HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2019 February 04.

Published in final edited form as:

Biometrics. 2018 December ; 74(4): 1171–1179. doi:10.1111/biom.12887.

Doubly Robust Matching Estimators for High Dimensional Confounding Adjustment

Joseph Antonelli[1,*], **Matthew Cefalu**[2], **Nathan Palmer**[3], and **Denis Agniel**[2,3]

[1]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, U.S.A.

[2]RAND Corporation, Santa Monica, California 90401, U.S.A.

[3]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, U.S.A.

Summary.

Valid estimation of treatment effects from observational data requires proper control of confounding. If the number of covariates is large relative to the number of observations, then controlling for all available covariates is infeasible. In cases where a sparsity condition holds, variable selection or penalization can reduce the dimension of the covariate space in a manner that allows for valid estimation of treatment effects. In this article, we propose matching on both the estimated propensity score and the estimated prognostic scores when the number of covariates is large relative to the number of observations. We derive asymptotic results for the matching estimator and show that it is doubly robust in the sense that only one of the two score models need be correct to obtain a consistent estimator. We show via simulation its effectiveness in controlling for confounding and highlight its potential to address nonlinear confounding. Finally, we apply the proposed procedure to analyze the effect of gender on prescription opioid use using insurance claims data.

Keywords

Causal inference; Double robustness; High-dimensional data; Lasso; Matching; Prognostic score; Propensity score

1. Introduction

The goal of many research studies is to estimate the effect of a treatment on an outcome. If the treatment is not randomized, as is the case in observational studies, then care must be taken to ensure valid estimation. One common issue faced is the selection of covariates. The omission of a single confounding variable may lead to biased inference, while inclusion of unnecessary covariates may inflate the variance. This issue becomes more pertinent as the number of covariates increases, and standard methods for confounding adjustment will fail if the number of covariates is large compared to the sample size.

* jla538@mail.harvard.edu.

Recommendations for covariate selection when estimating treatment effects are varied but can be loosely classified into three categories: (1) control for all observed covariates; (2) selection based on substantive knowledge (VanderWeele and Shpitser, 2011); and (3) data-driven approaches (van der Laan and Gruber, 2010; De Luna et al., 2011; Vansteelandt et al., 2012; Wang et al., 2012; Wilson and Reich, 2014; Zigler and Dominici, 2014). In the age of big data, where access to and use of electronic medical records, administrative databases, and large-scale genomic and imaging datasets is increasingly common, the number of covariates available for analysis continues to grow. When the number of covariates is large relative to the number of observations, controlling for all observed covariates becomes infeasible and selection based on substantive knowledge becomes impractical. As such, the focus of this article will be on data-driven approaches for dimension reduction.

A variety of methods allow the inclusion of a high-dimensional vector of covariates in regression given that a sparsity condition holds. Arguably the most popular, the lasso (Tibshirani, 1996) places a penalty on the absolute value of the coefficients for the covariates, which forces many of the coefficients to zero, leading to a more parsimonious model. Many similar penalization methods have been proposed (Fan and Li, 2001; Zou and Hastie, 2005; Zou, 2006, e.g.,). These approaches suffer in the context of effect estimation as they are designed for prediction. Estimation and prediction are very different statistical challenges, and the variables that one requires for valid estimation may be different than those needed for prediction. Fitting a lasso model for the outcome focuses only on the relationship of each variable with the outcome and ignores any association with the treatment. Such a method will tend to omit variables that are strongly associated with treatment but weakly associated with the outcome. Omission of these variables may lead to bias in the estimated effect.

Other methods have been developed to address this issue by performing variable selection or model averaging aimed at selection of confounders for use in effect estimation. Wang et al. (2012) proposed a Bayesian model averaging procedure that uses an informative prior to place more weight a priori on outcome models that include covariates associated with the exposure. Many ideas have built on this prior specification to address the issue of confounder selection and model uncertainty (Talbot et al., 2015; Wang et al., 2015; Cefalu et al., 2017). There also exists a small literature on dimension-preserving statistics that can be used in a similar manner as propensity scores to balance confounders between levels of a binary treatment. These approaches can be found in Ghosh (2011); Nelson and Noorbaloochi (2013); Lue (2015), and the references within. All of the aforementioned approaches have been shown to work well in identifying confounders or adjusting for confounding. However, none of these approaches can handle a high-dimensional vector of confounders.

Recent literature has focused on the scenario in which the number of confounders may exceed the number of observations. Wilson and Reich (2014) took a decision theoretic approach to confounder selection and showed that this approach had strong connections to the adaptive lasso, but with weights designed to select confounders instead of predictors. Belloni et al. (2014) and Farrell (2015) utilized standard lasso models on both the exposure and outcome, identifying confounders as variables that enter into either lasso model, then

fitting an unpenalized regression model or doubly robust estimator on the reduced set of covariates. Ertefaie et al. (2018) addressed the issue of selecting weak confounders in small sample sizes by penalizing a joint likelihood on the exposure and outcome. Regularization for effect estimation is adopted from a Bayesian perspective in Hahn et al. (2016) by reparameterizing the likelihood and using horseshoe priors on the regression coefficients. Ghosh et al. (2015) utilized penalization in the potential outcomes framework, though their goal is to identify covariates that modify treatment effects.

The approach proposed in this article, called *doubly robust matching*, aims to handle high-dimensional confounding by matching on both the propensity score (Rosenbaum and Rubin, 1983) and the prognostic score (Hansen, 2008). Recent work by Leacy and Stuart (2014) has shown that matching on both scores in low-dimensional settings can lead to improved inference over simply matching on the propensity score. We extend these ideas to higher dimensions by incorporating penalization into both the propensity score model and the prognostic score model. In addition, we demonstrate that matching on both scores simultaneously is doubly robust, in the sense that the treatment effect is consistently estimated if either the propensity score or the prognostic score is correctly specified. Using high-dimensional simulations, we show that our doubly robust matching estimator is superior to other doubly robust estimators because it is not sensitive to extreme propensity scores and it appears to be robust to misspecification of both scores.

