Biometrika (2020), **107**, 1, *pp.* 93–105
Printed in Great Britain
doi: 10.1093/biomet/asz050
Advance Access publication 22 October 2019

Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations

By YIXIN WANG

Department of Statistics, Columbia University, 1255 Amsterdam Ave, New York, New York 10027, U.S.A.
yixin.wang@columbia.edu

and JOSE R. ZUBIZARRETA

Department of Health Care Policy, Harvard University, 180 Longwood Avenue, Boston, Massachusetts 02115, U.S.A.
zubizarreta@hcp.med.harvard.edu

Summary

Weighting methods are widely used to adjust for covariates in observational studies, sample surveys, and regression settings. In this paper, we study a class of recently proposed weighting methods, which find the weights of minimum dispersion that approximately balance the covariates. We call these weights 'minimal weights' and study them under a common optimization framework. Our key observation is that finding weights which achieve approximate covariate balance is equivalent to performing shrinkage estimation of the inverse propensity score. This connection leads to both theoretical and practical developments. From a theoretical standpoint, we characterize the asymptotic properties of minimal weights and show that, under standard smoothness conditions on the propensity score function, minimal weights are consistent estimates of the true inverse probability weights. In addition, we show that the resulting weighting estimator is consistent, asymptotically normal and semiparametrically efficient. From a practical standpoint, we give a finite-sample oracle inequality that bounds the loss incurred by balancing more functions of the covariates than strictly needed. This inequality shows that minimal weights implicitly bound the number of active covariate balance constraints. Finally, we provide a tuning algorithm for choosing the degree of approximate balance in minimal weights. The paper concludes with an empirical study which suggests that approximate balance is preferable to exact balance, especially when there is limited overlap in covariate distributions. Further studies show that the root mean squared error of the weighting estimator can be reduced by as much as a half with approximate balance.

Some key words: Causal inference; Missing data; Observational study; Sample survey; Weighting.

1. Introduction

1.1. *Weighting methods for covariate adjustment*

Weighting methods are widely used to adjust for observed covariates, for example in observational studies of causal effects (Rosenbaum, 1987), in sample surveys and panel data with unit nonresponse (Robins et al., 1994), and in regression settings with missing and/or mismeasured

covariates (Hirano et al., 2003). Weighting methods are popular because they do not require explicit modelling of the outcome (Rosenbaum, 1987). As a result, they are part of the design stage as opposed to the analysis stage of the study (Rubin, 2008), which helps to maintain the objectivity of the study and preserve the validity of its tests (Rosenbaum, 2010). Furthermore, weighting methods are considered to be multipurpose in the sense that one set of weights can be used to estimate the mean of multiple outcomes (Little & Rubin, 2014).

Conventionally, the weights are estimated by modelling the propensities of receiving treatment or exhibiting missingness and then inverting the predicted propensities. However, with this approach it can be difficult to properly adjust for or balance the observed covariates, as the covariates are balanced only in expectation by the law of large numbers. In any particular dataset it may be difficult to balance covariates, especially if the dataset is small or the covariates are sparse (Zubizarreta et al., 2011). In addition, this approach can result in very unstable estimates when a few observations have very large weights (e.g., Kang & Schafer, 2007). To address these problems, a number of methods have been proposed recently. Instead of explicitly modelling the propensities of treatment or missingness, these methods directly balance the covariates. Some of them also minimize a measure of dispersion of the weights. Examples include the methods of Hainmueller (2012), Zubizarreta (2015), Chan et al. (2016), Zhao & Percival (2017), Wong & Chan (2018) and Zhao (2019). Earlier and related methods include those of Deville & Särndal (1992), Hellerstein & Imbens (1999), Imai & Ratkovic (2014) and Li et al. (2018). Two promising approaches that use similar weights together with outcome information are the ones proposed by Athey et al. (2018) and Hirshberg & Wager (2019). See Yiu & Su (2018) for a framework for constructing weights such that the association between the covariates and the treatment assignment is eliminated after weighting.

Most of these weighting methods balance covariates exactly rather than approximately. This is a subtle but important difference, because approximate balance can trade bias for variance, whereas exact balance cannot. Also, exact balance may not admit a solution while approximate balance may do so. For a fixed sample size, approximate balance may balance more functions of the covariates than exact balance.

In this paper, we study the class of weights of minimum dispersion that approximately balance the covariates. We call these weights minimal dispersion approximately balancing weights, or simply minimal weights. While it has been shown that instances of minimal weights work well in practice in both low- and high-dimensional settings (e.g., Zubizarreta, 2015; Athey et al., 2018; Hirshberg & Wager, 2019), and valuable theoretical results have been established (e.g., Athey et al., 2018; Hirshberg & Wager, 2019; Wong & Chan, 2018), important aspects of their theoretical properties and practical usage remain to be studied.

1.2. *Theoretical properties and practical considerations of minimal weights*

In this paper we study the class of minimal weights. The key observation is the connection between approximate covariate balance and shrinkage estimation of the inverse propensity score. This connection leads to both theoretical and practical developments.

From a theoretical standpoint, we first establish a connection between minimal weights and shrinkage estimation of the propensity score. We show that the dual of the minimal weights optimization problem is similar to parameter estimation in generalized linear models under ℓ_1 regularization. This connection allows us to establish the asymptotic properties of minimal weights by leveraging results on propensity score estimation. In particular, we show that under standard smoothness conditions, minimal weights are consistent estimates of the true inverse probability weights in both the ℓ_2- and the ℓ_∞-norms.

