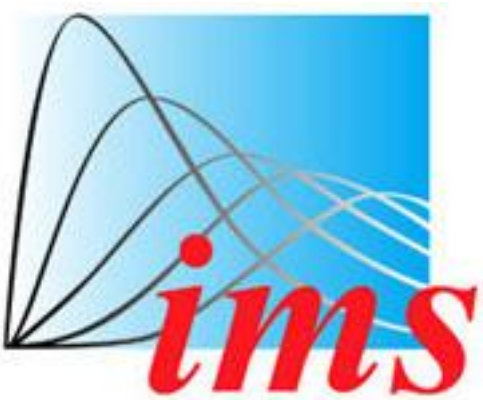

COVARIATE BALANCING PROPENSITY SCORE FOR A CONTINUOUS TREATMENT

Author(s): Christian Fong, Chad Hazlett and Kosuke Imai

Source: *The Annals of Applied Statistics*, March 2018, Vol. 12, No. 1 (March 2018), pp. 156-177

Published by: Institute of Mathematical Statistics

Stable URL: https://www.jstor.org/stable/10.2307/26542524

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/10.2307/26542524?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve and extend access to *The Annals of Applied Statistics*

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

The Annals of Applied Statistics
2018, Vol. 12, No. 1, 156–177
https://doi.org/10.1214/17-AOAS1101
© Institute of Mathematical Statistics, 2018

COVARIATE BALANCING PROPENSITY SCORE FOR A CONTINUOUS TREATMENT: APPLICATION TO THE EFFICACY OF POLITICAL ADVERTISEMENTS

BY CHRISTIAN FONG, CHAD HAZLETT, AND KOSUKE IMAI

Stanford University, University of California, Los Angeles and Princeton University

Propensity score matching and weighting are popular methods when estimating causal effects in observational studies. Beyond the assumption of unconfoundedness, however, these methods also require the model for the propensity score to be correctly specified. The recently proposed covariate balancing propensity score (CBPS) methodology increases the robustness to model misspecification by directly optimizing sample covariate balance between the treatment and control groups. In this paper, we extend the CBPS to a continuous treatment. We propose the covariate balancing generalized propensity score (CBGPS) methodology, which minimizes the association between covariates and the treatment. We develop both parametric and nonparametric approaches and show their superior performance over the standard maximum likelihood estimation in a simulation study. The CBGPS methodology is applied to an observational study, whose goal is to estimate the causal effects of political advertisements on campaign contributions. We also provide open-source software that implements the proposed methods.

1. Introduction. Propensity score methods are popular among researchers who wish to infer causal effects in observational studies [e.g., Rosenbaum and Rubin (1983, 1984, 1985), Robins, Hernán and Brumback (2000), Hirano, Imbens and Ridder (2003)]. Under the assumption of unconfoundedness, propensity score matching and weighting methods aim to balance observed covariates across different values of a treatment variable [e.g., Imbens (2004), Ho et al. (2007), Stuart (2010)].

Despite the popularity of propensity score methods, the vast majority of their applications have been confined to a binary treatment. In particular, propensity score methods have rarely been applied to a continuous treatment. This dearth of applications to nonbinary treatments is not due to a lack of available methods. For example, researchers proposed inverse-probability weighting, subclassification, and regression adjustment based on the estimated density of the realized continuous treatment given the covariates to formulate weights [e.g., Robins, Hernán and Brumback (2000), Imai and van Dyk (2004), Hirano and Imbens (2004)]. All

Received January 2017; revised June 2017.
Key words and phrases. Causal inference, covariate balance, generalized propensity score, inverse-probability weighting, treatment effect.

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

of these promising methods, however, presume the accurate estimation of the unknown generalized propensity score. Unfortunately, this is not a trivial assumption. Scholars have found that even in the case of binary treatment where relatively straightforward diagnostics tools are available, the empirical results can be sensitive to model misspecification [e.g., Smith and Todd (2005), Kang and Schafer (2007)]. This problem is exacerbated for a continuous treatment where checking covariate balance is more difficult and less intuitive because the treatment variable takes a continuum of values.

An important practical consequence of this complication is that applied researchers across social and medical sciences often dichotomize a continuous treatment in order to utilize propensity score methods [e.g., Donohue III and Ho (2007), Harder, Stuart and Anthony (2008), Boyd, Epstein and Martin (2010), Nielsen et al. (2011), De and Ratha (2012)]. In Section 2, we introduce an observational study, in which the goal is to estimate the causal effects of political advertisements on campaign contributions [Urban and Niebler (2014)]. In the original study, the authors dichotomized the number of advertisements, which is essentially a continuous variable, into a binary treatment using an arbitrary threshold of 1000 advertisements. Using this dichotomized treatment variable, they conducted propensity score matching by using the logistic regression to estimate the propensity score. The dichotomization of treatment variable results in the loss of information, which can compromise substantive insights gained from the data analysis.

In Section 3, we fill this gap between methodological and applied research and develop a new method to estimate the propensity score for a continuous treatment. In particular, we propose to directly minimize the association between a continuous treatment variable and covariates when estimating the generalized propensity score. In recent years, several researchers have proposed methods that estimate propensity scores by optimizing covariate balance [e.g., McCaffrey, Ridgeway and Morral (2004), Tan (2010), Hainmueller (2012), Graham, Pinto and Egel (2012), Imai and Ratkovic (2014), Chan, Yam and Zhang (2016), Zhu, Coffman and Ghosh (2015), Zubizarreta (2015), Hazlett (2016), Fan et al. (2016), Zhao (2016)]. While these methods improve the robustness of propensity score methods, most of them are not applicable to a continuous treatment. The only exception we find is the method of Zhu, Coffman and Ghosh (2015), which extends the generalized boosting method of McCaffrey, Ridgeway and Morral (2004) to a continuous treatment. In this paper, we extend the covariate balancing propensity score (CBPS) methodology of Imai and Ratkovic (2014) to a continuous treatment [see Fan et al. (2016) for the theoretical properties of CBPS, and Imai and Ratkovic (2015) and Zhao (2016) for other extensions]. We call this extension the Covariate Balancing Generalized Propensity Score (CBGPS) methodology. In generalizing the CBPS, we consider both parametric (Section 3.2) and nonparametric (Section 3.3) approaches.

Once researchers obtain the estimated propensity score using CBGPS, they can employ a variety of methods including regression adjustment and subclassification to estimate causal effects [e.g., Hirano and Imbens (2004), Imai and van Dyk

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

(2004)]. In this paper, we focus on the inverse-probability weighting [Robins, Hernán and Brumback (2000)], as it is directly related to the covariate balance measure used in the CBGPS estimation. (The nonparametric CBGPS directly produces these weights rather than first estimating a generalized propensity score.) In Section 4, we conduct a simulation study to evaluate the performance of the proposed methodology. We find that the CBGPS is more robust to misspecification than the standard maximum likelihood estimation. It also compares favorably to the recently proposed GBM method [Zhu, Coffman and Ghosh (2015)], which utilizes gradient boosted trees to flexibly estimate propensity scores while seeking to improve finite sample balance. The bias and root-mean-squared error of treatment effect estimates we obtain is similar to those of GBM (though our nonparametric estimator, npCBGPS, outperforms it somewhat), while the balance obtained by our approach is far better.

In Section 5, we reanalyze the motivating observational study introduced in Section 2, but without dichotomizing the continuous treatment variable. We first show that the proposed generalization of CBPS reduces the association between the treatment variable and covariates more effectively than the standard maximum likelihood estimation method. We then demonstrate that additional substantive insights can be obtained by analyzing the original continuous treatment variable rather than its dichotomized version. Finally, we offer concluding remarks in Section 6. The proposed CBGPS methodology is implemented through publicly available open-source software CBPS [Fong et al. (2017)].

2. The effect of political advertisements on campaign contributions. In this section, we introduce an observational study from political science that motivates our methodology. Urban and Niebler (2014) explored the potential causal link between advertising and campaign contributions. Presidential campaigns ordinarily focus their advertising efforts on competitive states, but if political advertising drives more donations, then it may be worthwhile for candidates to also advertise in noncompetitive states. The authors exploit the fact that media markets sometimes cross state boundaries. This means that candidates may inadvertently advertise in noncompetitive states when they purchase advertisements for media markets that mainly serve competitive states. By restricting their analysis to noncompetitive states, the authors attempt to isolate the effect of advertising from that of other campaigning, which do not incur these media market spillovers.