2. Methodology and Asymptotic Results

2.1. Notation and Framework

Suppose, we have collected N independent observations from (Y, W, X), where Y is the observed outcome, W is a binary treatment of interest, and X is a P−dimensional vector of covariates such that P may be larger than N. Let $Y(1)$ be the potential outcome under treatment and let $Y(0)$ be the potential outcome under control (Rubin, 1974). Our goal is the estimation of the average treatment effect defined as:

$$\tau = E[Y(1) - Y(0)], \quad (1)$$

where the expectation is over the population of interest.

For identification of the average treatment effect, we rely on the stable unit treatment value assumption (SUTVA), strong ignorability, and positivity. SUTVA is described elsewhere (Little and Rubin, 2000), but it can be understood as two conditions: the treatment received by one observation or unit does not affect the outcomes of other units and the potential outcomes are well-defined in the sense that there are not different versions of the treatment that lead to different potential outcomes. Strong ignorability and positivity are defined as:

Strong Ignorability: $Y(1), Y(0) \perp\!\!\!\perp W|X$

Positivity: $0 < \varphi(X) < 1$ for all X

where $\varphi(X) = P(W = 1|X)$ denotes the propensity score (Rosenbaum and Rubin, 1983). Under these assumptions, the average treatment effect is identified conditional on the propensity score:

$$\tau = E[E\{Y|W = 1, \varphi(X)\} - E\{Y|W = 0, \varphi(X)\}].$$

In addition, we define prognostic scores for each potential outcome as any scores $\Psi_0(X)$ and $\Psi_1(X)$ that satisfy the following conditions (Hansen, 2008):

$$Y(0) \perp\!\!\!\perp X \,|\, \Psi_0(X) \quad (2)$$

$$Y(1) \perp\!\!\!\perp X \,|\, \Psi_1(X) \quad (3)$$

For brevity, we restrict our attention to the case of no effect modification so that there is a single prognostic score, $\Psi(X)$, that satisfies (2) and (3). Under these assumptions, the average treatment effect is identified conditional on the prognostic score:

$$\tau = E[E\{Y|W = 1, \Psi(X)\} - E\{Y|W = 0, \Psi(X)\}].$$

Although many prognostic scores are possible, we consider $\Psi(X) = E[Y(0)|X]$. This will be a valid prognostic score if $Y(0)|X$ follows a generalized linear model or if $Y(0)$ only depends on the covariates through the mean. For a discussion of the implications of effect modification when using prognostic scores, see Section 2.5.

2.2. Identifiability and Double Robustness

In this section, we show that the average treatment effect is identified when conditioning on both the propensity score and the prognostic score. Interestingly, identifiability is maintained even if one of the two scores is incorrectly specified. This can be interpreted as a double robustness property, in which only one of the propensity score and prognostic score must be correctly specified to identify the average treatment effect.

Theorem 1. Assume that SUTVA, strong ignorability, and positivity hold. Further, assume there is no effect modification. Let $\varphi(X)$ be the true propensity score, let $\Psi(X)$ be a true prognostic score, and let $h(X)$ be any arbitrary function of X. Then,

$$Y(1), Y(0) \perp\!\!\!\perp W|\varphi(X), h(X) \text{ and } Y(1), Y(0) \perp\!\!\!\perp W|\Psi(X), h(X).$$

A proof of Theorem 1 can be found in Web Appendix A. Theorem 1 states that the treatment assignment is ignorable conditional on a correctly specified propensity or prognostic score and any other arbitrary function of the covariates. This result allows identification of the average treatment effect as:

$$\tau = E[E\{Y|\varphi(X), h(X), W = 1\} - E\{Y|\varphi(X), h(X), W = 0\}],$$

or

$$\tau = E[E\{Y|\Psi(X), h(X), W = 1\} - E\{Y|\Psi(X), h(X), W = 0\}].$$

Theorem 1 motivates and provides very strong justification for matching on both the propensity score and the prognostic score, as only one of the two scores must be correct to obtain valid estimates of the average treatment effect. Formal double robustness (i.e., consistency) based on matching is shown in Section 2.4. These results theoretically verify the simulation study by Leacy and Stuart (2014), which showed that bias remains small even when one of the scores is misspecified.

When the propensity score is misspecified, we leverage the assumption of no effect modification to identify the average treatment effect through the prognostic score. If there is effect modification, identifiability can be achieved by conditioning on two prognostic scores, one for each potential outcome, or by targeting the average treatment effect on the treated. For further discussion, see Section 2.5 and the Supplementary Materials.

2.3. Definition of the Doubly Robust Matching Estimator

Following Leacy and Stuart (2014), we propose to estimate the average treatment effect by matching on both the propensity score and the prognostic score. As indicated in Theorem 1, matching on both scores is doubly robust. For this reason, we will call our estimator the doubly robust matching estimator (DRME).

The DRME takes the form:

$$\tau(\theta) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i,\theta)} Y_j\right), \quad (4)$$

where M is the number of matches for each subject, $\mathcal{J}_M(i, \theta)$ is the set of matches for subject i, and θ is a set of parameters for the score models, adapting the notation from Abadie and Imbens (2006). Intuitively, the DRME finds M matches for each subject based on the similarity of their score values. The mean of the M matches is used to estimate the unobserved potential outcome for each subject, and the overall estimate of the average treatment effect is the mean difference between the potential outcomes for the N subjects in the data.