Next, we study the asymptotic properties of a linear estimator based on minimal weights. We show that the weighting estimator is consistent, asymptotically normal and semiparametrically efficient. This result is related to the work of Chan et al. (2016), Fan et al. (2016), Zhao & Percival (2017) and Zhao (2019) in that it establishes the asymptotic optimality of a similar weighting estimator. It differs, however, in that it encompasses both approximate balance and exact balance. The technical conditions required by our result are among the weakest in the literature; they are considerably weaker than those required by Hirano et al. (2003) and Chan et al. (2016), and are comparable to those in Fan et al. (2016).

From a practical standpoint, we address two problems related to minimal weights: choosing the number of basis functions and selecting the degree of approximate balance. We derive a finite-sample upper bound for the potential loss incurred by balancing too many basis functions of the covariates. This result shows that the loss due to balancing too many basis functions is hedged by minimal weights because the number of active balancing constraints is implicitly bounded.

Finally, we provide a tuning algorithm for calibrating the degree of approximate balance in minimal weights. This is a general problem in weighting and so this algorithm may be of independent interest. We conclude with four empirical studies which suggest that approximate balance is preferable to exact balance, especially when there is limited overlap in covariate distributions. These studies show that use of approximate balancing weights with the proposed tuning algorithm yields weighting estimators with considerably lower root mean squared error than their exact balancing counterparts.

2. A shrinkage estimation view of minimal weights

For simplicity of exposition, we focus on the problem of estimating a population mean from a sample with incomplete outcome data. We assume that the outcomes are missing at random (Little & Rubin, 2014). Under the closely related assumption of strong ignorability (Rosenbaum & Rubin, 1983), this problem is analogous to estimating an average treatment effect in an observational study. See Kang & Schafer (2007) for an example connecting the problems of causal inference and estimation with incomplete outcome data.

Consider a random sample of n units from a population of interest, where some of the units in the sample are missing due to nonresponse. Let Z_i be the response indicator such that $Z_i = 1$ if unit i responds and $Z_i = 0$ otherwise, for $i = 1, \ldots, n$. Write r for the total number of respondents. Denote by X_i the vector of observed covariates and Y_i the outcome of unit i.

Assume there is overlap; that is, the propensity score $\pi(x) = \mathrm{pr}(Z = 1 \mid X = x)$ satisfies $0 < \pi(x) < 1$. Furthermore, assume that the responses are missing at random. This assumption says that missingness can be fully explained by the observed covariates: $Y_i \perp\!\!\!\perp Z_i \mid X_i$ (Robins & Gill, 1997).

The goal is to estimate the population mean of the outcome, $\bar{Y} = E(Y_i)$. We use the linear estimator $\hat{Y}_w = \sum_{i=1}^{n} w_i Z_i Y_i$ for estimation, where the weights w_i adjust for or balance the observed covariates.

Conventionally, the weights w_i are obtained by fitting a model for the propensity score $\pi(x)$ and then inverting the predicted propensities. Despite being widely used, this approach has two problems in practice: first, balancing the covariates can be difficult due to misspecification of the propensity score model, if the sample size is small or if the covariates are sparse; second, the weighting estimator can be unstable due to the variability of the weights (see, e.g., Zubizarreta 2015 for a discussion).

To address these problems, several weighting methods have been proposed recently. These methods are encompassed by the following mathematical program:

$$\underset{w}{\text{minimize}} \quad \sum_{i=1}^{n} Z_i f(w_i) \tag{1}$$

$$\text{subject to} \quad \left| \sum_{i=1}^{n} w_i Z_i B_k(X_i) - \frac{1}{n} \sum_{i=1}^{n} B_k(X_i) \right| \leqslant \delta_k \quad (k = 1, \ldots, K), \tag{2}$$

where f is a convex function of the weights and $B_k(X_i)$ $(k = 1, \ldots, K)$ are smooth functions of the covariates. Typically, the functions B_k are basis functions for $E(Y_i)$ and are chosen as the moments of the covariate distributions, see Assumption 1 parts (iv) and (vi) below. Other common choices of B_k include spline bases (De Boor, 1972) and wavelet bases (Singh & Tiwari, 2006). The constants δ_k constrain the imbalances in B_k. They are summarized in the vector $\delta_{K\times 1} = (\delta_1, \ldots, \delta_K) \geqslant 0$. In (2) we can also constrain the weights to sum to unity, $\sum_{i=1}^{n} w_i = 1$, and to take positive values, $0 \leqslant w_i$ $(i = 1, \ldots, n)$. These two constraints together ensure that the weights do not extrapolate; that is, $0 \leqslant w_i \leqslant 1$ $(i = 1, \ldots, n)$. This is related to the sample boundedness property discussed in Robins et al. (2007), which requires the estimator to lie within the range of observed values of the outcome.

We call the class of weights that solve the above mathematical program minimal dispersion approximately balancing weights, or simply minimal weights. They have minimal dispersion because they explicitly minimize a measure of dispersion or extremity of the weights. They are approximate balancing weights because they have the flexibility to approximately, as opposed to exactly, balance covariates. This flexibility plays an important role in practice by trading bias for variance.

Special cases of minimal weights are the entropy balancing weights (Hainmueller, 2012) with $f(x) = x \log x$ and $\delta = 0$, the stable balancing weights (Zubizarreta, 2015) with $f(x) = (x - 1/r)^2$ and $\delta \in \mathbb{R}_0^+$, and the empirical balancing calibration weights (Chan et al., 2016) with $f(x) = D(x, 1)$ and $\delta = 0$, where $D(x, x_0)$ is a distance measure for a fixed $x_0 \in \mathbb{R}$ that is continuously differentiable in $x_0 \in \mathbb{R}$, nonnegative and strictly convex in x. With the exception of the stable balancing weights, these methods balance the covariates exactly by taking $\delta = 0$ and assuming that the optimization problem is feasible. Related methods that balance covariates approximately through a Lagrange relaxation of the balance constraints include those of Kallus (2017), Athey et al. (2018), Hirshberg & Wager (2019), Wong & Chan (2018) and Zhao (2019).

