Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2014)
76, *Part* 1, *pp.* 243–263

Covariate balancing propensity score

Kosuke Imai and Marc Ratkovic

Princeton University, USA

[Received April 2012. Final revision March 2013]

Summary. The propensity score plays a central role in a variety of causal inference settings. In particular, matching and weighting methods based on the estimated propensity score have become increasingly common in the analysis of observational data. Despite their popularity and theoretical appeal, the main practical difficulty of these methods is that the propensity score must be estimated. Researchers have found that slight misspecification of the propensity score model can result in substantial bias of estimated treatment effects. We introduce covariate balancing propensity score (CBPS) methodology, which models treatment assignment while optimizing the covariate balance. The CBPS exploits the dual characteristics of the propensity score as a covariate balancing score and the conditional probability of treatment assignment. The estimation of the CBPS is done within the generalized method-of-moments or empirical likelihood framework. We find that the CBPS dramatically improves the poor empirical performance of propensity score matching and weighting methods reported in the literature. We also show that the CBPS can be extended to other important settings, including the estimation of the generalized propensity score for non-binary treatments and the generalization of experimental estimates to a target population. Open source software is available for implementing the methods proposed.

Keywords: Causal inference; Instrumental variables; Inverse propensity score weighting; Marginal structural models; Observational studies; Propensity score matching; Randomized experiments

1. Introduction

The propensity score, which is defined as the conditional probability of receiving treatment given covariates, plays a central role in a variety of settings for causal inference. In their seminal article, Rosenbaum and Rubin (1983) showed that, if the treatment assignment is strongly ignorable given observed covariates, an unbiased estimate of the average treatment effect can be obtained by adjusting for the propensity score alone rather than a vector of confounders, which is often of high dimension. Over the next three decades, several new methods based on the propensity score have been developed and they have become an essential part of applied researchers' toolkits across disciplines. In particular, the propensity score is used to adjust for observed confounding through matching (e.g. Rosenbaum and Rubin (1985), Rosenbaum (1989) and Abadie and Imbens (2006)), subclassification (e.g. Rosenbaum and Rubin (1984), Rosenbaum (1991) and Hansen (2004)), weighting (e.g. Rosenbaum (1987), Robins *et al.* (2000) and Hirano *et al.* (2003)), regression (e.g. Heckman *et al.* (1998)) or their combinations (e.g. Robins *et al.* (1995), Ho *et al.* (2007) and Abadie and Imbens (2011)). Imbens (2004), Lunceford and Davidian (2004) and Stuart (2010) have provided comprehensive reviews of these and other methods.

Address for correspondence: Kosuke Imai, Department of Politics, Princeton University, Princeton, NJ 08544, USA.
E-mail: kimai@princeton.edu

 1369–7412/14/76243

Despite their popularity and theoretical appeal, a main practical difficulty of these methods is that the propensity score must be estimated. In fact, researchers have found that slight misspecification of the propensity score model can result in substantial bias of estimated treatment effects (e.g. Kang and Schafer (2007) and Smith and Todd (2005)). This challenge highlights the paradoxical nature of the propensity score—the propensity score is designed to reduce the dimension of covariates and yet its estimation requires modelling of high dimensional covariates. In practice, applied researchers search for an appropriate propensity score model specification by repeating the process of changing their model and checking the resulting covariate balance. Imai *et al.* (2008) called this the 'propensity score tautology'—the estimated propensity score is appropriate if it balances covariates. Although various methods have been proposed to refine and improve propensity score matching and weighting techniques (e.g. Robins *et al.* (1994) and Abadie and Imbens (2011)), we believe that it is also essential to develop a robust method for estimating the propensity score.

In this paper, we introduce the *covariate balancing propensity score* (CBPS) and show how to estimate the propensity score such that the resulting covariate balance is optimized. We exploit the dual characteristics of the propensity score as a covariate balancing score and the conditional probability of treatment assignment. Specifically, we use a set of moment conditions that are implied by the covariate balancing property (i.e. mean independence between the treatment and covariates after inverse propensity score weighting) to estimate the propensity score while also incorporating the standard estimation procedure (i.e. the score condition for the maximum likelihood) whenever appropriate. As will be shown, satisfying the score condition can be seen as a particular covariate balancing condition. The proposed CBPS estimation is then carried out within the familiar generalized method of moments (GMM) or empirical likelihood (EL) framework (Hansen, 1982; Owen, 2001).