The DRME depends intrinsically on the propensity and prognostic score models through the matching set $\mathcal{J}_M(i, \theta)$. In the rest of this section, we will clarify this relationship by defining θ and $\mathcal{J}_M(i, \theta)$. We propose the two following models for the propensity and prognostic scores:

$$\varphi(X) = P(W = 1 \mid X) = g\left(X'\gamma_w\right) \quad (5)$$

$$\Psi(X) = E(Y \mid W = 0, X) = f\left(X'\gamma_y\right), \quad (6)$$

where $f(\cdot)$ and $g(\cdot)$ are inverse link functions and θ in (4) is the concatenation of γ_w amd γ_y. Importantly, as discussed in Theorem 1, we only require that one of (5) and (6) hold. Let

$$\gamma_w^* = \max_{\gamma} E_W[W\log\{g(X'\gamma)\} + (1 - W)\log\{1 - g(X'\gamma)\}] \quad (7)$$

$$\gamma_y^* = \min_{\gamma} E_Y\{Y - f(X'\gamma)\}^2 \quad (8)$$

be the possibly mis-specified targets of estimation.

Standard methods can be used to estimate (5) and (6) when $P \ll N$. However, when the covariate space is high-dimensional, it is very challenging to perform estimation and inference without additional assumptions. To this end, we assume *sparsity*: that the true number of target parameters (and thus the true number of covariates required for valid effect estimation) is much smaller than the total number of observations. Letting $\|\cdot\|_0$ denote the number of nonzero elements of a vector, we define sparsity in our context as $\left\|\gamma_y^*\right\|_0 \leq s$ and $\left\|\gamma_w^*\right\|_0 \leq s$, where s is an integer satisfying $s \ll N$. For more details regarding sparsity and its effect on high-dimensional estimation, as well as conditions on the covariate design matrix X, see (Van de Geer, 2008; Bickel et al., 2009; Negahban et al., 2012). We note that the assumption of sparsity may be relaxed to allow s to depend on N with the consequence of having to adjust rates of convergence to depend on s.

For high-dimensional P, we propose to estimate (5) and (6) using lasso models. Let

$$\hat{\gamma}_y = \operatorname*{argmin}_{\gamma} \sum_{i:W_i = 0} \left(Y_i - f\left(X'_i\gamma\right)\right)^2 + \lambda_y \sum_{j=1}^{P} \left|\gamma_j\right| \quad (9)$$

$$\hat{\gamma}_w = \operatorname*{argmin}_{\gamma} \sum_{i=1}^{n} \left[W_i \log\left\{g\left(X'_i\gamma\right)\right\} + \left(1 - W_i\right)\log\left\{1 - g\left(X'_i\gamma\right)\right\}\right] + \lambda_w \sum_{j=1}^{P} \left|\gamma_j\right|, \quad (10)$$

where λ_y and λ_w may be chosen to agree with asymptotic results or via cross-validation. Then the estimated propensity score and prognostic score are given by $\hat{\varphi}(X) = g\left(X'\hat{\gamma}_w\right)$ and $\hat{\Psi}(X) = f\left(X'\hat{\gamma}_y\right)$, respectively. Letting $Z = \left[\hat{\varphi}(X), \hat{\Psi}(X)\right]$, we define the matching set as:

$$\mathcal{J}_M\left(i, \hat{\theta}\right) = \left\{ j = 1, \ldots, N : W_j = 1 - W_i, \left(\sum_{k: W_k = 1 - W_i} I\left(\left\| Z_i - Z_k \right\| < \left\| Z_i - Z_j \right\| \right) \right) \leq M \right\}.$$

In practice, calipers are frequently used to ensure good matches and to ensure no matches outside of the common support. This is well known to change the quantity being estimated, and a nice discussion of this can be found in Iacus et al. (2015), but it can help in small samples to reduce bias of the matching estimator. Asymptotically, fixed-width calipers do not drop observations when positivity holds; therefore, we will not incorporate calipers in our asymptotic results. We will utilize calipers in our simulation study and in the analysis of insurance claims data found in Section 4.

2.4. Consistency of the Doubly Robust Matching Estimator

In this section, we demonstrate the consistency of the DRME for estimating the average treatment effect and show that it is doubly robust, in the sense that the average treatment effect is consistently estimated when either the model for the propensity score or the model for the prognostic score is correctly specified. We do not require that both are correctly specified. These results hold for high dimensions at a rate no slower than the standard rate for high-dimensional estimators, again, when either of the two models is correctly specified.

Theorem 2. Assuming SUTVA, strong ignorability, positivity, no effect modification, the regularity conditions necessary for asymptotic consistency of the lasso, sparsity as defined in (2.3), that at least 1 of the 2 high dimensional models is correctly specified, and additional weak conditions on the distribution of the data available in the Web Appendix, then

$$\tau\left(\hat{\theta}\right) - \tau = O_p\left(\sqrt{\frac{\log P}{N}}\right).$$

Theorem 2 has several important implications. First, matching on both a high-dimensional propensity score and a high-dimensional prognostic score is consistent. Second, this consistency only requires one of the two models to be correctly specified. This is the matching version of the well known double robustness property from the inverse weighting literature (Bang and Robins, 2005). Third, this result directly implies that in low-dimensional settings we have root-N consistency and double robustness.

Sketch Proof of Theorem 2: First, note that we can write the error of our estimator as:

$$\tau\left(\hat{\theta}\right) - \tau = \left(\tau\left(\hat{\theta}\right) - \tau\left(\tilde{\theta}\right)\right) + \left(\tau\left(\tilde{\theta}\right) - \tau\right) \quad (11)$$

where $\tilde{\theta} = \left(\gamma_w^*, \gamma_y^*\right)$ denotes the probability limit of $\hat{\theta}$. It is important to note that $\tilde{\theta}$ here does not necessarily represent the true propensity and prognostic score parameters. The first component of (11) is the error that arises from needing to estimate the parameters of the two score models, while the second component is the error induced by the matching process.