The dynamics between the feasibility and the efficacy of covariate balancing constraints are central to estimation with incomplete outcome data. Tightening these constraints could make the optimization program infeasible, but relaxing them could compromise the removal of biases due to covariate imbalances.

Studying these dynamics, however, calls for an alternative formulation of problem (1)–(2) with a solution that is easier to characterize. Theorem 1 provides such a formulation. It expresses the dual problem of (1)–(2) as an unconstrained problem by leveraging the structure of minimal weights. Since problem (1)–(2) is convex, its optimal solution and the solution to the dual problem will be the same (Boyd & Vandenberghe, 2004). Dual formulations of balancing procedures have been studied by Zhao & Percival (2017) and Zhao (2019). Theorem 1 helps us to articulate the role of approximate balance constraints.

The dual formulation in Theorem 1 establishes a connection between minimal weights and shrinkage estimation of the propensity score. At a high level, minimal weights are implicitly

fitting a model for the inverse propensity score with ℓ_1 regularization; the model is a generalized linear model on $B_k(\cdot)$, the basis functions of the covariates.

THEOREM 1. *The dual of problem* (1)–(2) *is equivalent to the unconstrained optimization problem*

$$\underset{\lambda}{\text{minimize}} \; \frac{1}{n}\sum_{j=1}^{n}\left[-Z_j n\rho\{B(X_j)^{\mathrm{T}}\lambda\} + B(X_j)^{\mathrm{T}}\lambda\right] + |\lambda|^{\mathrm{T}}\delta, \tag{3}$$

where $\lambda_{K\times 1}$ *is the vector of dual variables associated with the* K *balancing constraints and* $B(X_j) = \{B_1(X_j), \ldots, B_K(X_j)\}$ *denotes the* K *basis functions of the covariates, with* $\rho(t) = t/n - t(h')^{-1}(t) + h\{(h')^{-1}(t)\}$ *and* $h(x) = f(1/n - x)$. *Moreover, the primal solution* w_j^* *satisfies*

$$w_j^* = \rho'\{B(X_j)^{\mathrm{T}}\lambda^{\dagger}\} \quad (j = 1, \ldots, n),$$

where $\lambda^{\dagger}$ *is the solution to the dual optimization problem.*

The proof is given in the Supplementary Material. The key to this result is the form of the constraints in (2). These box constraints allow us to eliminate the positivity constraints on the dual variables after a change of variables.

In Theorem 1, the function $\rho(\cdot)$ is a transformation of the measure of dispersion of the weights $f(\cdot)$ in (1). For example, when $f(x) = x\log x$, as in the entropy balancing weights (Hainmueller, 2012), we have $\rho(x) = -\exp(-x-1)$ and $\rho'(x) = \exp(-x-1)$, which implies a propensity score model of the form $\pi(x) = \exp\{B(x)^{\mathrm{T}}\lambda + 1\}$; and when $f(x) = (x - 1/r)^2$, as in the stable balancing weights (Zubizarreta, 2015), we have $\rho(x) = -x^2/4 + x/r$ and $\rho'(x) = -x/2 + 1/r$, which implies $\pi(x) = \{1/r - B(x)^{\mathrm{T}}\lambda/2\}^{-1}$. At a high level, the function ρ' can be seen as a link function in generalized linear models. With specific choices of ρ', (3) resembles a regularized version of the tailored loss function approach in Zhao (2019).

Problem (3) comes down to ℓ_1 shrinkage estimation. The inverse propensity score function is estimated as a generalized linear model on the basis functions B with link function ρ'. The dual variables in λ can be seen as the coefficients of the basis functions in the propensity score regression model. Estimation is regularized by the weighted ℓ_1-norm of the coefficients in λ. The loss function is

$$L(\lambda) = -Zn\rho\{B(x)^{\mathrm{T}}\lambda\} + B(x)^{\mathrm{T}}\lambda. \tag{4}$$

The expectation of this loss function is minimized when λ satisfies $\{n\pi(x)\}^{-1} = \rho'\{B(x)^{\mathrm{T}}\lambda\} = w^*$. This is the key equation connecting minimal weights to the propensity score $\pi(x)$.

Theorem 1 says that if the propensity score depends heavily on a given covariate, then problem (1)–(2) will try hard to balance this covariate by assigning it a large dual variable. The dual variables in λ can be interpreted as shadow prices of the covariate balance constraints (see Boyd & Vandenberghe, 2004, § 5.6). If a constraint has a high shadow price, then relaxing it a little will result in a large reduction in the optimization objective, and vice versa. On the other hand, the ℓ_1 penalty decreases the dependence of the weights on covariates that are hard to balance.

Theorem 1 is related to the dual formulation of covariate balancing scoring rules under regularization (Zhao, 2019). The two results have similarities, but differ in their objectives: here we use the dual formulation of problem (1)–(2) to analyse the asymptotic and finite-sample properties of minimal weights (see § 3 and § 4.1), whereas Zhao (2019) uses a related dual formulation to

show that increased regularization in covariate balancing scoring rules can deteriorate covariate balance.

3. ASYMPTOTIC PROPERTIES

Theorem 1 connects minimal weights to shrinkage estimation of the inverse propensity score function. In this section, we leverage this connection to characterize the asymptotic properties of minimal weights. We assume the following conditions and prove that minimal weights are consistent estimates of the inverse propensity score function $1/\pi(x)$.