The CBPS has several attractive characteristics. First, the CBPS estimation mitigates the effect of the potential misspecification of a parametric propensity score model by selecting parameter values that maximize the resulting covariate balance. In Section 3, we find that the CBPS dramatically improves the poor empirical performance of propensity score matching and weighting methods that were reported by Kang and Schafer (2007) and Smith and Todd (2005) respectively. Second, the CBPS can be extended to various other important settings in causal inference. In Section 4, we briefly describe the extension of the CBPS to the generalized propensity score for non-binary treatment (Imbens, 2000; Imai and van Dyk, 2004) and the generalization of experimental estimates to a target population (Cole and Stuart, 2010). Third, the CBPS inherits all the theoretical properties and methodologies that already exist in the GMM and EL literature. For example, GMM specification tests and moment selection procedures are directly applicable. Finally, because the methodology proposed simply improves the estimation of the propensity score, various propensity score methods such as matching and weighting can be implemented without modification.

The key idea behind the CBPS is that a single model determines the treatment assignment mechanism and the covariate balancing weights. This differs from several existing methods for automated covariate balancing (e.g. Diamond and Sekhon (2012), Iacus *et al.* (2011), Hainmueller (2012) and Ratkovic (2012)). In particular, Hainmueller (2012) proposed the entropy balancing method to construct a weight for each control observation such that the sample moments of observed covariates are identical between the treatment and weighted control groups. Unlike the entropy balancing method, however, the CBPS constructs balancing weights directly from the propensity score. In addition, Graham *et al.* (2012) proposed a similar covariate balancing method under the empirical likelihood framework, but the treatment assignment mechanism was not explicitly modelled. These methods resemble the weighting methods for

analysing sample surveys when some auxiliary information about population distribution is available (e.g. Deming and Stephan (1940), Little and Wu (1991), Hellerstein and Imbens (1999), Nevo (2003) and Chaudhuri *et al.* (2008)).

Furthermore, the method that was proposed by Tan (2010) uses the maximum likelihood estimate of the propensity score and achieves many desirable properties such as double robustness and sample boundedness by incorporating the outcome model. But, this method does not explicitly link the propensity score and covariate balancing weights. In contrast with the methods that were proposed by Tan (2010) and Graham *et al.* (2012), the CBPS focuses on the estimation of the propensity score without consulting the outcome data, which aligns with the original spirit of the propensity score methodology (Rubin, 2007). This separation from the outcome model enables the CBPS to be applicable to a variety of causal inference settings and to be used with the existing propensity score methods such as matching and weighting. Section 2.4 gives a more detailed discussion of the connections between the CBPS and these existing methods.

Finally, the method proposed can be implemented through the open source R package `CBPS` (Ratkovic *et al.*, 2012), which is available from the Comprehensive R Archive Network (`http://cran.r-project.org/package=CBPS`).

2. Methodology proposed

2.1. The set-up

Consider a simple random sample of N observations from a population $\mathcal{P}$. For each unit i, we observe a binary treatment variable T_i and a K-dimensional column vector of observed pretreatment covariates X_i whose support is denoted by $\mathcal{X}$. The propensity score is defined as the conditional probability of receiving the treatment given the covariates X_i. Following Rosenbaum and Rubin (1983), we assume that the true propensity score is bounded away from 0 and 1:

$$0 < \Pr(T_i = 1 | X_i = x) < 1 \qquad \text{for any } x \in \mathcal{X}. \tag{1}$$

We emphasize that in practice this assumption should be made with care. For example, when conducting a programme evaluation, researchers may exclude from their data set those individuals who are not eligible for the programme. Estimating the support of the propensity score is a difficult problem and is beyond the scope of this paper.

Rosenbaum and Rubin (1983) showed that if we further assume the ignorability of treatment assignment, i.e.