The treatment of interest, the number of political advertisements aired in each zipcode, takes a range of values from 0 to 22,380 (with the average number of advertisements equal to 1903), and hence can essentially be considered as a continuous variable. Urban and Niebler dichotomized this political advertising variable by examining whether a zip code received more than 1000 advertisements or not. According to this operationalization, 5230 of 16,265 zip codes are classified

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

as "treated." Using this dichotomized treatment variable, the authors then conduct one-to-one nearest neighbor propensity score matching after using logistic regression to estimate the propensity score. The observed confounders included in this analysis are median household income, percent black, percent Hispanic, percent college graduates, and population density. In addition, the authors employ several different matching methods as robustness checks, including kernel matching. The authors found that advertising in noncompetitive states led to a statistically and substantively significant increase in the level of campaign contributions.

However, dichotomization makes the interpretation of the results difficult because the naive interpretation—the reported estimate represents the effect of airing 1000 advertisements instead of 0—is incorrect. Additionally, balancing covariates on the dichotomized treatment variable does not guarantee that the covariates will be balanced on the underlying continuous treatment variable. The estimate may be biased by this hidden imbalance. Elsewhere in their paper, Urban and Niebler (2014) estimate the dose-response curve using the original nonbinary treatment variable without matching. Thus, it is clear that the authors are interested in the underlying treatment variable rather than its dichotomized version. The goal of this paper is to develop a method to reliably estimate the generalized propensity score when the treatment is not binary.

3. The proposed methodology. The motivating application in Section 2 highlights the need for a methodology to estimate the propensity score for general treatment regimes. Currently, fitting a parametric model under the framework of maximum likelihood is the most commonly used method for a continuous treatment [e.g., Robins, Hernán and Brumback (2000), Hirano and Imbens (2004), Imai and van Dyk (2004)]. In this section, we first aim to improve the parametric estimation of generalized propensity score (Section 3.2). Specifically, we extend the covariate balancing propensity score (CBPS) methodology of Imai and Ratkovic (2014) to a continuous treatment and call this new methodology the Covariate Balancing Generalized Propensity Score (CBGPS). We then develop a nonparametric approach, which is refereed to as the nonparametric CBGPS (npCBGPS) (Section 3.3). The key feature of both approaches is that they estimate the generalized propensity score such that the resulting covariate balance is optimized.

3.1. *Notation and assumptions*. Suppose that we have a continuous treatment T_i for unit i whose support is $\mathcal{T} \subseteq \mathcal{R}$. Consider also observed covariates $\mathbf{X}_i \in \mathcal{R}^K$ where K is the number of pretreatment covariates. We assume a sample of observations $\{Y_i, \mathbf{X}_i, T_i\}$ for $i \in \{1, \ldots, N\}$ is drawn independently from a common joint distribution $f(Y, \mathbf{X}, T)$.

Throughout this paper, we maintain the strong ignorability and common support assumptions with respect to the original nonbinary treatment variable,

$$T_i \perp\!\!\!\perp Y_i(t) \mid \mathbf{X}_i \quad \text{and} \quad p(T_i = t \mid \mathbf{X}_i) > 0 \qquad \text{for all } t \in \mathcal{T}, \tag{1}$$

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

where $Y_i(t)$ is the potential outcome given the treatment value $T_i = t$, and $\mathbf{X}_i$ is a vector of observed pretreatment covariates. Note that the potential outcomes must be defined with respect to the original treatment variable in order to satisfy the stable unit treatment value assumption or SUTVA [Rubin (1990)]. Furthermore, the conditional distribution of treatment $p(T_i \mid \mathbf{X}_i)$ is called the *generalized propensity score* [Imbens (2000), Hirano and Imbens (2004), Imai and van Dyk (2004)]. Finally, as part of the SUTVA, we assume no interference among units. Throughout this paper, we maintain these assumptions. The main quantity of interest is the dose-response function, $\mathbb{E}(Y_i(t))$.

We begin by centering and orthogonalizing the covariates,

$$\mathbf{X}_i^* = \mathbf{S}_{\mathbf{X}}^{-1/2}(\mathbf{X}_i - \overline{\mathbf{X}}),$$

where $\overline{\mathbf{X}} = \sum_{i=1}^N \mathbf{X}_i/N$ and $\mathbf{S}_{\mathbf{X}} = \sum_{i=1}^N (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top/(N-1)$ are the sample mean vector and sample covariance matrix of $\mathbf{X}$, respectively. Similarly, we transform the treatment variable,

$$T_i^* = s_T^{-1/2}(T_i - \overline{T}),$$

where $\overline{T} = \sum_{i=1}^N T_i/N$ and $s_T = \sum_{i=1}^N (T_i - \overline{T})^2/(N-1)$ are the sample mean and variance of T, respectively. The transformed covariates $\mathbf{X}^*$ and treatment T^* thus have zero mean and unit variance. In addition, the covariates are uncorrelated with each other.

3.2. *Parametric approach.* We first consider a parametric approach by balancing covariates such that weighted correlation between $\mathbf{X}^*$ and T^* is minimized. The weight is given by $f(T_i^*)/f(T_i^* \mid \mathbf{X}_i^*)$ where the numerator is a required stabilizing factor [Robins, Hernán and Brumback (2000)]. Formally, the covariate balancing condition is given by the weighted cross moment between these centered variables,

$$\begin{aligned}\mathbb{E}\left(\frac{f(T_i^*)}{f(T_i^* \mid \mathbf{X}_i^*)} T_i^* \mathbf{X}_i^*\right) &= \int \left\{ \int \frac{f(T_i^*)}{f(T_i^* \mid \mathbf{X}_i^*)} T_i^* \, dF(T_i^* \mid \mathbf{X}_i^*) \right\} \mathbf{X}_i^* \, dF(\mathbf{X}_i^*) \\ &= \mathbb{E}(T_i^*)\mathbb{E}(\mathbf{X}_i^*) = 0.\end{aligned}$$

For the parametric CBGPS, we follow a common practice of assuming a homoskedastic linear model as done in our application [e.g., Robins, Hernán and Brumback (2000), Hirano and Imbens (2004), Imai and van Dyk (2004)]. Then the generalized propensity score is given by the following conditional normal density:

$$f_\theta(T_i^* \mid \mathbf{X}_i^*) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(T_i^* - \mathbf{X}_i^{*\top}\beta)^2\right],$$

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

where $\theta = (\beta, \sigma^2)$. In addition, we follow a typical parametric modeling approach described by Robins, Hernán and Brumback (2000) and assume the marginal distribution to be standard normal (due to centering and scaling), that is, $T_i^* \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Then the stabilizing weight is given by

$$\frac{f(T_i^*)}{f_\theta(T_i^* \mid \mathbf{X}_i^*)} = \sigma \exp\left[\frac{1}{2\sigma^2}(T_i^* - \mathbf{X}_i^{*\top}\beta)^2 - \frac{T_i^{*2}}{2}\right].$$

Under the method of moments framework, we have the following moment conditions if we combine the score condition for σ^2 and the covariate balancing conditions so that θ is just identified:

$$\mathbb{E}\{\mathbf{m}_\theta(T_i, \mathbf{X}_i)\} = \mathbb{E}\begin{pmatrix} \frac{1}{\sigma^2}(T_i^* - \mathbf{X}_i^{*\top}\beta)^2 - 1 \\ \sigma \exp\left[\frac{1}{2\sigma^2}(T_i^* - \mathbf{X}_i^{*\top}\beta)^2 - \frac{T_i^{*2}}{2}\right] T_i^* \mathbf{X}_i^* \end{pmatrix} = 0. \tag{2}$$

The estimate of θ, which we denote by $\hat{\theta}$, is obtained by numerically solving this equation.