We examine the asymptotic behavior of each component separately and details can be found in Web Appendix B. Loosely speaking, the first component is no slower than the $\sqrt{\frac{\log P}{N}}$ rate inherited from estimating the parameters of the lasso models and does not require correct specification of either model. The second component requires at least one of the score models to be correctly specified and has the $N^{-1/2}$ rate from matching on two scores, which follows directly from Abadie and Imbens (2006). Combining the rates of convergence from the two components gives the final result.

2.5. Implications of Effect Modification

The results of the prior sections were derived under the assumption of no effect modification. If we relax this assumption, then conditioning on a single prognostic score $\Psi(X)$ is no longer sufficient for the identification of the average treatment effect. Instead, a separate prognostic score is needed for each potential outcome as defined in (2) and (3). Theorem 1 is easily relaxed to allow effect modification by conditioning on both of these prognostic scores, $\Psi_0(X)$ and $\Psi_1(X)$, in addition to the propensity score. A proof of this result is provided in Web Appendix A. However, the rate of convergence in Theorem 2 may potentially suffer when matching on more than two scores because, in general, matching on more than two scores is consistent at a rate slower than root-N (Abadie and Imbens, 2006).

An alternative to matching on multiple prognostic scores in the presence of effect modification is to target the average treatment effect on the treated (ATT). The ATT is defined as $E[Y(1) - Y(0) \mid W = 1]$. Interestingly, the prognostic score for the potential outcome under control, $\Psi_0(X)$, defined in (2) is sufficient for identification of the ATT. This implies that both Theorems 1 and 2 hold for the ATT when matching on the propensity score and the prognostic score under control. Thus, regardless of the presence or absence of effect modification, our results show that matching on an estimated propensity and prognostic score in high-dimensionsis consistent for the ATT with rate no slower than $\sqrt{\frac{\log P}{N}}$.

2.6. Estimation of Standard Errors

Measuring the uncertainty of the doubly robust matching estimator is difficult due to the high-dimensional nature of the models used to estimate the propensity and prognostic scores. Limiting distributions for estimators based on propensity score matching have only recently been developed (Abadie and Imbens, 2016), and the estimation of uncertainty around lasso estimates is an ongoing topic of research. Combining the two to provide a limiting distribution from which inference can be performed is a difficult task and a topic of further research. Here, we provide an approximation to the standard error and assess its ability to provide valid confidence intervals through simulation. Conditional on the matches, any matching estimator can be written as a weighted average of the observed data:

$$\hat{\tau} = \frac{\sum_{i=1}^{n} W_i R_i Y_i}{\sum_{i=1}^{n} W_i R_i} - \frac{\sum_{i=1}^{n} (1 - W_i) R_i Y_i}{\sum_{i=1}^{n} (1 - W_i) R_i}, \quad (12)$$

where R_i is the weight given to subject i in the estimator. In the case of the doubly robust matching estimator described in Section 2.3, $R_i = 1 + \frac{K_i}{M}$, where K_i is the number of times subject i is used as a match. Therefore, given an estimate of the residual variance, $Var(Y|W,X)$, which we denote $\hat{\sigma}^2$, we can approximate the standard error as:

$$\widehat{se}(\hat{\tau}) = \frac{\hat{\sigma}^2 \sum_{i=1}^{n} W_i R_i^2}{\left(\sum_{i=1}^{n} W_i R_i\right)^2} + \frac{\hat{\sigma}^2 \sum_{i=1}^{n} (1 - W_i) R_i^2}{\left(\sum_{i=1}^{n} (1 - W_i) R_i\right)^2}. \quad (13)$$

For this article, we will estimate the residual variance by fitting a lasso model to the outcome regressed against the treatment and covariates and taking the average squared residual from the fitted model. This estimate of the variance does not account for uncertainty in estimation of the two scores. Therefore, this standard error underestimates the true standard error and may lead to anti-conservative interval estimates. We assess the performance of the estimation of standard errors in Section 3.3.

3. Simulation Study

We compare the DRME with several competing approaches, including standard regression models. Details of the data-generating mechanisms are left to the relevant sections, but in all cases, we simulate a binary treatment W, a continuous outcome Y, and independent standard normal covariates X. We consider the following set of estimation techniques:

(1) Naive approach that compares the mean of Y in the treated and control groups (Naive)

(2) Estimating the true model (Oracle)

(3) Fitting a lasso model to the outcome only, penalizing the potential confounders, but not penalizing the treatment effect (Outcome lasso)

(4) The double post selection approach of Belloni et al. (2014). This involves fitting the treatment model in equation 10 and an outcome model as in the outcome lasso approach. The union of the covariates with nonzero coefficients from these models is then used to fit a standard linear model. (Double Post Selection)

(5) Inverse probability weighted estimator with a lasso propensity score model (lasso IPW)

(6) Doubly robust approach of Farrell (2015) that uses the same working models as the double post selection, but fits a doubly robust estimator using the resulting covariates (Farrell)

(7) Doubly robust estimator from Bang and Robins (2005) fit with lasso models for propensity score and outcome models (lasso DR)

(8) Matching approach that matches on a high-dimensional propensity score estimated from a lasso regression of W on all the covariates (Propensity Score Matching)

(9) Matching approach that matches on a high-dimensional prognostic score estimated from a lasso regression of Y on all the covariates in the controls only (Prognostic Score Matching)

(10) Matching approach that matches on both the high-dimensional propensity and prognostic scores (Doubly Robust Matching)

For all matching procedures, full matching was used with calipers of 0.5 standard deviations on each of the matching variables to ensure that no poor matches were used in the data set. Throughout, the **glmnet** R package was used to perform the lasso. All tuning parameters were chosen through cross-validation using the default arguments of the function cv.glmnet.