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i | X_i \tag{2}$$

where $Y_i(t)$ represents the potential outcome under the treatment status $t \in \{0, 1\}$, then the treatment assignment is ignorable given the (true) propensity score $\pi(X_i)$:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i | \pi(X_i). \tag{3}$$

This implies that the unbiased estimation of treatment effects is possible by conditioning on the propensity score alone rather than the entire covariate vector X_i, which is often of high dimension. This dimension reduction property led to the subsequent development of various propensity score methods, including matching and weighting. Although matching exactly on the propensity score is typically impossible, methods have been developed to reduce the bias due to imperfect matching (Abadie and Imbens, 2011) or to obtain a consistent estimate via weighting (Robins *et al.*, 1994).

In observational studies, however, the propensity score is unknown and must be estimated from the data. Typically, researchers assume a parametric propensity score model $\pi_\beta(X_i)$,

$$\Pr(T_i = 1|X_i) = \pi_\beta(X_i) \tag{4}$$

where $\beta \in \Theta$ is an L-dimensional column vector of unknown parameters. For example, a popular choice is the logistic model

$$\pi_\beta(X_i) = \frac{\exp(X_i^{\mathrm{T}}\beta)}{1+\exp(X_i^{\mathrm{T}}\beta)} \tag{5}$$

in which case we have $L = K$. Researchers then maximize the empirical fit of the model so that the estimated propensity score predicts the observed treatment assignment well. This is often done by maximizing the log-likelihood function:

$$\hat{\beta}_{\mathrm{MLE}} = \arg\max_{\beta\in\Theta} \sum_{i=1}^{N} T_i \log\{\pi_\beta(X_i)\} + (1-T_i)\log\{1-\pi_\beta(X_i)\}. \tag{6}$$

Assuming that $\pi_\beta(\cdot)$ is twice continuously differentiable with respect to β, this implies the first-order condition

$$\frac{1}{N}\sum_{i=1}^{N} s_\beta(T_i, X_i) = 0, \qquad s_\beta(T_i, X_i) = \frac{T_i\,\pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\,\pi'_\beta(X_i)}{1-\pi_\beta(X_i)}, \tag{7}$$

and $\pi'_\beta(X_i) = \partial\pi(X_i)/\partial\beta^{\mathrm{T}}$. We emphasize that equation (7) can also be interpreted as the condition that balances a particular function of covariates, i.e. the first derivative of $\pi_\beta(X_i)$.

The major difficulty of this standard approach is that the propensity score model may be misspecified, yielding biased estimates of treatment effects. Although in theory a more complex non-parametric model can be used (e.g. McCaffrey *et al.* (2004)), the high dimensionality of covariates X_i may pose a challenge. To address this issue, we develop CBPS estimation as a method that is robust to mild misspecification of the parametric propensity score model. We achieve this robustness by exploiting the dual characteristics of the propensity score as a covariate balancing score and the conditional probability of treatment assignment.

We operationalize the covariate balancing property by using inverse propensity score weighting:

$$\mathbb{E}\left\{\frac{T_i\tilde{X}_i}{\pi_\beta(X_i)} - \frac{(1-T_i)\tilde{X}_i}{1-\pi_\beta(X_i)}\right\} = 0 \tag{8}$$

where $\tilde{X}_i = f(X_i)$ is an M-dimensional vector-valued measurable function of X_i specified by the researcher. The score condition given in equation (7) represents a particular choice of $f(\cdot)$, i.e. $\tilde{X}_i = \pi'_\beta(X_i)$, thereby giving more weights to covariates that are predictive of the treatment assignment according to the propensity score model. However, equation (8) must hold for *any* function of covariates so long as the expectation exists. Indeed, if the propensity score model is misspecified, maximizing the likelihood might not balance covariates. Thus, by setting $\tilde{X}_i = X_i$, we can ensure that the first moment of each covariate is balanced even when the model is misspecified. Similarly, $\tilde{X}_i = (X_i^{\mathrm{T}} X_i^{2\mathrm{T}})^{\mathrm{T}}$ will balance both the first and the second moments. Currently, the literature offers little theoretical guidance on what covariates should be included and we do not resolve this problem in this paper.