One advantage of this parametric approach is that we can derive the asymptotic variance of the estimated causal effects by taking into account the estimation uncertainty of the generalized propensity score. This avoids the use of a more computationally intensive procedure such as bootstrap. To illustrate this feature, suppose that we wish to estimate the average causal effects via the weighted linear regression of Y_i on a set of covariates $\mathbf{Z}_i$, which may include a subset of $\mathbf{X}_i$ as well as the intercept and the treatment variable, for example, $\mathbf{Z}_i = (1, T_i, \mathbf{X}_i^\top)^\top$. The weight is given by $f(T_i^*)/f_{\hat{\theta}}(T_i^* \mid \mathbf{X}_i^*)$ where $\hat{\theta}$ is obtained via the parametric CBGPS methodology described above. If we let δ denote the vector of regression coefficients, then the moment condition is given by

$$\mathbb{E}\{\mathbf{s}_{(\theta,\delta)}(Y_i, T_i, \mathbf{X}_i)\} = \mathbb{E}\left\{\frac{f(T_i^*)}{f_\theta(T_i^* \mid \mathbf{X}_i^*)}(Y_i - \mathbf{Z}_i^\top \delta)\mathbf{Z}_i\right\} = 0. \tag{3}$$

To derive the asymptotic variance of the weighted linear least squares estimator $\hat{\delta}$, we view it as the method of moments estimator based on equations (2) and (3) [Newey and McFadden (1994), Theorem 6.1]. Then the asymptotic variance of $\hat{\delta}$ is given by

$$\begin{aligned}\mathbf{S}_\delta^{-1}\mathbb{E}\Big[&\{\mathbf{s}_{(\theta,\delta)}(Y_i, T_i, \mathbf{X}_i) - \mathbf{S}_\theta \mathbf{M}^{-1}\mathbf{m}_\theta(T_i, \mathbf{X}_i)\} \\ &\times \{\mathbf{s}_{(\theta,\delta)}(Y_i, T_i, \mathbf{X}_i) - \mathbf{S}_\theta \mathbf{M}^{-1}\mathbf{m}_\theta(T_i, \mathbf{X}_i)\}^\top\Big]\mathbf{S}_\delta^{-1^\top},\end{aligned}$$

where

$$\mathbf{S}_\delta = \mathbb{E}\left\{\frac{\partial}{\partial\delta}\mathbf{s}_{(\theta,\delta)}(Y_i, T_i, \mathbf{X}_i)\right\} = -\mathbb{E}\left\{\frac{f(T_i^*)}{f_\theta(T_i^* \mid \mathbf{X}_i)}\mathbf{Z}_i\mathbf{Z}_i^\top\right\},$$

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

$$\begin{aligned}
\mathbf{S}_\theta &= \mathbb{E}\left\{\frac{\partial}{\partial\theta}\mathbf{s}_{(\theta,\delta)}(Y_i, T_i, \mathbf{X}_i)\right\} \\
&= \mathbb{E}\left(-\frac{1}{\sigma^2}\frac{f(T_i^*)}{f_\theta(T_i^* \mid \mathbf{X}_i)}(T_i^* - \mathbf{X}_i^{*\top}\beta)(Y_i - \mathbf{Z}_i^\top\delta)\mathbf{Z}_i\mathbf{X}_i^{*\top}\right. \\
&\qquad \left.\frac{1}{2\sigma^2}\frac{f(T_i^*)}{f_\theta(T_i^* \mid \mathbf{X}_i)}\left\{1 - \frac{1}{\sigma^2}(T_i^* - X_i^{*\top}\beta)^2\right\}(Y_i - \mathbf{Z}_i^\top\delta)\mathbf{Z}_i\right), \\
\mathbf{M} &= \mathbb{E}\left\{\frac{\partial}{\partial\theta}\mathbf{m}_\theta(T_i, \mathbf{X}_i)\right\} \\
&= \begin{pmatrix} -\frac{2}{\sigma^2}(T_i^* - \mathbf{X}_i^{*\top}\beta)\mathbf{X}_i^{*\top} \\ -\frac{1}{\sigma^2}\frac{f(T_i^*)}{f_\theta(T_i^* \mid \mathbf{X}_i)}T_i^*(T_i^* - \mathbf{X}_i^{*\top}\beta)\mathbf{X}_i^*\mathbf{X}_i^{*\top} \\ -\frac{1}{\sigma^4}(T_i^* - \mathbf{X}_i^{*\top}\beta)^2 \\ \frac{1}{2\sigma^2}\frac{f(T_i^*)}{f_\theta(T_i^* \mid \mathbf{X}_i)}\left\{1 - \frac{1}{\sigma^2}(T_i^* - \mathbf{X}_i^{*\top}\beta)^2\right\}T_i^*\mathbf{X}_i^* \end{pmatrix}.
\end{aligned}$$

Finally, note that the asymptotic variance of $\hat{\theta}$ is given by

$$\text{avar}(\hat{\theta}) = (\mathbf{M}^\top\mathbf{M})^{-1}\mathbf{M}^\top\mathbb{E}\{\mathbf{m}_\theta(T_i, \mathbf{X}_i)\mathbf{m}_\theta(T_i, \mathbf{X}_i)^\top\}\mathbf{M}(\mathbf{M}^\top\mathbf{M})^{-1}. \tag{4}$$

In practice, the estimation of this asymptotic variance in a finite sample may suffer from numerical instability with $\mathbf{M}$ being near singular. To address this issue, we add a small constant, for example, 0.01, to the diagonal elements of $\mathbf{M}$ whenever the ratio of smallest and largest eigenvalues of $\mathbf{M}$ falls below a certain threshold.

3.3. *Nonparametric approach*. We next consider a nonparametric extension of the CBGPS we refer to as npCBGPS. This method does not involve direct estimation of the generalized propensity score, and thus does not require the model to be correctly specified. Rather, we use an empirical likelihood approach to chose weights that represent the stabilizing inverse generalized propensity score, and simultaneously ensure balancing conditions (zero correlation with the treatment) are met in the sample. Owing to the potential need for extreme weights to achieve these conditions in some samples, we further develop the method to allow some degree of finite sample imbalance. This makes the approach suitable in a wide range of conditions in which the investigator would prefer not to choose a functional form for the propensity score, but at a computational cost.

3.3.1. *The formulation*. We begin by defining the stabilizing weight as

$$w_i = \frac{f(T_i^*)}{f(T_i^* \mid \mathbf{X}_i^*)}, \tag{5}$$

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

where no parametric restriction is placed on the generalized propensity score $f(T_i^* \mid \mathbf{X}_i^*)$, nor on the marginal distribution of treatments $f(T_i^*)$. This stabilizing weight is already normalized in that its expectation taken over joint density $f(T_i^*, \mathbf{X}_i^*)$ equals unity:

$$\mathbb{E}(w_i) = \int\int \frac{f(T_i^*)}{f(T_i^* \mid \mathbf{X}_i^*)} f(T_i^*, \mathbf{X}_i^*)\, dT_i^*\, d\mathbf{X}_i^* = 1. \tag{6}$$

In the current framework, the covariate balancing conditions are derived such that after weighting with w_i, T_i^* and $\mathbf{X}_i^*$ are uncorrelated (hence, the original variables, T_i and $\mathbf{X}_i$, are also uncorrelated). Specifically, we have shown that the covariate balancing conditions is equal to

$$\mathbb{E}(w_i T_i^* \mathbf{X}_i^*) = \mathbb{E}(T_i^*)\mathbb{E}(\mathbf{X}_i^*) = 0. \tag{7}$$

Similarly, it can be shown that weighting with w_i also preserves the marginal means of $\mathbf{X}_i^*$ and T_i^*. This provides two additional covariate balancing conditions, $\mathbb{E}(w_i \mathbf{X}_i^*) = \mathbb{E}(\mathbf{X}_i^*) = 0$ and $\mathbb{E}(w_i T_i^*) = \mathbb{E}(T_i^*) = 0$. Altogether, the constraints on the mean of w_i, on the marginal means of $\mathbf{X}^*$ and T^*, and on the crossproducts $\mathbf{X}^* T^*$ give rise to the sample conditions,

$$\sum_{i=1}^{N} w_i g(\mathbf{X}_i^*, T_i^*) = 0 \quad \text{and} \quad \sum_{i=1}^{N} w_i - N = 0, \tag{8}$$

where $g(\mathbf{X}_i^*, T_i^*) = (\mathbf{X}_i^*, T_i^*, \mathbf{X}_i^* T_i^*)^\top$, whose dimensionality is $2K + 1$. Although we do not discuss it in detail, in general, categorical and multidimensional treatments can be accommodated analogously within this framework.