3.1. Linear Confounding

First, we explore the scenario where the true treatment and outcome models are linear in the covariates on the logit and identity scales, respectively, that is, equations 5 and 6 hold. We set $N = 200$ and $P = 1000$ and simulate the treatment and outcome from the following models:

$$W \sim Bernoulli\left\{\frac{exp\left(0.4X_1 + 0.9X_2 - 0.4X_3 - 0.7X_4 - 0.3X_5 + 0.6X_6\right)}{1 + exp\left(0.4X_1 + 0.9X_2 - 0.4X_3 - 0.7X_4 - 0.3X_5 + 0.6X_6\right)}\right\} \quad (14)$$

$$Y \sim Normal\left(-2 + W + 0.9X_1 - 0.9X_2 + 0.2X_3 - 0.2X_4 + 0.9X_7 - 0.9X_8, \sigma^2 = 1\right). \quad (15)$$

Therefore, covariates 1 through 4 are confounders, and notably covariates 3 and 4 are "weak" confounders in the sense that they have small associations with the outcome. Covariates 5 and 6 are instruments only associated with the treatment, while covariates 7 and 8 are predictive only of the outcome. The true average treatment effect is 1, which coincides with the regression coefficient for W in (15).

Table 1 shows the absolute bias, standard deviation (SD), and mean squared error (MSE) from this simulation for each of the estimators. The absolute bias is calculated as the absolute difference between the mean of the 1000 estimates and the truth. Matching on the propensity score, matching on the prognostic score, outcome lasso, and lasso IPW all result in substantial bias (more than 20%). Each of these approaches relies on a single model,

which appears to be undesirable in this high-dimensional setting. The double post selection, doubly robust matching, and Farrell estimators rely on two models, increasing the chance of adjusting for the important confounders. This is verified in the simulation as all three have smaller biases (7.5%, 8.5% and 8.7%) and have MSEs that compare favorably to the oracle outcome model. The double post selection approach has the smallest MSE in this setting, due to the linear relationship between the covariates and outcome. The double post selection model is fitting the correct outcome model in this case as it relies on the linearity assumption, while the doubly robust matching procedure does not (i.e., it is a nonparametric matching estimator).

3.2. Nonlinear Confounding

It is also of interest to examine the performance of the respective approaches when either the treatment or outcome models are nonlinear functions of the confounders. In this setting, approaches that assume the covariates enter into the models linearly will be misspecified and will no longer validly estimate the causal effect of interest. We use the same simulation framework as in Section 3.1, only now we simulate the treatment and outcome from the following models:

$$W \sim Bernoulli\left(\frac{exp\left(0.3X_1^2 + 0.5X_1^3 - 0.3X_2^4 + 0.4X_3^2\right)}{1 + exp\left(0.3X_1^2 + 0.5X_1^3 - 0.3X_2^4 + 0.4X_3^2\right)}\right) \quad (16)$$

$$Y \sim Normal\left(-2 + W - 0.5X_1 + 0.5X_2^2 + 0.4X_2^3 + 0.3X_3^2\sigma^2 = 1\right). \quad (17)$$

The first three covariates are confounders while the rest are noise. Each of the confounders has a nonlinear association with the treatment, outcome, or both. It is important to note that all of the estimated models are the same as in Section 3.1, which assume that the covariates enter into the systematic component of the models linearly. The lone exception is the oracle model, which again takes the form of the true regression function.

Table 2 shows the results of the simulation across 1000 replications. We see that none of the approaches, with the exception of the true model, are able to estimate the treatment effect without bias. Importantly, the doubly robust matching estimator proposed in this article substantially reduces the bias relative to any other approach. The bias is 6.7% for the doubly robust matching estimator, while the next smallest bias is 25.3% for the prognostic score matching estimator. The doubly robust matching estimator also has the lowest MSE of the non-oracle estimators at 0.133.

3.3. Investigation of Estimated Standard Errors

In this section, we assess the viability of our proposed estimate of the standard error for the DRME to obtain valid inference and interval coverage. We simulated data under the same scenario as Section 3.1 and varied $N \in \{200, 500, 1000, 2000\}$ and $P \in \{200, 500, 1000, 2000\}$, while keeping track of 95% interval coverage. Table 3 shows the confidence interval

coverage probabilities for each combination of N and P. We see that the estimated variance from (13) generally achieves near nominal coverages, with the lone exception being when the sample size is very small and the number of covariates is very large. This is the most difficult scenario in which the uncertainty in estimation of the propensity and prognostic scores is the highest. It is important to note, however, that the doubly robust matching estimator is somewhat biased in this scenario as seen in Table 1. This means that coverage below 95% is not solely due to underestimation of standard errors, but also due to the bias in the estimator. We have found that if we perform a bias correction on the confidence intervals using the empirically estimated bias, then coverage increases from 89% to 92% when $N = 200$ and $P = 2000$, indicating only slightly anti-conservative standard errors.

3.4. Sensitivity to Assumptions and Data Generating Mechanisms

Web Appendices C–G provide a number of additional simulation results assessing the performance of the doubly robust matching estimator. We investigated scenarios with different strengths of confounding, different nonlinear data generating mechanisms, scenarios where only one of the two models is misspecified, and different sample sizes and covariate dimensions. We find that the proposed estimator performs quite well across all scenarios, particularly when both models are misspecified. When only one of the two models is misspecified, it performs competitively with any of the existing doubly robust estimators, while it greatly reduces MSE when both models are incorrect. We empirically confirmed our theoretical results by showing consistency when one of the models is misspecified, and by finding that the MSE of the estimator converges at a rate faster than $\sqrt{\frac{\log P}{N}}$.

