}\{(\alpha + \sum_{p=1}^{P} \beta_j X_{ip} + \sum_{p,p'=1}^{P} \beta_{pp'} X_{ip} X_{ip'})\}$, where X_{ip} denotes the pth observed covariate of unit i. We zero out the estimated coefficients with p-values smaller than 0.25, and retain the rest of the coefficients. Finally, we generate the treatment assignment indicator Z_i for the simulated data set using a thresholding model $Z_i = \mathbf{1}\{Z_i^* > 0\}$, where $Z_i^* = (\alpha + \sum_{p=1}^{P} \beta_j X_{ip} + \sum_{p,p'=1}^{P} \beta_{pp'} X_{ip} X_{ip'})/c + \text{Unif}(-0.5, 0.5)$, setting $c = 50$ to induce limited overlap.

Next, we simulate the potential outcomes $\{Y_i(0), Y_i(1)\}$. Again, we begin by fitting a linear regression model with all possible second-order interaction terms to the original treated and control outcomes in the sample. The model is $Y_i = \alpha' + \sum_{p=1}^{P} \beta_j' X_{ip} + \sum_{p,p'=1}^{P} \beta_{pp'}' X_{ip} X_{ip'} + \beta_t' T_i + \epsilon_i$, $\epsilon_i \sim N(0, 1)$. As with the treatment assignment, we zero out all estimated coefficients with p-values smaller than 0.25, and predict the potential treated and control outcomes on the entire sample using the fitted model. We then generate the observed outcomes using $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. This data-generating process yields a simulated data set with the same size (122 units) as the original data set in Redding and Sturm (2008) and outcomes determined by three covariates and four associated relevant second-order terms. Thus, if we balance up to the kth univariate moment of the covariates, we need to balance $3 \cdot k$ basis functions.

Using these simulated data sets, we evaluate the MSE of matching for balance in estimating the ATE. We vary the number of basis functions that we balance by setting the bases to be the moments of the covariates and increasing their order. We balance the first, second, and third moments of the covariates. We set the level of imbalance to be 0.1 standard deviations of the corresponding moment.

Figure 1 shows that as we increase the number of balancing basis functions, the MSE of matching for balance decreases. Furthermore, because the simulated data set has limited overlap, matching for balance achieves a lower MSE than the standard augmented inverse propensity weighted estimator (AIPW; Robins, Rotnitzky and Zhao (1994)), because the approximate balance constraints trade variance for bias. Nonetheless, when covariate distributions have limited overlap, this improvement in MSE comes at a cost. As we show below, the resulting confidence intervals may exhibit lower than nominal coverage, owing to the approximate covariate balance. This suggests exploring separate imbalance tolerances for estimation and for inference. Finally, because of the nonlinearity of the inverse propensity score, including third order-basis functions improves the MSE, despite the data generating-process only involving second-order terms.

Figure 1 also corroborates the discussion in Section 3 that matching for balance may not admit a solution if we aim to balance too many basis functions. For example, in Figure 1, matching for balance admits a solution if we balance the first three moments of the covariates, but not if we also try to balance the fourth moment for the given tolerance level of covariate imbalance. Note that in such cases, $\sqrt{n}$-consistency may not hold, because the asymptotic results derived in Section 3 only apply when $K = o(n^{1/2})$. These asymptotic results do not apply if matching for balance does not admit a solution when K increases in this order and not enough matches can be found.

Figure 2a shows that as we increase the number of balancing basis functions, the average number of matches M decreases. The reason is that balancing more basis functions implies solving a more constrained optimization problem; hence, the average M decreases. A similar phenomenon appears in Figure 2b, where the average standardized covariate balance

$$\frac{1}{K}\sum_{k=1}^{K}\left[\left|\sum_{i=1}^{n}\sum_{p=1}^{n}\frac{Z_i(1-Z_j)m_{ij}\{B_k(X_i)-B_k(X_j)\}}{\sum_{i=1}^{n}\sum_{p=1}^{n}Z_i(1-Z_j)m_{ij}}\right|/\mathrm{sd}\{B_k(X_i)\}\right]$$

increases with the number of balancing basis functions. (In Figure 2b, the order of the balancing basis functions equal to zero represents covariate balance before matching.) Both metrics (the average number of matches and the average

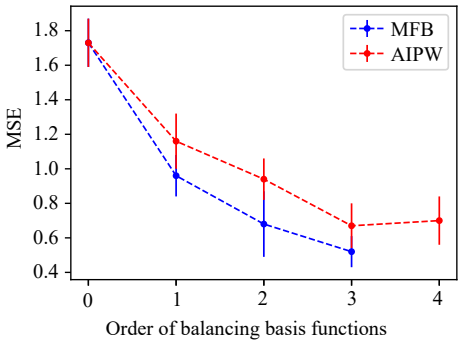
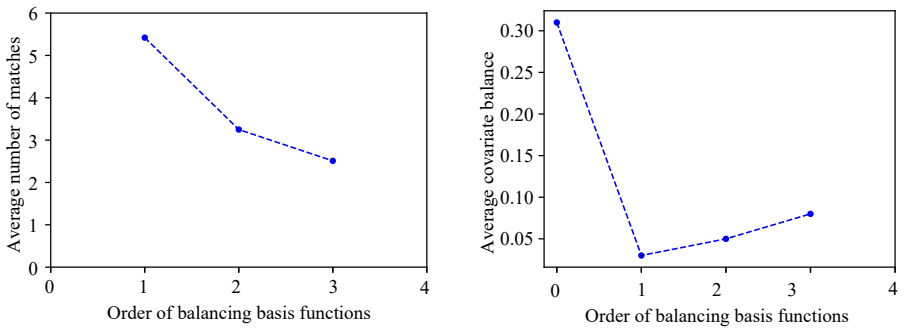