We now choose weights w_i that satisfy the moment conditions given in equation (8) while maximizing the empirical likelihood of observing the data. That is, through equation (5), we can express the joint density of each observation in relation to the weights as $f(T_i^*, \mathbf{X}_i^*) = \frac{1}{w_i} f(T_i^*) f(\mathbf{X}_i^*)$. The likelihood function for the whole sample is thus $\prod_{i=1}^{N} f(T_i^*, \mathbf{X}_i^*) = \prod_{i=1}^{N} \frac{1}{w_i} f(T_i^*) f(\mathbf{X}_i^*)$. We wish to maximize the empirical likelihood of the data by choosing w_i, but also require w_i to satisfy the constraints in 8 above [Owen (2001)]. Thus, we maximize

$$\prod_{i=1}^{N} f(T_i^*, \mathbf{X}_i^*) = \prod_{i=1}^{N} \frac{1}{w_i} f(T_i^*) f(\mathbf{X}_i^*)$$

subject to the following constraints:

$$\sum_{i=1}^{N} w_i g(\mathbf{X}_i^*, T_i^*) = 0, \qquad \sum_{i=1}^{N} w_i = N, \qquad \sum_{i=1}^{N} w_i \mathbf{X}_i^* T_i^* = 0$$

and

$$w_i > 0 \qquad \text{for all } i.$$

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

This is equivalent to maximizing

$$\sum_{i=1}^{N} \log f(T_i^*) + \log f(\mathbf{X}_i^*) - \log w_i$$

in which only the final term involves w_i. Therefore, the estimation of the npCBGPS reduces to the simply finding

$$\underset{w \in \mathbb{R}^N}{\operatorname{argmin}} \sum_{i=1}^{N} \log w_i$$

subject to the above constraints. We note that w_i chosen this way estimates stabilizing inverse (generalized) propensity scores weights, with the desired covariate balancing properties built in through the constraints.

3.3.2. *The numerical algorithm.* We follow an approach similar to the standard Lagrange multiplier technique for numerically solving this optimization problem [Owen (2001)]. Construct the Lagrangian,

$$\mathcal{L}(w_i, \lambda, \gamma) = \sum_{i=1}^{N} \log w_i + \lambda \left(N - \sum_{i=1}^{N} w_i \right) + \gamma^\top \sum_{i=1}^{N} w_i g(\mathbf{X}_i^*, T_i^*),$$

where λ and γ are Lagrange multipliers. The first-order conditions are given by

$$\frac{\partial \mathcal{L}}{\partial w_i} = \frac{1}{w_i} - \lambda + \gamma^\top g(\mathbf{X}_i^*, T_i^*) = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = N - \sum_{i=1}^{N} w_i = 0, \qquad \frac{\partial \mathcal{L}}{\partial \gamma} = \sum_{i=1}^{N} w_i g(\mathbf{X}_i^*, T_i^*) = 0.$$

We now sum $w_i \frac{\partial \mathcal{L}}{\partial w_i} = 0$ over i to obtain $\lambda = 1$. Plugging this into $\frac{\partial \mathcal{L}}{\partial w_i}$ and solving for w_i yields

$$w_i = \frac{1}{1 - \gamma^\top g(\mathbf{X}_i^*, T_i^*)}.$$

Thus, the minimization of $\sum_{i=1}^{N} \log w_i$ can be done over γ, which has only $(2K+1)$ dimensions rather than N. As a result, our constrained optimization problem is solved by the unconstrained maximization,

$$\underset{\gamma \in \mathbb{R}^K}{\operatorname{argmax}} \sum_{i=1}^{N} \log(1 - \gamma^\top g(\mathbf{X}_i^*, T_i^*)).$$

This optimization is relatively straightforward, and is well handled by the standard BFGS procedure. At the solution, if it exists, γ corresponds to nonnegative values of w_i. However, during the optimization, the argument inside the logarithmic

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

function can be nonpositive. To handle this, when the argument to the logarithmic function falls below $1/N$, we instead use the second order Taylor series approximation to the log around the point $1/N$. Once a solution is reached, we evaluate w_i by using the original formula, $1/\{1 - \gamma^\top g(\mathbf{X}_i^*, T_i^*)\}$.

Since the the empirical likelihood is not generally convex, there is no guarantee that the optimization procedure described here finds the global optimum. This contrasts with some other methods such as entropy balancing and stable weights that solve a convex optimization problem [Hainmueller (2012), Zubizarreta (2015)]. Although these methods are not based on likelihood inference, they may still possess appealing statistical properties [Zhao and Percival (2017)]. Future research may consider the extension of these methods to a continuous treatment.

3.3.3. *A penalized imbalance approach.* The flexibility of the nonparametric approach comes with a cost. In many practical situations, the numerical algorithm described above fails to find a solution. This failure occurs especially when the number of covariates is large and/or the treatment T_i is strongly predicted by X_i. In such cases, we may wish to avoid forcing the covariate balancing conditions given in equation (7) to hold exactly in sample. Instead, we allow some finite sample correlation, but penalize the degree of sample imbalance that remains.

Specifically, we consider the sample weighted correlation, $\eta = \frac{1}{N}\sum_{i=1}^N w_i \times \mathbf{X}_i^* T_i^*$. The above empirical likelihood approach maximized $f(T_i^*, \mathbf{X}_i^* \mid \eta)$ where $\eta = 0$. However, replacing our original constraint that $\sum_{i=1}^n w_i \mathbf{X}_i^* T_i^* = 0$ with the constraint that $\sum_{i=1}^n w_i \mathbf{X}_i^* T_i^* - \eta = 0$ allows us to obtain and maximize the empirical likelihood conditional on any other level of weighted sample correlation, η. Conditional on η, the Lagrangian for the likelihood maximization problem is then

$$\mathcal{L}(w_i, \lambda, \gamma \mid \eta) = \sum_{i=1}^N \log w_i + \lambda\left(N - \sum_{i=1}^N w_i\right) + \gamma^\top\left(\sum_{i=1}^n w_i \mathbf{X}_i^* T_i^* - \eta\right).$$

The optimization of the dual is then

$$\underset{\gamma \in \mathbb{R}^K}{\operatorname{argmax}} \sum_{i=1}^N \log\bigl[1 - \gamma^\top\bigl(g(\mathbf{X}_i^*, T_i^*) - \eta\bigr)\bigr]. \tag{9}$$

We do, however, maintain the exact constraint that $\frac{1}{N}\sum_{i=1}^N T_i^* w_i = 0$ and $\sum_i^N \mathbf{X}_i^* w_i = 0$ because these are centering choices that are not costly and without which the constraints $\sum_{i=1}^N w_i \mathbf{X}_i^* T_i^* = 0$ would no longer correspond to the zero-correlation conditions.

What remains is to determine the appropriate penalty for a given level of imbalance, η, so that we can choose both w_i and η according to a single penalized optimization objective. To motivate our choice of penalty, we consider the likelihood of jointly observing both the data and the selected level of finite sample

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

imbalance, $f(T_i^*, \mathbf{X}_i^*, \eta)$, which factors into $f(T_i^*, \mathbf{X}_i^* \mid \eta) f(\eta)$. The first component, $f(T_i^*, \mathbf{X}_i^* \mid \eta)$, is given as above by empirical likelihood. For $f(\eta)$, we assume

$$\eta = \frac{1}{N} \sum_{i=1}^{N} w_i \mathbf{X}_i^* T_i^* \sim \mathcal{N}_K(0, \rho \mathbf{I}_K),$$

where $\rho > 0$ is a tuning parameter or penalty that can be set by the investigator to determine how severely finite sample imbalances should be penalized. We discuss the appropriate choice of ρ in Section 3.3.4 below.