Assumption 1. The following conditions hold:

(i) the minimizer $\lambda^\circ = \arg\min_{\lambda\in\Theta} E[-Zn\rho\{B(X_i)^{\mathrm{T}}\lambda\} + B(X_i)^{\mathrm{T}}\lambda]$ is unique, where Θ is the parameter space for λ;

(ii) $\lambda^\circ \in \mathrm{int}(\Theta)$, where Θ is a compact set and $\mathrm{int}(\cdot)$ stands for the interior of a set;

(iii) there exists a constant $0 < c_0 < 1/2$ such that $c_0 \leqslant n\rho'(v) \leqslant 1 - c_0$ for any $v = B(x)^{\mathrm{T}}\lambda$ with $\lambda \in \mathrm{int}(\Theta)$; also, there exist constants $c_1 < c_2 < 0$ such that $c_1 \leqslant n\rho''(v) \leqslant c_2 < 0$ in some small neighbourhood $\mathcal{B}$ of $v^* = B(x)^{\mathrm{T}}\lambda^\dagger$;

(iv) there exists a constant C such that $\sup_{x\in\mathcal{X}} \|B(x)\|_2 \leqslant CK^{1/2}$ and $E\{B(X_i)B(X_i)^{\mathrm{T}}\} \leqslant C$;

(v) the number of basis functions K satisfies $K = o(n)$;

(vi) there exist constants $r_\pi > 1$ and λ_1^* such that the true propensity score function satisfies $\sup_{x\in\mathcal{X}} |m^*(x) - B(x)^{\mathrm{T}}\lambda_1^*| = O(K^{-r_\pi})$ where $m^*(\cdot) = (\rho')^{-1}[1/\{n\pi(x)\}]$;

(vii) $\|\delta\|_2 = O_{\mathrm{p}}\{K^{1/2}(\log K)/n + K^{1/2-r_\pi}\}$.

In Assumption 1, (i) and (ii) are standard regularity conditions for consistency of minimum risk estimators. Condition (iii) makes it possible for consistency of $\lambda^\dagger$ to be translated into consistency of the weights. In particular, the fact that ρ'' is bounded implies that the derivative of the inverse propensity score function is bounded. This condition is satisfied by common choices of f in problem (1)–(2), including the variance, the mean absolute deviation and the negative entropy of the weights. Condition (iv) is a standard technical assumption that restricts the magnitude of the basis functions; see also Assumption 4.1.6 of Fan et al. (2016) and Assumption 2(ii) of Newey (1997). This condition is satisfied by many classes of basis functions, including the regression spline bases, trigonometric polynomial bases and wavelet bases (Newey, 1997; Horowitz & Mammen, 2004; Chen, 2007; Belloni et al., 2015; Fan et al., 2016). Condition (v) controls the growth rate of the number of basis functions relative to the number of units. Condition (vi) is a uniform approximation condition on the inverse propensity score function. It requires the basis $B(x)$ to be complete or $m^*(x)$ to be well approximated by a linear model on $B(x)$. For splines and power series, this assumption is satisfied by $r_\pi = s/d$, where s is the number of continuous derivatives of $m^*(\cdot)$ that exist and d is the dimension of x with a compact domain (Newey, 1997). Condition (vii) quantifies the extent to which the equality covariate balancing constraints can be relaxed such that the consistency of the resulting weight estimates is maintained.

Under these assumptions, we can prove that minimal weights are consistent for the inverse propensity score function.

THEOREM 2. *Let $\lambda^\dagger$ be the solution to (1)–(2) and $w^*(x) = \rho'\{B(x)^{\mathrm{T}}\lambda^\dagger\}$. Then, under the conditions in Assumption 1, we have:*

(i) $\sup_{x\in\mathcal{X}} |nw^*(x) - 1/\pi(x)| = O_{\mathrm{p}}\{K(\log K)/n + K^{1-r_\pi}\} = o_{\mathrm{p}}(1)$;

(ii) $\|nw^*(x) - 1/\pi(x)\|_{P,2} = O_{\mathrm{p}}\{K(\log K)/n + K^{1-r_\pi}\} = o_{\mathrm{p}}(1)$.

The proof, given in the Supplementary Material, consists of two steps. First we show that $\lambda^\dagger$, the solution to the dual problem, is close to λ_1^* in the ℓ_2-norm. Consistency of the weights then follows from the Lipschitz property of ρ' and the bounds on the basis functions in Assumption 1. In the special case of exact balance ($\delta = 0$), Theorem 2 is related to a result in Fan et al. (2016, Appendix D, p. 46). This connection stems from Theorem 1, as minimal weights are estimating the inverse propensity score.

We now assume the following additional conditions and prove that the resulting weighting estimator is consistent and semiparametrically efficient for the mean outcome.

Assumption 2. The following conditions hold:

(i) $E\{|Y_i - Y(X_i)|\} < \infty$, where $Y(x) = E(Y_i \mid X = x)$;
(ii) $E(Y_i^2) < \infty$, where $\bar{Y} = E(Y_i)$ is the population mean of the outcome;
(iii) there exist $r_y > 1/2$ and λ_2^* such that the outcome model $Y(x) = E(Y_i \mid X = x)$ satisfies $\sup_{x\in\mathcal{X}} |Y(x) - B(x)^{\mathrm{T}}\lambda_2^*| = O(K^{-r_y})$;
(iv) $m^*(\cdot) \in \mathcal{M}$ and $Y(\cdot) \in \mathcal{H}$, where $m^*(\cdot) = (\rho')^{-1}[1/\{n\pi(x)\}]$, $Y(\cdot)$ is the mean outcome function, and $\mathcal{M}$ and $\mathcal{H}$ are two sets of smooth functions satisfying $\log n_{[]}\{\varepsilon, \mathcal{M}, L_2(P)\} \leqslant C(1/\varepsilon)^{1/k_1}$ and $\log n_{[]}\{\varepsilon, \mathcal{H}, L_2(P)\} \leqslant C(1/\varepsilon)^{1/k_2}$ for a positive constant C and $k_1, k_2 > 1/2$, with $n_{[]}\{\varepsilon, \mathcal{S}, L_2(P)\}$ denoting the covering number of the set $\mathcal{S}$ by ε-brackets;
(v) $n^{0.5(r_\pi + r_y - 0.5)^{-1}} = o(K)$.