Figure 1. The matching for balance (MFB) estimator achieves a lower MSE than the augmented inverse probability weighting (AIPW) estimator when the data have limited overlap. Increasing the number of balancing basis functions improves the MSE of MFB until its optimization problem becomes infeasible. The bars indicate ±1 standard deviation across 100 simulations.

absolute standardized mean difference in covariates) decrease as we increase the number of basis functions that we balance. However, the MSE still improves, because the treatment and control groups after matching are more similar in ways that are relevant to the propensity score and outcome models. This illustrates the importance of balancing more basis functions than just means when the propensity score and outcome models are non-linear on the covariates. Although balancing a high number of basis functions can be difficult with most matching methods (because they do not directly target covariate balance), with matching for balance, covariate balance on the basis functions is obtained by construction. Subject to these covariate balance requirements, the matching ratio M in matching for balance is optimized, resulting in the largest possible $1 : M/M : 1$ matching ratio for the data at hand.

Finally, we evaluate the coverage probabilities of matching for balance and AIPW. We focus on balancing the first two moments of the covariates, and vary the imbalance tolerance δ such that $\delta_k = \delta \cdot \text{sd}\{B_k(X_i)\}$. In Figure 3, we show the coverage probabilities of the 95% confidence intervals constructed based on Theorem 1. The figure shows that when the imbalance tolerance is small ($\delta \leq 0.01$), the confidence intervals have close to nominal coverage. As the imbalance tolerance increases, the average number of matches increases, but the coverage probability degrades as the matched sample exhibits worse balance. In contrast, the AIPW estimator achieves close to nominal coverage. In such cases, matching for balance trades the imbalance tolerance δ for the matching ratio M in order to exchange bias for variance and achieve a lower MSE, but it can compromise



(a) Average number of matches. (b) Average covariate balance.

Figure 2. Average number of matches and average standardized covariate balance of matching for balance. Although both measures decrease as we balance more basis functions, the MSE still improves, because the covariate distributions of the treatment and control groups are more similar in ways that are relevant to the propensity score and outcome models.

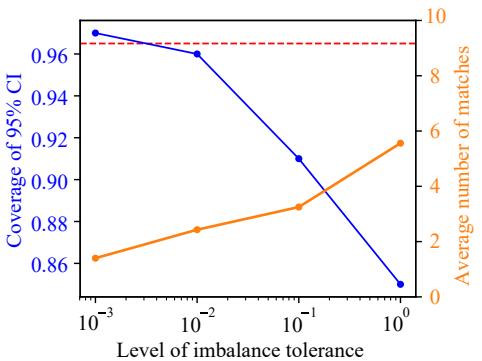


Figure 3. The confidence intervals of matching for balance achieve nominal coverage when the level of imbalance tolerance δ is small. Increasing δ trades variance for bias, but can degrade the coverage probabilities of the corresponding confidence intervals. In contrast, AIPW (the dashed red line) achieves close to nomimal coverage.

coverage probabilities.

An interesting direction for future research is to use a larger value of δ for estimation, and a smaller value for inference, such that confidence intervals are not necessarily centered at the point estimate. This is analogous to what is sometimes done in analyses of regression discontinuity designs (Calonico, Cattaneo and Titiunik (2014)). In general, how to select the parameter δ is an open question in causal inference. In the context of weighting, two proposals are provided by Zhao (2019) and Wang and Zubizarreta (2020), where δ is selected to generalize

covariate balance across cross-validation and bootstrap samples, respectively.

5. Concluding Remarks

We have analyzed a recent class of matching methods that targets aggregate covariate balance. After all, covariate balance is the main diagnostic that investigators carry out in practice. As discussed, matching for balance does not preclude finding close unit matches, because it can be followed by matching for homogeneity in order to minimize covariate distances between matched units, while preserving aggregate covariate balance (see Zubizarreta, Paredes and Rosenbaum (2014)).

Under suitable conditions, we have shown that this class of matching methods yields a simple difference-in-means estimator that is asymptotically optimal: it is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient. As discussed in the simulation study section, these conditions can be stringent in practice, because they require that the imbalance tolerance δ decreases as the sample size increases and there needs to exist a matching solution for such values of δ. These results complement the fundamental results of Abadie and Imbens (2006), who showed that a similar estimator is not, in general, $\sqrt{n}$-consistent for nearest neighbor matching with replacement.