Having given $f(\eta)$ this form, we can now choose w_i so as to maximize $f(\mathbf{X}^*, T^* \mid \eta) f(\eta)$ over the sample. The log (penalized) likelihood maximization problem becomes

$$\underset{w \in \mathbb{R}^N, \eta \in \mathbb{R}^K}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \log w_i - \log f(\eta) \right] = \underset{w \in \mathbb{R}^N, \eta \in \mathbb{R}^K}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \log w_i + \frac{1}{2\rho} \eta^\top \eta \right]$$

subject to the modified balance constraints, $\sum_{i=1}^{N} w_i \mathbf{X}_i^* T_i^* = \eta$ and the additional constraints as before, that is, $\sum_{i=1}^{N} w_i - N = 0$, $\sum_{i=1}^{N} w_i \mathbf{X}^* = 0$, and $\sum_{i=1}^{N} w_i T_i^* = 0$.

The optimization is now more difficult because of the additional parameter η, which is multidimensional. We first consider the optimization with respect to w_i given the value of η, which is given by equation (9). For the "outer" optimization over η, we initialize η_0 to the unweighted correlation of $\mathbf{X}_i^*$ and T_i^*. This returns a solution with all equal weights. We then reparameterize η as $\alpha \eta_0$ for the scalar α, and then line search over $\alpha \in [0, 1]$,

$$\underset{w \in \mathbb{R}^N, \alpha \in [0,1]}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \log w_i + \frac{1}{2\rho} (\alpha \eta_0)^\top (\alpha \eta_0) \right]. \tag{10}$$

We thus do not search all possible values of η, but rather those that correspond to equal proportional imbalance reductions.

3.3.4. *Choice of* ρ. We note that by the central limit theorem, for $\mathbf{X}_i^*$ and T_i^* that have zero correlation in expectation, the distribution of finite sample imbalance one would see, ignoring weights, would be $\mathcal{N}_K(0, \frac{1}{N} \mathbf{I}_K)$. Thus, $\rho = \frac{1}{N}$ would seem the correct choice. However, the weights complicate this and without further assumptions, this result does not hold on the weighted correlation. Moreover, our general aim in developing covariate balancing scores is to achieve better balance (lower correlation) than one expects by chance alone had $\mathbf{X}_i^*$ and T_i^* been uncorrelated in expectation.

Thus, we consider ρ a tuning parameter which the investigator can manipulate [using the `corprior` argument in the `npCBGPS()` function in the `CBPS` package]. This can be increased to ensure the allowable finite sample imbalance

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

is large enough to prevent extreme weights, while being small enough to result in weights that achieve fine balance. However, as a starting point, we provide a default option of $\rho = 0.1/N$. We choose this because it implies better balance than would be expected by chance, and because in practice it has provided a reasonable tolerance and generally allows convergence. To interpret this value, suppose we had only 10 observations. Then, setting $\rho = 0.1/N = 0.01$ implies that we expect the finite sample correlation of T_i and $\mathbf{X}_i$ (that are uncorrelated in the population) to have a distribution such that 95% of the time it falls in $[-1.96\sqrt{\rho}, 1.96\sqrt{\rho}] = [-0.196, 0.196]$. At $N = 100$, this default ρ implies that 95% of the time we expect to find a finite sample correlation in the range $[-0.062, 0.062]$. By making the allowable finite imbalance somewhat generous but lower than we expect by chance alone given uncorrelated treatment and covariates, this choice ensures the algorithm converges almost without exception while improving balance. Once run, researchers may check balance and opt to reduce ρ to obtain finer balance. If ρ is set too small and convergence fails, the weights will no longer sum to one, and balance will be poor.

4. Simulation studies. To examine the finite sample properties of the proposed estimators, we conduct simulation studies under four different scenarios. We vary whether the true treatment assignment is correctly specified, and whether the data generating process for the outcome is linear in the covariates or not. Our guiding motivation in designing the simulation settings is to use simple data generating processes that reveal the failings of each estimator, while still replicating the fundamental difficulties researchers will face in practice. For example, when a nonlinearity is required in some simulation settings below, a very simple one proves sufficient in which a measured covariate (e.g., X) enters the treatment (and outcome) model not linearly but as $(X + c)^2$ for a small constant c.

We are also sensitive to the concern that investigators will often attempt to adjust for many or all available measured pretreatment covariates, of which only some are actually important. Our aim of obtaining balance on all covariates can thus be costly, as some of these covariates may actually be irrelevant. We thus ensure this cost is borne in the simulations shown below. Specifically, all simulations are run assuming the investigator has 10 (correlated) covariates to consider, though only up to five appear in the true treatment assignment model. Moreover, only four of these appear in the true outcome model, meaning that some variables that are deliberately imbalanced due to their role in treatment assignment will actually be irrelevant in the outcome model. This ensures that we pay a penalty for methods that expend effort to balance all of these covariates, despite some of them not explicitly influencing the outcome.

4.1. *Four data generating processes.* For all data generating processes, $K = 10$ covariates are drawn independently from a multivariate normal distribution with mean 0, variance 1, and covariances of 0.2, though not all these covariates

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

appear in the treatment or outcome equations. In the first simulation setting, we assume that both the treatment assignment and outcome models are linear in the covariates as given, with true data generating process given by

$$T_i = X_{i1} + X_{i2} + 0.2X_{i3} + 0.2X_{i4} + 0.2X_{i5} + \xi_i, \tag{11}$$

$$Y_i = X_{i2} + 0.1X_{i4} + 0.1X_{i5} + 0.1X_{i6} + T_i + \varepsilon_i, \tag{12}$$

where $\xi_i \overset{\text{i.i.d.}}{\sim} N(0, 4)$ and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 25)$ are error terms, and the true average treatment effect is set to 1. Note that because the outcome model is linear in covariates, weights that achieve mean covariate balance will be sufficient for the (weighted) difference in means estimator to be unbiased for the average treatment effect.

Under the second simulation setting, we introduce misspecification of the treatment assignment model by including a nonlinear term in the true data generating process,

$$T_i = (X_{i2} + 0.5)^2 + 0.4X_{i3} + 0.4X_{i4} + 0.4X_{i5} + \xi_i,$$

where $\xi_i \sim N(0, 2.25)$ and while the outcome equation remains the same as equation (12).

The third simulation setting correctly specifies the treatment assignment model, returning to equation (11). However, it now uses an outcome that is not linear in the covariates as given. Specifically,

$$Y_i = 2(X_{i2} + 0.5)^2 + T_i + 0.5X_{i4} + 0.5X_{i5} + 0.5X_{i6} + \varepsilon_i.$$

Finally, the fourth simulation setting uses the nonlinear data generating process for the treatment, ensuring misspecification of the treatment assignment in later analysis, while also including a nonlinearity in the outcome data generating process:

$$T_i = (X_{i2} + 0.5)^2 + 0.4X_{i3} + 0.4X_{i4} + 0.4X_{i5} + \xi_i,$$

$$Y_i = 2(X_{i2} + 0.5)^2 + T_i + 0.5X_{i4} + 0.5X_{i5} + 0.5X_{i6} + \varepsilon_i.$$

4.2. *Results*. For each of the above four settings, we run 500 independent Monte Carlo simulations and examine the covariate balance across these replications. We try five weighting approaches for purposes of examining the resulting balance and ATE estimates. First, we use the original, unweighted observations (`Unweighted`). Then we use methods the derive weights from the maximum likelihood estimate of the generalized propensity score (`MLE`), from the exactly-identified CBGPS, the nonparametric npCBGPS, and finally the gradient-boosting approach of Zhu, Coffman and Ghosh (2015) (`GBM`), which selects the number of trees such that covariate balance is optimized. For the `GBM`, we present the results based on the Pearson correlation coefficients as the measures of covariate balance so that it is comparable to our methods.