If the average treatment effect for the treated is of interest, we may wish to weight the control group observations such that their (weighted) covariate distribution matches with that of the treatment group. In this case, the moment condition becomes

$$\mathbb{E}\left\{ T_i \tilde{X}_i - \frac{\pi_\beta(X_i)(1-T_i)\tilde{X}_i}{1-\pi_\beta(X_i)} \right\} = 0. \tag{9}$$

Several remarks are in order. First, the score condition in equation (7) represents another covariate balancing condition where $\tilde{X}_i = \pi'_\beta(X_i)$, thereby placing a greater emphasis on covariates with strong predictive power. Second, as discussed in Section 2.2, it is possible to overidentify the propensity score model by incorporating both the score and the covariate balancing conditions. Third, we note that the covariate balancing property follows directly from the definition of the propensity score and does not require the ignorability assumption that is given in equation (2). Thus, the application of the CBPS will improve the balance of observed covariates regardless of whether there are unmeasured confounders (though the resulting treatment effect estimates may be biased unless the ignorability assumption holds). Finally, following Rubin (2007), we separate the estimation of the propensity score from the analysis of outcome data. Indeed, the goal of the CBPS is to improve the commonly used parametric estimation of the propensity score regardless of statistical methods that are used to estimate treatment effects.

2.2. Estimation and inference

We estimate the CBPS by using the moment conditions based on the covariate balancing property under the GMM or EL framework (see Hayashi (2000) and Owen (2001) for an accessible introduction of each method). The CBPS is said to be *just identified* when the number of parameters equals the number of moment conditions. For example, consider the logistic regression given in equation (5). The just-identified CBPS results if we use the covariate balancing conditions given in equation (8) or (9) alone by setting $X_i = f(X_i)$. In contrast, if we combine this with the score condition given in equation (7), then the CBPS is *overidentified* because the number of moment conditions $L + M$ exceeds that of model parameters L. In the literature, it is suggested that the overidentifying restrictions generally improve asymptotic efficiency but may result in a poor finite sample performance. In our simulation and empirical studies (Section 3), we shall examine the performance of both a just-identified and an overidentified CBPS.

Formally, define the sample analogue of the covariate balancing moment condition given in equation (8) as

$$\frac{1}{N}\sum_{i=1}^{N} w_\beta(T_i, X_i)\tilde{X}_i, \qquad w_\beta(T_i, X_i) = \frac{T_i - \pi_\beta(X_i)}{\pi_\beta(X_i)\{1-\pi_\beta(X_i)\}}. \tag{10}$$

If the average treatment effect for the treated rather than the average treatment effect is the quantity of interest, we use the sample analogue of equation (9). In this case, the weight becomes

$$w_\beta(T_i, X_i) = \frac{N}{N_1}\frac{T_i - \pi_\beta(X_i)}{1-\pi_\beta(X_i)}. \tag{11}$$

For the overidentified CBPS, we follow Hansen (1982) and use the following efficient GMM estimator,

$$\hat{\beta}_{\mathrm{GMM}} = \arg\min_{\beta\in\Theta} \bar{g}_\beta(T, X)^{\mathrm{T}} \Sigma_\beta(T, X)^{-1} \bar{g}_\beta(T, X) \tag{12}$$

where $\bar{g}_\beta(T, X)$ is the sample mean of the moment conditions,

$$\bar{g}_\beta(T, X) = \frac{1}{N}\sum_{i=1}^{N} g_\beta(T_i, X_i), \tag{13}$$

and $g_\beta(T_i, X)$ combining all moment conditions,

$$g_\beta(T_i, X_i) = \begin{pmatrix} s_\beta(T_i, X_i) \\ w_\beta(T_i, X_i)\tilde{X}_i \end{pmatrix}. \tag{14}$$

We assume that $\pi_\beta(\cdot)$ and $f(\cdot)$ satisfy the standard regularity conditions of the GMM estimator (Newey and McFadden, 1994). For example, if the conditional distribution of T_i given X_i belongs to the exponential family, then it is sufficient to assume that the expectations of $\tilde{X}_i$ and $w_\beta(T_i, X_i)$ exist and that all moment conditions are satisfied at a unique value of β.