In Assumption 2, (i) and (ii) are standard regularity conditions which ensure that the estimators have finite moments, and (iii) is a uniform approximation condition similar to Assumption 1(vi), but on the mean outcome function $Y(x) = E(Y \mid X = x)$. Condition (iv) requires that the complexity of the function classes $\mathcal{M}$ and $\mathcal{H}$ do not increase too quickly as ε approaches zero. 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 $s/d > 1/2$ (van der Vaart & Wellner, 1996; Fan et al., 2016); see also Assumption 4.1.7 in Fan et al. (2016). Condition (v) controls the rate at which K can increase with respect to n. In particular, the rate depends on the sum of r_π and r_y, the approximation errors of the propensity score and outcome functions, respectively. This assumption relates to the product structure of error bounding in doubly robust estimation; see, for example, Kennedy (2016, equation (41)).

THEOREM 3. *Suppose that Assumptions 1 and 2 hold. Then*

$$n^{1/2}(\hat{Y}_{w^*} - \bar{Y}) \xrightarrow{d} N(0, V_{\mathrm{opt}})$$

in distribution, where $V_{\mathrm{opt}} = \mathrm{var}\{Y(X_i)\} + E\{\mathrm{var}(Y_i \mid X_i)/\pi(X_i)\}$ *is the semiparametric efficiency bound. If, in addition,* $r_y > 1$, *then the estimator*

$$\hat{V}_K = \frac{1}{n}\sum_{i=1}^{n}\Bigg[nZ_i w_i Y_i - \sum_{i=1}^{n} w_i Y_i - B(X_i)^{\mathrm{T}}\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^{\mathrm{T}} B(X_i)\right\}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^{\mathrm{T}} Y_i\right\}(nZ_i w_i - 1)\Bigg]^2$$

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

The proof can be found in the Supplementary Material. It uses empirical process techniques as in Fan et al. (2016) and involves the standard decomposition of $\hat{Y}_{w^*} - \bar{Y}$ into four components,

where three of them converge to zero in probability and the fourth is asymptotically normal and semiparametrically efficient. Each of the first three components can be controlled by the bracketing numbers of the function classes to which the inverse propensity score function and the outcome function belong. Assumption 2(ii) provides this control.

We conclude this section on asymptotic properties with a discussion of the uniform approximability conditions, Assumption 1(vi) and Assumption 2(iii). These assumptions depend both on the smoothness of the propensity score and outcome functions, and on the dimension d of the covariates. Suppose the propensity score and outcome functions both belong to the Hölder class with smoothness parameter s on the domain $[0, 1]^d$. Assumption 1(vi) and Assumption 2(iii) are among the weakest in the literature, as they require only $s/d > 1$ for the propensity score function and $s/d > 1/2$ for the outcome function. They are weaker than the assumptions in Hirano et al. (2003), which require $s/d > 7$ for the propensity score function and $s/d > 1$ for the outcome function, as well as those in Chan et al. (2016), which require $s/d > 13$ for the propensity score function and $s/d > 3/2$ for the outcome function. They are comparable to the conditions in Fan et al. (2016), which require $s/d > 1/2$ for the propensity score function and $s/d > 1/2$ for the outcome function, plus the requirement that the sum of these two ratios do not exceed $3/2$. To establish these results under weak assumptions, we use Bernstein's inequality as in Fan et al. (2016) and leverage the particular structure of minimal weights.

4. PRACTICAL CONSIDERATIONS

4.1. *The loss due to balancing too many functions of the covariates is bounded*

An important question that arises in practice relates to the cost of balancing too many basis functions of the covariates. In other words, practitioners are concerned about how big the loss will be if they balance more basis functions than needed. This is a valid concern because Theorem 1 implies that, for each basis function B_k we balance, we are implicitly including a similar term in the inverse propensity score model. Therefore, balancing too many basis functions could result in estimation loss due to fitting an overly complex model. The following oracle inequality provides reassurance by showing that this loss is bounded.

THEOREM 4. *Let $\lambda^\dagger$ be the solution to the dual of the minimal weights problem* (3)*, and let $\lambda^\ddagger$ be the solution to the dual of the exact balancing weights problem with the number of active constraints $\|\lambda^\ddagger\|_0$ capped by some constant $C_0 > 0$. Then, under suitable technical conditions,*

$$E\{L(\lambda^\dagger) - L(\lambda_1^*)\} \leqslant 3E\{L(\lambda^\ddagger) - L(\lambda_1^*)\} + c_0\|\lambda^\ddagger\|_0,$$

where λ_1^* is the oracle solution as in Assumption 1(vi)*, $L(\cdot)$ is the dual loss as in* (4)*, and c_0 is a positive constant depending on the number of basis functions K.*

See the Supplementary Material for technical details. This oracle inequality bounds $E\{L(\lambda^\dagger) - L(\lambda_1^*)\}$, the excess risk of the minimal weights estimator relative to the oracle estimator λ_1^*. The optimal dual loss $L(\lambda)$ is equal to the optimal primal loss $\sum_{i=1}^n Z_i f(w_i)$ in (1), because the optimization problem (1)–(2) is convex. A smaller excess risk translates into a smaller estimation error of the causal effect estimator.

This inequality compares the linear weighted estimator with two versions of minimal weights: one with approximate balance and the other with exact balance. The exact balancing version caps the number of exact balancing constraints at C_0. The inequality shows that the two estimators have similar risks.