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

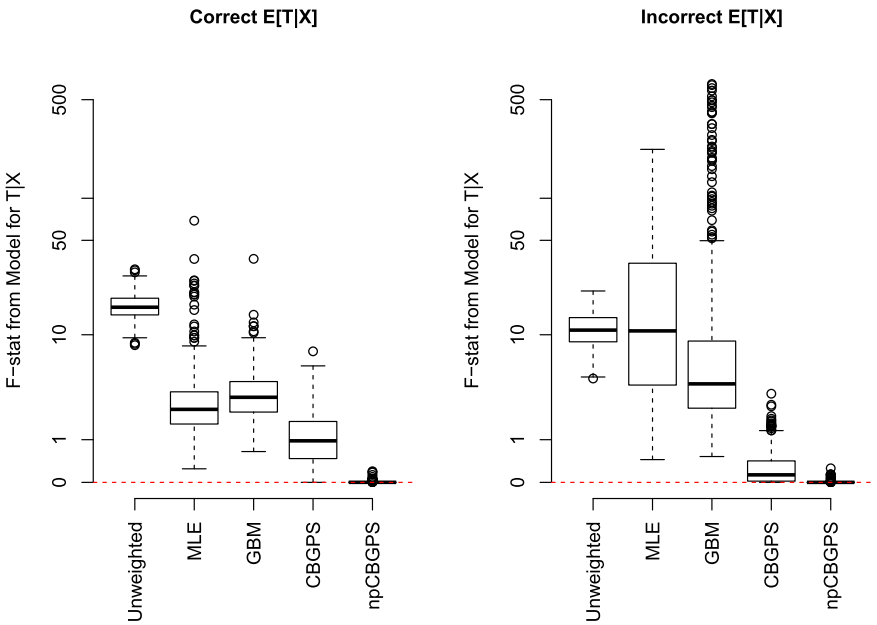

FIG. 1. *Covariate balance for simulation studies*: *F-statistics obtained from the regression of* T_i *on* $\mathbf{X}_i$ *with weights determined by each method. The* `MLE` *estimator achieves unreliable balance even when the treatment assignment is correctly specified* (*left*), *and is highly unstable when treatment assignment is misspecified* (*right*). *By contrast, weighting with either* CBGPS *or* npCBGPS *produces excellent balance* (*F-stats near zero*) *on nearly every iteration of every simulation scenario. The* `GBM` *methods provides little control over imbalance.*

Figure 1 visualizes the degree of covariate balance achieved by each method when applied to data from either the correctly specified (*left*) or incorrectly specified (*right*) treatment assignment. The plots show the distribution of F-statistics obtained from the regression of T_i on $\mathbf{X}_i$, to give a global summary of covariate balance, across the 500 simulations. We note the limitation that, because we linearly regress T_i on $\mathbf{X}_i$ to produce the F-statistics shown, they only indicate the quality of balance on the covariate means, and not on their higher or multivariate moments.

The estimates using weights based on `MLE` are not reliable especially when the treatment assignment is misspecified (*right*). In this case, `MLE` makes balance far worse on some iterations than it was without adjustment (`Unweighted`). By contrast, weighting with either CBGPS or npCBGPS produces F-statistics very close to zero on nearly every iteration of every simulation scenario. These proposed methods also outperform `GBM`. The poor balance achieved by `GBM` relative to the CBGPS and npCBGPS methods is not surprising: the `GBM` approach attempts to optimize balance by estimating a propensity score, but the only scope for improving balance is in the choice of how many trees are employed by the gradient boosting algorithm. This does not provide as direct control over finite sample

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

imbalance as can be obtained by directly satisfying balance constraints as in the CBGPS and npCBGPS methods. Under an incorrectly specified propensity score, GBM outperforms MLE, but the balance remains poor and highly unstable—in many cases worse than using no weights at all.

Covariate balance is important because imbalances may lead to biased causal effect estimates. Figure 2 shows the distribution of the estimated ATE. Here, we use a (weighted) difference in means estimator with the weights determined by each method. The uncertainty resulting from each procedure is apparent in the variability of estimates across Monte Carlo replicates. The dotted horizontal line shows the true ATE of 1.

We find, first, that the distribution of unweighted difference in means estimate (Unweighted) has no overlap with the truth under any of the four simulations. Second, when both the treatment assignment and outcome models are misspecified, all methods fail although the bias is particularly severe for the estimates based on MLE (*bottom right*). Third, adjustment by MLE produces low bias estimates whenever the treatment assignment is correctly specified, regardless of whether the outcome is also correctly specified (*top left*) or incorrectly specified (*bottom left*). However, the MLE procedure generates more widely varying estimates than either CBGPS or npCBGPS. This is unsurprising, as we saw that balance was not as finely controlled by MLE. Moreover, when the outcome is linear in $\mathbf{X}$ but the treatment assignment is misspecified, both CBGPS and npCBGPS are able to recover good estimates, while MLE fails widely (*top right*). This is the expected behavior. The balancing criteria added to npCBGPS and CBGPS allow it to circumvent misspecified treatment assignments.

We also find that GBM performs well when only the propensity score is misspecified, as the GBM technique is sufficiently flexible to still estimate a reasonable propensity score. The improved covariate balance achieved by the CBGPS and npCBGPS under misspecified propensity scores helps with the estimation of ATE, but only produces unbiased estimates when the outcome is linear in $\mathbf{X}$ because balance is achieved only on the covariate means and not necessarily on higher moments. While GBM is able to estimate the propensity score reasonably well even when it is nonlinear in $\mathbf{X}$, this is evidently not enough to ensure good estimates when the outcome model is also misspecified. We note however that GBM tends to produce less biased ATE estimates than the npCBGPS and CBGPS in the doubly-misspecified case, when the sample size is much larger. For example, while $N = 200$ in this exercise, at $N = 1000$, the bias of GBM drops to approximately 0.5, about half that of CBGPS. With a large sample size, GBM appears to more accurately estimate the generalized propensity score.

Finally, using the first data generating process with no misspecification (where we expect the least bias), we examine the coverage rate of confidence intervals constructed using the asymptotic variance formula given in equation (4). After 10,000 iterations, we find that the coverage rate of the resulting 95% confidence interval is quite accurate at 95.5%.

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

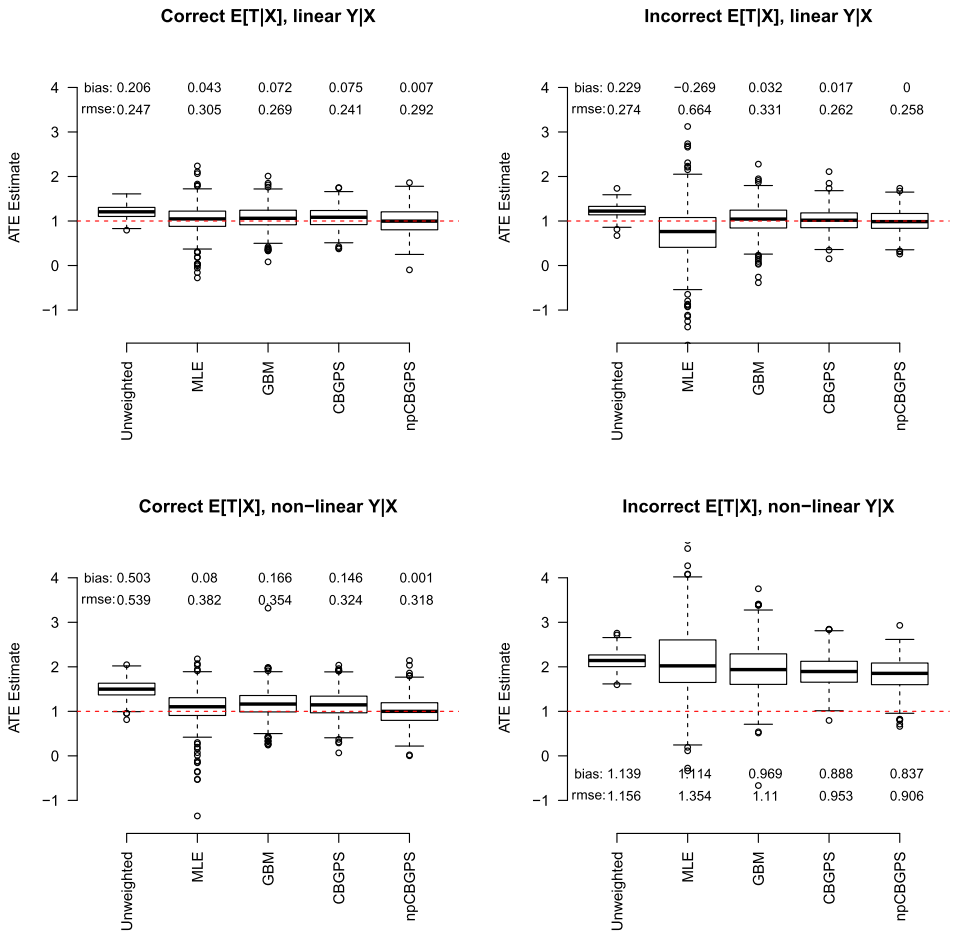

FIG. 2. *Difference in means estimates of ATE using weights determined by each method. True ATE* (1) *shown by dotted horizontal line. The simple difference in means estimate* (`Unweighted`) *has no overlap with the truth under any of the four simulations. Adjustment by* `MLE` *produces low bias estimates whenever the treatment assignment is correctly specified* (*top left and bottom left*), *but with greater uncertainty than* CBGPS *or* npCBGPS. *When the outcome is linear in* **X** *but the treatment assignment is misspecified, both* CBGPS *and* npCBGPS *are able to recover good estimates, while* `MLE` *fails substantially* (*top right*). *However, when the outcome is also nonlinear in* **X** (*bottom right*) *all methods fail.* `GBM` *generally performs similarly to* CBGPS *while* npCBGPS *generally outperforms it slightly.*