4. Analysis of Post-Surgical Prescription Opioid Use

In this section, we investigate the difference between males and females in the amount of opioids prescribed to commercially insured individuals after surgery. The United States is currently experiencing an epidemic of opioid dependence and abuse. If males or females were systematically prescribed more opioids, that could have significant effects on downstream addiction and could indicate an area for policy intervention. There is some controversy regarding causal estimates of immutable characteristics such as gender. While there exist studies aiming to estimate the causal effect of gender (Boyd et al., 2010), others have argued against this because one can not intervene or manipulate gender (Holland, 1986). Greiner and Rubin (2011) argue that it is relevant to estimate the causal effect of the perception of gender, rather than gender itself, as this could hypothetically be intervened upon. Regardless as to whether causal effects of gender are well defined, we believe that it is interesting to identify differences across gender after controlling for baseline characteristics.

To investigate this question, surgeries were ascertained from a de-identified administrative database of insurance claims at Aetna, Inc., a large, national commercial managed healthcare company. This database includes all 37,651,619 million members with medical and pharmacy insurance coverage between 2008 and 2016. Data includes all medical and pharmacy claims during the study period, as well as basic demographic information. Surgeries were identified via International Classification of Disease, version 9 (ICD-9)

procedure codes. Members were required to have six months of medical coverage, as well as three months of pharmacy coverage, before and after surgery. If a member had multiple surgeries that met the inclusion criteria, we only analyzed the first one. A total of $N = 205,934$ surgeries were included.

The outcome of interest is the total days supply of opioids for which the member filled a prescription in the 90 days following surgery. Opioids were identified in the database as drugs associated with the following common primary ingredients: codeine, fentanyl, hydrocodone, hydromorphone, morphine, oxycodone, oxymorphone, or tramadol. Injected drugs were excluded. Due to a heavy right-skew in the total days supply, we log-transformed total supply. Because observed differences in opioid prescription between sexes could be due to systematic over- or under-prescription, as well as due to differences in age, surgery types, or overall health, we considered a broad range of potential confounders: surgery date, surgery type, birth year, patient relationship to insurance subscriber, and all pre-surgical diagnosis codes observed within 6 months of the surgery date. Diagnosis codes that occur in less than 50 members and diagnosis codes that occur more than four times as often in one sex than the other were excluded. In total, there were 3696 covariates included in this analysis.

One advantage of using matching-based procedures is the ability to assess balance of the covariates before and after matching by looking at the absolute standardized difference in means between the males and females for each covariate. Figure 1 shows the absolute standardized difference for each covariate before matching and after doubly robust matching. We exclude propensity score matching from the figure since it is very similar to doubly robust matching, though there are small differences that can be seen in Table 4. Both approaches lead to absolute standardized differences less than 0.1 for all covariates, indicating that they successfully removed differences between males and females with respect to the observed covariates. Table 4 indicates that doubly robust matching is doing exceptionally well with regards to covariate balance, as it obtains a lower maximum and average absolute standardized difference across the covariates than propensity score matching.

We estimate the difference in the log-total days supply of opioids using the same set of approaches evaluated in Section 3. The outcome lasso, lasso DR, and lasso IPW approaches do not have existing approaches for uncertainty assessment, and therefore we do not include confidence intervals for these estimators. Tuning parameters for the treatment and outcome lasso models were chosen via cross validation. In total, 467 and 1823 covariates had nonzero coefficients in the outcome and treatment models, respectively. Table 5 presents the estimated difference between males and females in this population of commercially insured patients.

The naive approach estimates a statistically significant difference between males and females, indicating that males receive about 0.06 log-days fewer (or about 95% of the supply of females). With the exception of the lasso IPW estimator, all of the estimators that adjust for the high-dimensional set of confounders attenuate the difference between males and females markedly. The doubly robust matching procedure estimates that males, on average

and after controlling for the observed covariates, receive a supply that is for 0.007 fewer log-days, or about 99.3% as long, and the confidence interval runs from 0.01 log-days more supply to 0.025 log-days less supply. Regardless of statistical significance, this estimated difference is not practically meaningful.

5. Discussion

In this article, we propose a doubly robust matching estimator for high-dimensional confounding adjustment. This work has extended the literature on confounding adjustment in two important ways. First, we have shown theoretically that matching on a high-dimensional score such as a propensity or prognostic score is consistent with rate no slower than the usual rate for high-dimensional estimators, $\sqrt{\frac{\log P}{N}}$. This shows that matching presents a useful approach when trying to estimate treatment effects in difficult, high-dimensional settings. Second, matching on both the propensity score and the prognostic score is doubly robust; that is, as long as one of the two scores is correctly specified, the matching procedure is consistent.

Many existing approaches in the literature require correct modeling of the relationship between the outcome and confounders. Models that allow for complex interactions and nonlinearities in the confounders are very difficult (if not impossible) to implement in high dimensions. Our simulations suggest that the doubly robust matching estimator is fairly robust when both the propensity score model and the prognostic score model are misspecified, indicating that simple models for the scores can be used to remove much of the bias. This extra robustness of matching estimators has been seen before as Leacy and Stuart (2014) found matching estimators to be the most robust to model misspecification and Waernbaum (2012) found matching estimators to be more robust than doubly robust estimators under misspecification. Our results show that these ideas extend to high-dimensional settings and further justify the use of matching estimators.

One limitation of the doubly robust matching estimator is the current inability to derive the asymptotic variance. While this is certainly a theoretical limitation, we illustrated that reasonable confidence intervals can be constructed by ignoring the uncertainty in the estimation of the models used to create matches.

6. Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding for this work was provided by National Institutes of Health (ES000002, ES024332, ES007142, ES026217 P01CA134294, R01GM111339, R35CA197449, P50MD010 428) and National Institute on Drug Abuse (R01DA040721). Joseph Antonelli and Matthew Cefalu are co-lead authors.