More specifically, when there are few active covariate balancing constraints, $\|\lambda^{\ddagger}\|_0$ will be small. The inequality then says that the excess risk of approximate balancing in minimal weights is of the same order as the excess risk of exact balancing with its number of balancing constraints capped. Therefore, balancing covariates approximately can be seen as implicitly capping the number of active balancing constraints.

At a high level, this oracle inequality bounds the loss of balancing too many functions of the covariates with minimal weights. Fundamentally, the approximate balancing constraints in problem (1)–(2) are performing ℓ_1 regularization in the inverse propensity score estimation problem. This sparse behaviour of the balancing constraints is common in practice; for example, it can be seen in the 2010 Chilean post-earthquake survey data of Zubizarreta (2015, Fig. 1).

4.2. *A tuning algorithm for choosing the degree of approximate balance δ*

Another practical question that arises with minimal weights is how to choose the degree of approximate balance δ. Similar to the regularization parameter accompanying the ℓ_1-norm in lasso estimation, δ is a tuning parameter that the investigator needs to choose. In our setting, choosing δ is particularly hard; since there are no outcomes, there is not a clear out-of-sample target to optimize toward. For choosing δ, we propose Algorithm 1.

Algorithm 1. Choosing δ in minimal weights.

For each δ in a grid $\mathcal{D} \subset [0, K^{-1/2}]$ of candidate imbalances
  Compute $\{w_i\}_{i=1}^n$ by solving problem (1)
  For each $j \in \{1, ..., J\}$
    Draw a bootstrap sample $\mathcal{S}_j$ from the original data
    Evaluate covariate balance C_j on the sample $\mathcal{S}_j$,
      $C_j := \sum_{k=1}^K ||\{\sum_{i \in \mathcal{S}_j} w_i Z_i B_k(X_i)\}/(\sum_{i \in \mathcal{S}_j} w_i Z_i) - \sum_{i=1}^n B_k(X_i)/n||_2 / \mathrm{sd}\{B_k(X)\}$
  Compute the mean covariate balance, $\bar{C}(\delta) := \sum_{j=1}^J C_j / J$
Output $\delta^* = \arg\min_{\delta \in \mathcal{D}} \bar{C}(\delta)$

The key idea behind Algorithm 1 is to use the covariate balance in the bootstrapped samples as a proxy for how well the target parameters are estimated. The intuition is that in theory the true inverse propensity score weights will balance the population as well as samples from the population. Therefore, if the weights are well calibrated and robust to sampling variation, they will have this same property. To this end, we evaluate the covariate balance on bootstrapped samples C_S with the weights computed from the original dataset. In the following subsection, we show that the value of δ selected by Algorithm 1 often coincides with or is close to the optimal δ that gives the smallest root mean squared error in estimating the target parameters. We recommend choosing values of δ smaller than $K^{-1/2}$, because larger values are likely to violate the conditions in Assumption 1.

4.3. *Empirical studies*

We illustrate the performance of minimal weights in four empirical studies. In these four studies we choose δ with Algorithm 1 and consider three dispersion measures of the weights: the sum of absolute deviations, $f(w) = |w - \bar{w}|$; the variance, $f(w) = (w - 1/r)^2$ (Zubizarreta, 2015);

Table 1. *Root mean squared error for (a) the average treatment effect and (b) the average treatment effect on the treated. The lowest error for each measure of dispersion is shown in italics; a hyphen indicates that exact balancing does not admit a solution. In the case of bad overlap, balancing covariates approximately reduces the error of the average treatment effect on the treated by a half compared to exact balance*

(a)	Good overlap		Bad overlap		(b)	Good overlap		Bad overlap	
Dispersion	Exact	Approx.	Exact	Approx.	Dispersion	Exact	Approx.	Exact	Approx.
Abs. Dev.	0.19	*0.18*	—	*0.27*	Abs. Dev.	*0.10*	*0.10*	0.24	*0.08*
Variance	*0.16*	0.17	—	*0.26*	Variance	*0.09*	*0.09*	0.18	*0.07*
Neg. Ent.	*0.16*	*0.16*	—	*0.27*	Neg. Ent.	0.10	*0.09*	0.20	*0.10*

Approx., approximate balancing; Abs. Dev., sum of absolute deviations; Neg. Ent., negative entropy.

and the negative entropy, $f(w) = w \log w$ (Hainmueller, 2012). We find that minimal weights with approximate balance *admit* a solution in cases where exact balance does not. Approximate balancing also achieves considerably lower root mean squared error than exact balancing when there is limited overlap in covariate distributions.

The results of three of the simulation studies are reported in the Supplementary Material: the Kang & Schafer (2007) example, the LaLonde (1986) dataset, and the Wong & Chan (2018) simulation. Here we present a simulation study based on the right heart catheterization dataset of Connors et al. (1996).

The right heart catheterization dataset was first used to study the effectiveness of right heart catheterization in the initial care of critically ill patients. The dataset contains 2998 observations and 77 variables, including covariates, a treatment indicator, and the outcome. Balancing the 75 available covariates exactly is not feasible in most of the simulated datasets, so for comparison purposes we restrict the analyses to the 23 covariates listed in Table 1 of Connors et al. (1996). We generate the datasets and calculate the minimal weights, with both exact and approximate balance, using only these 23 covariates.

Based on this dataset, we generate 1000 simulated datasets as follows. We construct the treatment indicator Z_i as $Z_i = \mathbb{1}_{\{Z_i^* > 0\}}$ with $Z_i^* = (\alpha + \beta X_i)/c + \text{Unif}(-0.5, 0.5)$, where X_i denotes the observed covariates. In the model for Z_i^*, α and β are obtained by fitting a logistic regression to the original treatment indicator in the original dataset. We simulate two scenarios, one with good overlap ($c = 10$) and another with bad overlap ($c = 1$). For both scenarios, we generate pairs of potential outcomes $\{Y_i(0), Y_i(1)\}$ by fitting a regression model to the original treated and control outcomes and predicting on the entire sample. We obtain the observed outcome by letting $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.