5. Empirical application. We apply the proposed CBGPS methodology to the observational study described in Section 2. The treatment variable, that is, the number of political advertisements aired in each zip code, has a skewed distribution in which 12.1% of the observations are 0. To make our assumption that T^* is normally distributed more reasonable, though far from perfect, we search across

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

Box–Cox transformations of the form $\{(T+1)^{\lambda}-1\}/\lambda$ [taking $\log(T+1)$ when $\lambda = 0$] to find a transformation of the treatment whose marginal distribution is the closest to the standard normal. We use $\lambda = -0.16$, which yields the greatest correlation between the sample quantiles of the transformed treatment variable and the corresponding theoretical quantiles of the standard normal distribution.

The pretreatment covariates $\mathbf{X}$ in the generalized propensity score model include the log population, population density, log median income, percent Hispanic, percent black, percent over age 65, percent college graduates, and a binary indicator of whether it is possible to commute to the zip code from a competitive state. This list includes all of the variables used in the original analysis as well as several variables available in the author's data set that were not included in the original propensity score estimation. We add the squares of all nonbinary pretreatment covariates to the model in order to balance both their first and second moments. The outcome model includes the treatment variable (on its transformed scale), the square of the treatment variable, and unit fixed effects for states.

Figure 3 shows two metrics of covariate balance. The left plot shows the Pearson correlations between each covariate (including the square terms) and the transformed treatment variable in the original unweighted sample, after propensity score matching on the dichotomized treatment variable in which the logistic regression is used to estimate the propensity score (as in the original analysis), and after weighting based on the estimated generalized propensity score in four ways (`MLE`, `GBM`, CBGPS, and npCBGPS). Matching based on the dichotomized treatment variable (second left boxplot) only slightly improves the covariate balance with respect to the original treatment variable (far left). Weighting based on the `MLE` of the generalized propensity score (middle) makes the covariate balance far worse than in the original sample. While weighting based on the `GBM` substantially improves covariate balance, both parametric CBGPS (second right) and npCBGPS (far right) virtually eliminates the imbalance.

The right plot presents the F-statistics calculated by regressing the transformed treatment variable on each pre-treatment covariate one at a time. The pattern is essentially the same as the one for correlation. Using all covariates in a single regression, the F-statistic is 29.3 in the original sample, 38.3 in the post-matching sample, 215.3 with `MLE` weighting, 2.60 with `GBM` weighting, 9.33×10^{-5} with parametric CBGPS weighting, and 0.406 with nonparametric CBGPS weighting.

Table 1 shows the absolute Pearson correlation balance metric on a variable-by-variable basis using the original scale. Even though CBGPS and npCBGPS orthogonalize the covariates before optimizing balance, they still achieve superior balance for every covariate on their original scales. In contrast, `MLE` worsens the covariate balance for virtually all variables. While `GBM` improves balance for each variable, the degree of improvement is less than npCBGPS and CBGPS, which virtually eliminate the imbalance.

We use the bootstrap to obtain confidence intervals for the average dose response while incorporating uncertainty over the choice of weights. Each of the 5000 bootstrap replicates finds the best Box–Cox transformation for the outcome,

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

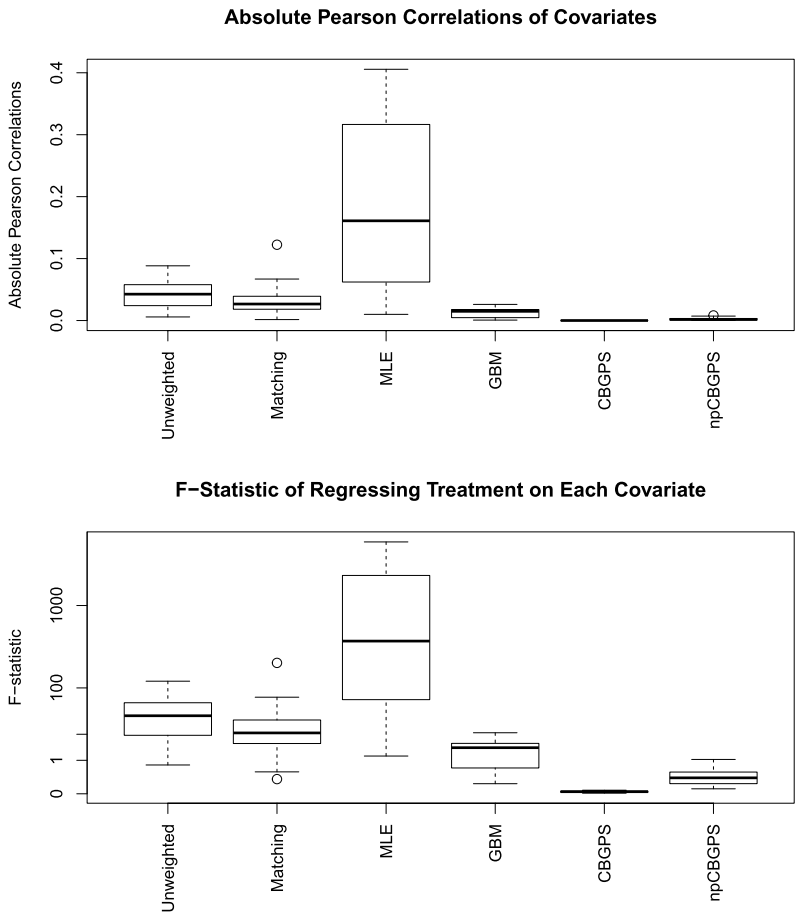

FIG. 3. *Two measures of covariate imbalance in the Urban and Niebler* (2014) *data. The figure presents the absolute Pearson correlation between the treatment and each covariate after weighting as well as the F-statistic from the regression of the treatment on each covariate after weighting* (*fourth root scale*). *Weighting via MLE* (*middle left*) *yields worse covariate balance than the original unweighted sample* (*far left*) *or the matched sample using the dichotomized treatment* (*second left*). *Weighting via GBM* (*middle right*) *yields better balance, but weighting via* CBGPS (*second right*) *and* npCBGPS (*far right*) *improves the covariate balance most.*

fits the treatment model for the bootstrapped sample using each of the four procedures, and estimates the outcome model with these weights. The outcome model regresses contributions on the treatment, the square of the treatment, and state-level fixed effects (to approximate the within-state matching employed by Urban and Niebler). The covariate imbalance in these bootstrap iterations follows the same pattern as in the full sample. The interquartile range for the F-statistic in the full post-weighting treatment model is $(11.00, 45.88)$ for `MLE`, $(2.61, 4.62)$ for `GBM`, $(7.4 \times 10^{-5}, 3.31)$ for parametric CBGPS, and $(0.44, 0.99)$ for npCBGPS.