References

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

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

Bang H and Robins JM (2005). Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–973. [PubMed: 16401269]

Belloni A, Chernozhukov V, and Hansen C (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies 81, 608–650.

Bickel PJ, Ritov Y, and Tsybakov AB (2009). Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics 37, 1705–1732.

Boyd CL, Epstein L, and Martin AD (2010). Untangling the causal effects of sex on judging. American Journal of Political Science 54, 389–411.

Cefalu M, Dominici F, Arvold N, and Parmigiani G (2017). Model averaged double robust estimation. Biometrics 73, 410–421. [PubMed: 27893927]

De Luna X, Waernbaum I, and Richardson TS (2011). Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika 98, 861–875.

Ertefaie A, Asgharian M, and Stephens DA (2018). Variable selection in causal inference using a simultaneous penalization method. Journal of Causal Inference 6.1.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96, 1348–1360.

Farrell MH (2015). Robust inference on average treatment effects with possibly more covariates than observations. Journal of Econometrics 189, 1–23.

Ghosh D (2011). Propensity score modelling in observational studies using dimension reduction methods. Statistics & Probability Letters 81, 813–820. [PubMed: 21617766]

Ghosh D, Zhu Y, and Coffman DL (2015). Penalized regression procedures for variable selection in the potential outcomes framework. Statistics in Medicine 34, 1645–1658. [PubMed: 25628185]

Greiner DJ and Rubin DB (2011). Causal effects of perceived immutable characteristics. Review of Economics and Statistics 93, 775–785.

Hahn PR, Carvalho C, and Puelz D (2016). Bayesian regularized regression for treatment effect estimation from observational data. Available at SSRN.

Hansen BB (2008). The prognostic analogue of the propensity score. Biometrika 95, 481–488.

Holland PW (1986). Statistics and causal inference. Journal of the American statistical Association 81, 945–960.

Iacus SM, King G, and Porro G (2015). A theory of statistical inference for matching methods in applied causal research. URL: http://gking.harvard.edu/publications/how-Coarsening-Simplifies-Matching-Based-Causal-Inference-Theory.

Leacy FP and Stuart EA (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: A simulation study. Statistics in Medicine 33, 3488–3508. [PubMed: 24151187]

Little RJ and Rubin DB (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. Annual Review of Public Health 21, 121–145.

Lue H-H (2015). An inverse-regression method of dependent variable transformation for dimension reduction with nonlinear confounding. Scandinavian Journal of Statistics 42, 760–774.

Negahban S, Yu B, Wainwright MJ, and Ravikumar PK (2012). A unified framework for high-dimensional analysis of m−estimators with decomposable regularizers. Statistical Science, 1348–1356.

Nelson D and Noorbaloochi S (2013). Information preserving sufficient summaries for dimension reduction. Journal of Multivariate Analysis 115, 347–358.

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology 66, 688.

Talbot D, Lefebvre G, and Atherton J (2015). The bayesian causal effect estimation algorithm. Journal of Causal Inference 3, 207–236.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B (Methodological) 58, 267–288.

Van de Geer SA (2008). High-dimensional generalized linear models and the lasso. The Annals of Statistics 36, 614–645.

van der Laan MJ and Gruber S (2010). Collaborative double robust targeted maximum likelihood estimation. The international journal of biostatistics 6.

VanderWeele TJ and Shpitser I (2011). A new criterion for confounder selection. Biometrics 67, 1406–1413. [PubMed: 21627630]

Vansteelandt S, Bekaert M, and Claeskens G (2012). On model selection and model misspecification in causal inference. Statistical Methods in Medical Research 21, 7–30. [PubMed: 21075803]

Waernbaum I (2012). Model misspecification and robustness in causal inference: Comparing matching with doubly robust estimation. Statistics in Medicine 31, 1572–1581. [PubMed: 22359267]

Wang C, Dominici F, Parmigiani G, and Zigler CM (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. Biometrics 71, 654–665. [PubMed: 25899155]

Wang C, Parmigiani G, and Dominici F (2012). Bayesian effect estimation accounting for adjustment uncertainty. Biometrics 68, 661–671. [PubMed: 22364439]

Wilson A and Reich BJ (2014). Confounder selection via penalized credible regions. Biometrics 70, 852–861. [PubMed: 25123966]

Zigler CM and Dominici F (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. Journal of the American Statistical Association 109, 95–107. [PubMed: 24696528]

Zou H (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 67, 301–320.