In both scenarios, we compare the root mean squared errors of the estimated average treatment effects on the entire and treated populations using both minimal weights with Algorithm 1 and minimal weights with exact balance, i.e., with $\delta = 0$. The results are presented in Fig. 1 and the Supplementary Material.

Table 1(a) presents the root mean squared error of minimal weights in estimating the average treatment effect. When the data exhibit bad overlap, minimal weights provide good estimates, whereas their exact balancing counterparts do not *admit* a solution. With good overlap, minimal weights with approximate balancing perform similarly to exact balancing.

Table 1(b) shows the results for the average treatment effect on the treated. In this case, both exact and approximate balance admit solutions under bad overlap. The table shows that approximate balance can markedly reduce the root mean squared error relative to exact balance. While in a low-dimensional regime we balance fewer basis functions than the total number of

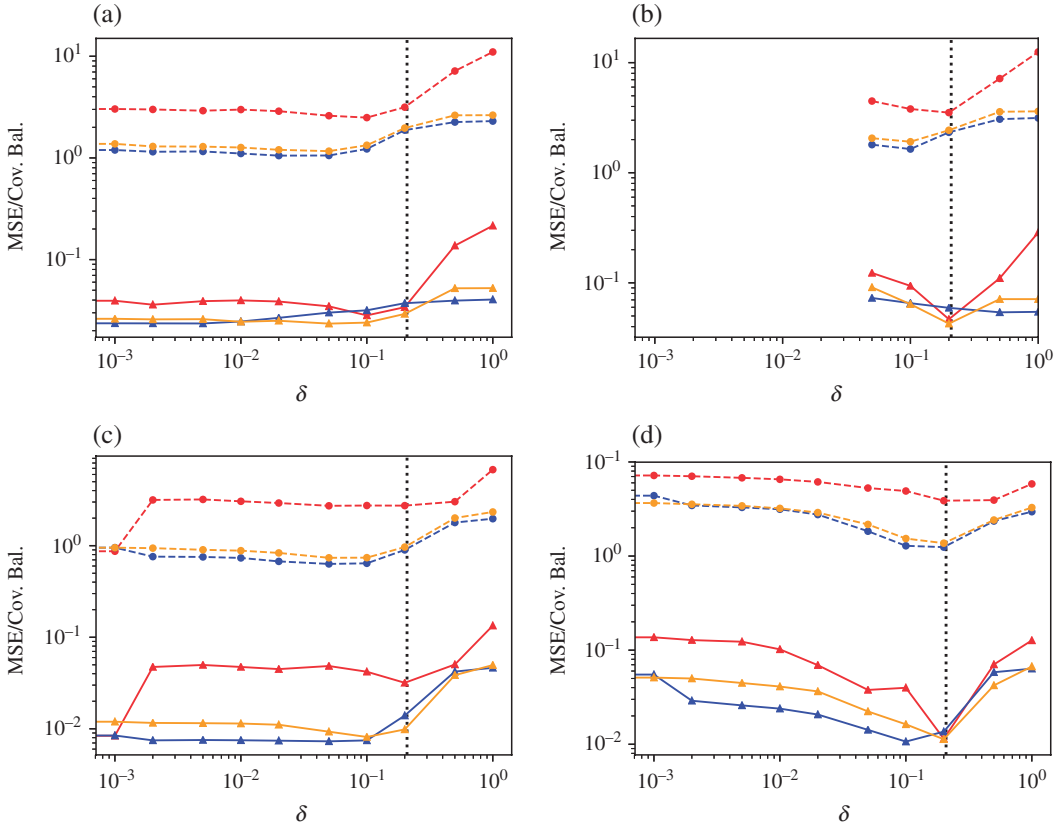

Fig. 1. Mean squared error and bootstrapped covariate balance for different values of the tuning parameter δ: (a) good overlap, average treatment effect; (b) bad overlap, average treatment effect; (c) good overlap, average treatment effect on the treated; (d) bad overlap, average treatment effect on the treated. In each panel, δ starts at 0 on the horizontal axis, and the vertical dotted line indicates $\delta = K^{-1/2}$, where K is the number of basis functions balanced. The δ value selected according to the bootstrapped covariate balance, as in Algorithm 1, often coincides with or is close to the optimal δ with the smallest error. We recommend choosing values of δ smaller than $K^{-1/2}$, as greater values are likely to violate the conditions in Assumption 1. MSE, mean squared error; Cov. Bal., covariate balance; Abs. Dev., sum of absolute deviations; Neg. Ent., negative entropy. Abs. Dev. Cov. Bal., ●-; Abs. Dev. MSE, ▲-; Variance Cov. Bal., ●-; Variance MSE, ▲-; Neg. Ent. Cov. Bal., ●-; Neg. Ent. MSE, ▲-.

observations, approximate balance, or ℓ_1 regularization, still helps to reduce the error. The reason is that approximate balance trades bias for variance. In fact, when there is bad overlap, traditional weighting estimators which use weights that balance covariates exactly tend to have high variance as they rely heavily on a few observations. In such cases, approximate balance can pull back from those observations and trade bias for variance to reduce the overall error.

Figure 1 shows that the root mean squared error of the effect estimates is sensitive to the choice of δ. Moreover, the value of δ selected by Algorithm 1 often coincides with the optimal value of δ that produces the lowest mean squared error; see the solid lines in Fig. 1. Again, Algorithm 1 selects the value of δ that minimizes the bootstrapped covariate balance, dashed lines in Fig. 1. We observe that when δ achieves the lowest bootstrapped covariate balance, i.e., the dashed lines, it also attains the lowest error, solid lines. In each panel of the figure, the dotted line indicates a value of δ equal to $K^{-1/2}$, where K is the number of basis functions of the covariates being balanced. We recommend choosing values of δ smaller than $K^{-1/2}$ for Assumption 1(vii) required by Theorem 3 to hold.