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

TABLE 1

Variable-by-variable comparison of the absolute Pearson correlations presented in Figure 3. Maximum likelihood estimation (MLE) makes the covariate imbalance worse for virtually every variable. While GBM improves balance for each variable, the degree of improvement is less than npCBGPS *and* CBGPS*, which virtually eliminates the imbalance*

	Unweighted	**MLE**	**GBM**	**CBGPS**	**npCBGPS**
log(Population)	−0.059	−0.034	0.016	0.000	−0.001
% Over 65	0.006	−0.162	−0.004	−0.000	0.000
log(Income + 1)	−0.021	−0.384	0.014	−0.000	−0.001
% Hispanic	−0.043	0.053	0.007	0.000	−0.002
% Black	−0.076	0.295	−0.003	0.000	0.003
Population density	−0.088	0.405	0.016	−0.000	0.008
% College graduates	−0.032	−0.145	0.018	−0.000	0.004
Can commute	0.054	0.161	0.027	−0.000	0.003
log(Population)^2	−0.057	−0.049	0.018	0.000	−0.000
% Over 65^2	0.010	−0.071	−0.001	0.000	−0.001
log(Income + 1)^2	−0.028	−0.338	0.018	−0.000	−0.001
% Hispanic^2	−0.013	−0.010	0.006	0.000	0.001
% Black^2	−0.057	0.291	−0.007	0.000	0.003
Population density^2	−0.072	0.406	0.003	−0.000	0.003
% College graduates^2	−0.028	−0.079	0.022	0.000	0.007

Table 2 shows the estimated average effect of 1000 ads on campaign contributions as well as its standard error and 95% confidence interval (the same dose that was investigated in Urban and Niebler's dichotomized analysis). For the sake of comparison, the original results based on dichotomized matching are presented, showing that the estimated average effect is positive and statistically significant. Note, however, that the Abadie and Imbens (2006) standard errors do not account for the uncertainty regarding propensity score estimation (although they do ac-

TABLE 2

The estimated effect of 1000 *political advertisements on campaign contributions. The standard errors from dichotomized matching (in the first row) are obtained from the Abadie and Imbens* (2006) *standard errors, which, unlike the other procedures, do not account for uncertainty in the estimates of the propensity scores. The standard errors and confidence intervals for the other estimates are based on* 5000 *bootstrap replicates*

Method	**Estimate**	**Standard error**	**95% confidence interval**
Matching (original)	6800	1655	(3556, 10,043)
MLE	477	4629	(−345, 17,532)
GBM	11,162	2555	(6105, 16,095)
CBGPS	4935	3865	(−1032, 13,989)
npCBGPS	6518	3668	(−415, 13,840)

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

count for the uncertainty of the matching procedure). The standard errors and 95% confidence intervals for the other estimates are based on 5000 bootstrap replicates.

We have already shown that the propensity score estimated by MLE produced poor balance in this sample, so its estimates may be severely biased. The point estimate based on MLE is much smaller than those based on the other estimates, and its bootstrap distribution is quite skewed due to the existence of extreme weights. The weighting estimate based on GBM, which achieves a moderate level of balance, yields a point estimate that is far larger than any other estimator and is statistically significant. In contrast, the estimates based on the CBGPS and npCBGPS, both of which achieve excellent covariate balance, are of moderate magnitude and yield the 95% confidence intervals that contain zero. Finally, the estimates based on the CBGPS and npCBGPS have narrower confidence intervals and smaller standard errors than MLE does. Thus, these methods appear to yield more efficient estimates than the standard MLE method.

One advantage of CBGPS is that it is possible to obtain standard errors that account for uncertainty in the estimation of weights without relying on a computationally intensive method such as bootstrap. In the current case, we obtain the asymptotic 95% confidence interval of $(-2028, 11{,}898)$ with the standard error of 3552. This is quite similar to the bootstrap confidence interval, but it can be computed much more quickly.

6. Concluding remarks. Despite advances in generalizing propensity score methods to nonbinary treatments, applied researchers often dichotomize nonbinary treatment variables in order to utilize propensity score methods. One reason for this gap between statistical theory and practice is the absence of a reliable method for estimating the generalized propensity score. In this paper, we extend the covariate balancing propensity score (CBGPS) of Imai and Ratkovic (2014) to a continuous treatment. We estimate the generalized propensity score such that the resulting covariate balance is optimized. Our empirical analyses show that the proposed methodology results in better covariate balance than the standard method and can yield substantive insights which may be difficult to obtain by analyzing a dichotomous treatment. We also find that CBGPS reduces sensitivity to misspecification of the generalized propensity score model. Finally, we consider a nonparametric extension of the CBGPS methodology based on maximizing the empirical likelihood of the data given the desired moment constraints. While computationally more demanding, this method avoids the distributional assumptions made in the parametric CBGPS approach.

Acknowledgments. The proposed method is implemented through open-source software CBPS [Fong et al. (2017)], which is freely available as an R package at the Comprehensive R Archive Network (CRAN https://cran.r-project.org/package=CBPS). We thank Marc Ratkovic and Dylan Small for their comments and suggestions. The replication archive is available as Fong, Hazlett and Imai (2017).

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

REFERENCES

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325

BOYD, C. L., EPSTEIN, L. and MARTIN, A. D. (2010). Untangling the causal effects of sex on judging. *Amer. J. Polit. Sci.* **54** 389–411.

CHAN, K. C. G., YAM, S. C. P. and ZHANG, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 673–700. MR3506798

DE, P. K. and RATHA, D. (2012). Impact of remittances on household income, asset, and human capital: Evidence from Sri Lanka. *Migr. Dev.* **1** 163–179.

DONOHUE III, J. J. and HO, D. E. (2007). The impact of damage caps on malpractice claims: Randomization inferences with difference-in-differences. *J. Empir. Leg. Stud.* **4** 69–102.

FAN, J., IMAI, K., LIU, H., NING, Y. and YANG, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. Technical report, Princeton Univ.

FONG, C., HAZLETT, C. and IMAI, K. (2018). Replication data for: Covariate balancing propensity score for a continuous treatment. DOI:10.7910/DVN/AIF4PI.

FONG, C., RATKOVIC, M., HAZLETT, C. and IMAI, K. (2017). CBPS: R package for covariate balancing propensity score. Available at the Comprehensive R Archive Network (CRAN): https://CRAN.R-project.org/package=CBPS.

GRAHAM, B. S., PINTO, C. and EGEL, D. (2012). Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.* **79** 1053–1079.

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

HARDER, V. S., STUART, E. A. and ANTHONY, J. C. (2008). Adolescent cannabis problems and young adult depression: Male–female stratified propensity score analyses. *Am. J. Epidemiol.* **168** 592–601.

HAZLETT, C. (2016). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. Technical report, Depts. Statistics and Political Science, Univ. California Los Angeles.

HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*: *An Essential Journey with Donald Rubin's Statistical Family* 73–84. Wiley, New York.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15** 199–236.

IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. MR3153941

IMAI, K. and RATKOVIC, M. (2015). Robust estimation of inverse probability weights for marginal structural models. *J. Amer. Statist. Assoc.* **110** 1013–1023.

IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866.

IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710.

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms

McCaffrey, D. F., Ridgeway, G. and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics, Vol. IV. Handbooks in Econom.* **2** 2111–2245. North-Holland, Amsterdam. MR1315971

Nielsen, R. A., Findley, M. G., Davis, Z. S., Candland, T. and Nielson, D. L. (2011). Foreign aid shocks as a cause of violent armed conflict? *Amer. J. Polit. Sci.* **55** 219–232.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton, FL.

Robins, J. M., Hernán, M. Á. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

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

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

Rubin, D. B. (1990). Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statist. Sci.* **5** 472–480.

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

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97** 661–682. MR2672490

Urban, C. and Niebler, S. (2014). Dollars on the sidewalk: Should US presidential candidates advertise in uncontested states? *Amer. J. Polit. Sci.* **58** 322–336.

Zhao, Q. (2016). Covariate balancing propensity score by tailored loss functions. Technical report. Dept. Statistics, Stanford Univ.

Zhao, Q. and Percival, D. (2017). Entropy balancing is doubly robust. *J. Causal Inference* **5** 20160010.

Zhu, Y., Coffman, D. L. and Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *J. Causal Inference* **3** 25–40.

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

C. Fong
Graduate School of Business
Stanford University
Stanford, California 94305
USA
E-mail: christianfong@stanford.edu

C. Hazlett
Departments of Statistics and
Political Science
University of California, Los Angeles
Los Angeles, California 90095
USA
E-mail: chazlett@ucla.edu

K. Imai
Department of Politics
and
Center for Statistics and
Machine Learning
Princeton University
Princeton, New Jersey 08544
USA
E-mail: kimai@princeton.edu

This content downloaded from
169.226.227.41 on Sun, 27 Feb 2022 22:18:02 UTC
All use subject to https://about.jstor.org/terms