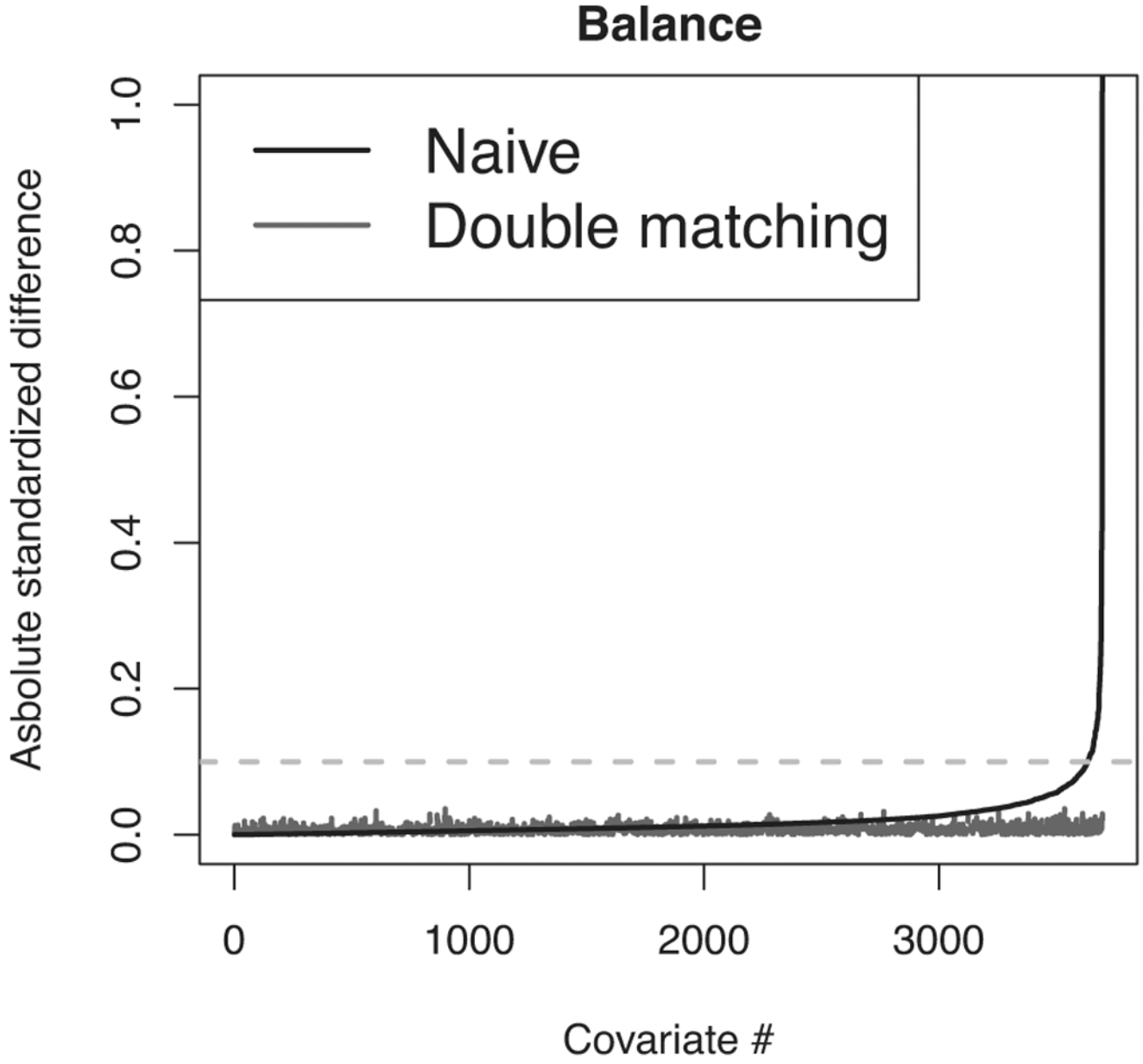

Figure 1.
Balance of covariates before matching and after doubly robust matching. Balance is measured in terms of the absolute standardized difference in means between the treated and control groups.

Table 1

Absolute bias, standard deviation, and mean squared error from the simulation of Section 3.1 across 1000 replications

Type	Absolute bias	SD	MSE
Oracle	0.002	0.160	0.026
Naive	0.490	0.285	0.321
Outcome Lasso	0.290	0.176	0.115
Double post selection	0.075	0.198	0.045
Lasso IPW	0.365	0.243	0.192
Farrell	0.087	0.370	0.145
Lasso DR	0.226	0.169	0.080
Propensity score matching	0.255	0.466	0.282
Prognostic score matching	0.242	0.240	0.116
Doubly robust matching	0.085	0.232	0.061

Table 2

Absolute bias, standard deviation, and mean squared error from the simulation of Section 3.2 across 1000 replications

Type	Absolute bias	SD	MSE
Oracle	0.007	0.026	0.026
Naive	0.611	0.296	0.461
Outcome Lasso	0.607	0.272	0.442
Double post selection	0.365	0.281	0.212
Lasso IPW	0.570	0.287	0.407
Farrell	0.414	0.350	0.294
Lasso DR	0.553	0.268	0.378
Propensity score matching	0.368	0.426	0.317
Prognostic score matching	0.253	0.276	0.140
Doubly robust matching	0.067	0.359	0.133

Table 3

Coverage probabilities for a variety of data dimensions using the proposed standard error estimate

	P **= 200**	***P*** **= 500**	***P*** **= 1000**	***P*** **= 2000**
N = 200	0.948	0.927	0.958	0.887
N = 500	0.965	0.961	0.954	0.968
N = 1000	0.957	0.961	0.964	0.971
N = 2000	0.950	0.928	0.964	0.939

Table 4

Illustration of the balance of the covariates before and after matching. Mean is the average absolute standardized difference across all covariates. Unbalanced mean is the same metric, except averaged only over covariates who had a naive balance greater than 0.1. Maximum is the largest absolute standardized difference across all covariates.

	Absolute standardized difference		
Type	**Mean**	**Unbalanced Mean**	**Maximum**
Naive	0.02	0.19	1.91
Propensity score matching	0.01	0.02	0.07
Doubly robust matching	0.01	0.01	0.04

Table 5

Estimated difference between males and females in the log-total days supply of opioids for which members filled a prescription in the 90 days following surgery. Negative estimates indicate that males fill prescriptions with a supply of fewer days.

Estimator	Difference (95% CI)	Standard error
Naive	−0.055 (−0.065, −0.045)	0.005
Outcome lasso	−0.017 (−, −)	-
Double Post Selection	−0.016 (−0.026, −0.006)	0.005
Lasso IPW	−0.066 (−, −)	-
Farrell	−0.017 (−0.068, 0.034)	0.026
Lasso DR	−0.009 (−, −)	-
Propensity score matching	0.006 (−0.012, 0.024)	0.009
Prognostic score matching	0.011 (0.001, 0.021)	0.005
Doubly robust matching	−0.007 (−0.025, 0.010)	0.009