In general, minimal weights tuned with Algorithm 1 exhibit better empirical performance in the right heart catheterization dataset than their exact balancing counterparts. Empirical studies with the Kang & Schafer (2007) example, the LaLonde (1986) dataset, and the Wong & Chan (2018) simulation show a similar pattern. See the Supplementary Material for details.

5. FUTURE RESEARCH

The theoretical results developed in this work can be extended to matching, where covariates are balanced approximately, but with weights that encode an assignment between matched units (e.g., Rubin, 1973; Rosenbaum, 1989; Hansen, 2004; Abadie & Imbens, 2006; Zubizarreta, 2012; Diamond & Sekhon, 2013). The tuning algorithm used to select the degree of approximate balance can also be extended to matching. Promising directions for future work include doubly robust estimation (Robins & Rotnitzky, 1995), where propensity score modelling weights can be replaced by minimal weights (see Athey et al., 2018; Hirshberg & Wager, 2019). Also, minimal weights can be extended to instrumental variables and regression discontinuity settings, where model-based inverse probability weights are used for covariate adjustments.

ACKNOWLEDGEMENT

We thank the editor, the associate editor and two reviewers for their insightful comments. We are grateful to David Blei, Zach Branson, Ambarish Chattopadhyay, Xinkun Nie, Stefan Wager, Anna Zink and Qingyuan Zhao for their valuable feedback on the manuscript. We also acknowledge support from the Alfred P. Sloan Foundation. Zubizarreta is also affiliated with the Department of Statistics at Harvard University.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of the theorems and lemmas, further details on the empirical studies, and code for the simulations.

REFERENCES

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

ATHEY, S., IMBENS, G. W. & WAGER, S. (2018). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *J. R. Statist. Soc.* B **80**, 597–623.

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. & KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Economet.* **186**, 345–66.

BOYD, S. & VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

CHAN, K. C. G., YAM, S. C. P. & ZHANG, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Statist. Soc.* B **78**, 673–700.

CHEN, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics*, vol. 6B. Amsterdam: Elsevier, pp. 5549–632.

CONNORS, A. F., SPEROFF, T., DAWSON, N. V., THOMAS, C., HARRELL, F. E., WAGNER, D., DESBIENS, N., GOLDMAN, L., WU, A. W., CALIFF, R. M. et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *J. Am. Med. Assoc.* **276**, 889–97.

DE BOOR, C. (1972). On calculating with B-splines. *J. Approx. Theory* **6**, 50–62.

DEVILLE, J.-C. & SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Statist. Assoc.* **87**, 376–82.

Diamond, A. & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Statist.* **95**, 932–45.
Fan, J., Imai, K., Liu, H., Ning, Y. & Yang, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. https://imai.fas.harvard.edu/research/files/CBPStheory.pdf.
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. Am. Statist. Assoc.* **99**, 609–18.
Hellerstein, J. K. & Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Rev. Econ. Statist.* **81**, 1–14.
Hirano, K., Imbens, G. W. & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–89.
Hirshberg, D. A. & Wager, S. (2019). Augmented minimax linear estimation. *arXiv:* 1712.00038v5.
Horowitz, J. L. & Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *Ann. Statist.* **32**, 2412–43.
Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *J. R. Statist. Soc.* B **76**, 243–63.
Kallus, N. (2017). Generalized optimal matching methods for causal inference. *arXiv:* 1612.08321v3.
Kang, J. D. Y. & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with Discussion). *Statist. Sci.* **22**, 523–39.
Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research*. Basel: Springer, pp. 141–67.
LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* **76**, 604–20.
Li, F., Morgan, K. L. & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *J. Am. Statist. Assoc.* **113**, 390–400.
Little, R. J. & Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. Hoboken, New Jersey: John Wiley & Sons.
Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Economet.* **79**, 147–68.
Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when 'inverse probability' weights are highly variable. *Statist. Sci.* **22**, 544–59.
Robins, J. M. & Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statist. Med.* **16**, 39–56.
Robins, J. M. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Assoc.* **90**, 122–9.
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. (1987). Model-based direct adjustment. *J. Am. Statist. Assoc.* **82**, 387–94.
Rosenbaum, P. R. (1989). Optimal matching for observational studies. *J. Am. Statist. Assoc.* **84**, 1024–32.
Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York: Springer.
Rosenbaum, P. R. & 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**, 159–83.
Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Statist.* **2**, 808–40.
Singh, B. N. & Tiwari, A. K. (2006). Optimal selection of wavelet basis function applied to ECG signal denoising. *Digit. Sig. Proces.* **16**, 275–87.
van der Vaart, A. W. & Wellner, J. A. (1996). Weak convergence. In *Weak Convergence and Empirical Processes*. New York: Springer, pp. 16–28.
Wong, R. K. & Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika* **105**, 199–213.
Yiu, S. & Su, L. (2018). Covariate association eliminating weights: A unified weighting framework for causal effect estimation. *Biometrika* **105**, 709–22.
Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Ann. Statist.* **47**, 965–93.
Zhao, Q. & Percival, D. (2017). Entropy balancing is doubly robust. *J. Causal Infer.* **5**. DOI: 10.1515/jci-2016-0010.
Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Am. Statist. Assoc.* **107**, 1360–71.
Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Am. Statist. Assoc.* **110**, 910–22.
Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H. & Rosenbaum, P. R. (2011). Matching for several sparse nominal variables in a case-control study of readmission following surgery. *Am. Statistician* **65**, 229–38.

[*Received on* 20 *December* 2017. *Editorial decision on* 19 *March* 2019]