Matching for balance exemplifies how tools from modern optimization (e.g., Jünger et al. (2009) and Bixby (2012)) can play a central role in the design of observational studies in general (e.g., Rosenbaum (2002) and Imbens and Rubin (2015)) and in matching for covariate balance in particular (e.g., Zubizarreta et al. (2013) and Keele, Titiunik and Zubizarreta (2015)). A natural future research direction is to augment matching for balance as in doubly robust estimators (Robins, Rotnitzky and Zhao (1994); see also Rubin (1979); Abadie and Imbens (2011); Athey, Imbens and Wager (2018) for related approaches). Another promising direction is to build on the methods described in Rosenbaum (2017) on evidence factors and sensitivity analysis for interpretable analyses of matched observational studies.

Acknowledgments

We thank Ambarish Chattopadhyay, Eric Cohn, Bijan Niknam, the editors, and three anonymous reviewers for their helpful comments and suggestions. This work was supported through a Patient-Centered Outcomes Research Institute (PCORI) Project Program Award (ME-2019C1-16172) and grants from the Alfred P. Sloan Foundation (G-2018-10118, G-2020-13946), the Office of Naval Re-

search, and the National Science Foundation (NSF-CHE-2231174).

References

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

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *J. Bus. Econom. Statist.* **29**, 1–11.

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *J. Amer. Statist. Assoc.* **107**, 833–843.

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

Athey, S., Imbens, G. W. and Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **80**, 597–623.

Baiocchi, M., Small, D. S., Lorch, S. and Rosenbaum, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105**, 1285–1296.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186**, 345–366.

Bixby, R. E. (2012). A brief history of linear and mixed-integer programming computation. *Documenta Mathematica*, 107–121.

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* **82**, 2295–2326.

Chan, K. C. G., Yam, S. C. P. and Zhang, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **78**, 673–700.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* **6**, 5549–5632.

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

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Stat.* **95**, 932–945.

Fan, J., Imai, K., Liu, H., Ning, Y. and Yang, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. Technical report.

Fogarty, C. B., Mikkelsen, M. E., Gaieski, D. F. and Small, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Amer. Statist. Assoc.* **111**, 447–458.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20**, 25–46.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99**, 609–618.

Hirshberg, D. A. and Wager, S. (2021). Augmented minimax linear estimation. *Ann. Statist.* **49**, 3206–3227.

Horowitz, J. L. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *Ann. Statist.* **32**, 2412–2443.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76**, 243–263.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, New York.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* **47**, 5–86.

Jünger, M., Liebling, T. M., Naddef, D., Nemhauser, G. L., Pulleyblank, W. R., Reinelt, G. et al. (2009). *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*. Springer-Verlag, Berlin.

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.* **21**, 1–54.

Keele, L., Titiunik, R. and Zubizarreta, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J. R. Statist. Soc. Ser. A* **178**, 223–239.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79**, 147–168.

Neyman, J. (1923, 1990). On the application of probability theory to agricultural experiments. *Statist. Sci.* **5**, 465–472.

Nikolaev, A. G., Jacobson, S. H., Cho, W. K. T., Sauppe, J. J. and Sewell, E. C. (2013). Balance optimization subset selection: An alternative approach for causal inference with observational data. *Oper. Res.* **61**, 398–412.

Pimentel, S. D., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Amer. Statist. Assoc.* **110**, 515–527.

Redding, S. J. and Sturm, D. M. (2008). The costs of remoteness: Evidence from German division and reunification. *Am. Econ. Rev.* **98**, 1766–97.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Stat. Med.* **16**, 285–319.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846–866.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist. Assoc.* **84**, 1024–1032.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer-Verlag, New York.

Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59**, 147–152.

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

Rosenbaum, P. R. (2017). *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press, Cambridge.

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

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74**, 318–328.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75**, 591–593.

van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak Convergence and Empirical Processes* (Edited by A. W. van der Vaart and J. A. Wellner), 16–28. Springer, New York.

Visconti, G. and Zubizarreta, J. R. (2018). Handling limited overlap in observational studies with cardinality matching. *Obs. Stud.* **4**, 217–249.

Wang, Y. and Zubizarreta, J. R. (2020). Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. *Biometrika* **107**, 93–105.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Ann. Statist.* **47**, 965–993.

Zhao, Q. and Percival, D. (2017). Entropy balancing is doubly robust. *J. Causal Inference* **5**.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107**, 1360–1371.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110**, 910–922.

Zubizarreta, J. R., Paredes, R. D. and Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Appl. Stat.* **8**, 204–231.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S. A. and Rosenbaum, P. R. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Ann. Appl. Stat.* **7**, 25–50.

Yixin Wang

Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, USA.

E-mail: yixin.wang@columbia.edu

José R. Zubizarreta

Departments of Health Care Policy, Biostatistics, and Statistics, Harvard University, Cambridge, MA 02138-2901, USA.

E-mail: zubizarreta@hcp.med.harvard.edu

(Received September 2019; accepted September 2021)