We use the 'continuous updating' GMM estimator (Hansen *et al.*, 1996), which, unlike the two-step optimal GMM estimator, is invariant and has better finite sample properties. Our choice of a consistent covariance estimator for $g_\beta(T_i, X_i)$ is given by

$$\Sigma_\beta(T, X) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\{g_\beta(T_i, X_i)\, g_\beta(T_i, X_i)^{\mathrm{T}} | X_i\} \tag{15}$$

where we integrate out the treatment variable T_i conditional on the pretreatment covariates X_i. We find that this covariance estimator outperforms the sample covariance of moment conditions because the latter does not penalize large weights. In particular, in the case of the logistic regression propensity score model, i.e. $\pi_\beta(X_i) = \mathrm{logit}^{-1}(X_i^{\mathrm{T}}\beta)$, we have the expression for $\Sigma_\beta(T, X)$

$$\Sigma_\beta(T, X) = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \pi_\beta(X_i)\{1-\pi_\beta(X_i)\}X_i X_i^{\mathrm{T}} & X_i\tilde{X}_i^{\mathrm{T}} \\ \tilde{X}_i X_i^{\mathrm{T}} & [\pi_\beta(X_i)\{1-\pi_\beta(X_i)\}]^{-1}\tilde{X}_i\tilde{X}_i^{\mathrm{T}} \end{pmatrix} \tag{16}$$

when we use the average treatment effect weight given in equation (10), or

$$\Sigma_\beta(T, X) = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \pi_\beta(X_i)\{1-\pi_\beta(X_i)\}X_i X_i^{\mathrm{T}} & N\,\pi_\beta(X_i)X_i\tilde{X}_i^{\mathrm{T}}/N_1 \\ N\,\pi_\beta(X_i)\tilde{X}_i X_i^{\mathrm{T}}/N_1 & N^2\,\pi_\beta(X_i)/[N_1^2\{1-\pi_\beta(X_i)\}]\tilde{X}_i\tilde{X}_i^{\mathrm{T}} \end{pmatrix} \tag{17}$$

if the average treatment effect for the treated weight given in equation (11) is used. With a set of reasonable starting values (e.g. $\hat{\beta}_{\mathrm{MLE}}$), we find that the gradient-based optimization method works well in terms of speed and reliability.

For estimating the just-identified CBPS, we still use equation (12) without the score condition and find the optimal value of parameter β such that this objective function equals 0.

Alternatively, we can apply the EL framework to the above moment conditions (Qin and Lawless, 1994) where the profile empirical likelihood ratio function is given by

$$R(\beta) = \sup\left\{ \prod_{i=1}^{n} n\,p_i \,|\, p_i \geqslant 0,\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i\, g_\beta(T_i, X_i) = 0 \right\}. \tag{18}$$

This EL estimator shares many of the attractive properties of the above continuous updating GMM estimator, notably invariance and higher order bias properties.

2.3. Specification test

One advantage of the overidentified CBPS is that we can use the test of overidentifying restrictions as a specification test for the propensity score model. This is not possible if the propensity score is estimated on the basis of either the score condition or the covariate balancing condition alone. Under the GMM framework, we can use Hansen's J-statistic

$$J = N\{\bar{g}_{\hat{\beta}_{\mathrm{GMM}}}(T, X)^{\mathrm{T}}\Sigma_{\hat{\beta}_{\mathrm{GMM}}}(T, X)^{-1}\,\bar{g}_{\hat{\beta}_{\mathrm{GMM}}}(T, X)\} \xrightarrow{\mathrm{d}} \chi^2_{L+M} \tag{19}$$

where the null hypothesis is that the propensity score model is correctly specified. If the propensity score model is correct, any deviation of this statistic from 0 should be within the range of sampling error. We emphasize that the failure to reject this null hypothesis does not necessarily imply the correct specification of the propensity score model because it may simply mean that the test lacks the power (Imai *et al.*, 2008). Nevertheless, the test could be useful for detecting the model misspecification. Within the EL framework, a similar specification test can be conducted on the basis of the likelihood ratio.

2.4. Related methods

The idea to weight observations and to balance covariates can be found in both the causal inference and the survey sampling literatures. Although there is no single best measure of balance, in our context, weighting is more attractive for balancing covariates than matching because the latter requires exact matching on the estimated propensity score, which is typically impossible (Abadie and Imbens, 2011). We do not use balance measures based on matching because matching is a non-smooth function of data and so the standard GMM or EL theory may not be applicable (Abadie and Imbens, 2006, 2008).

An important difference between the CBPS and the existing methods, however, is that the former exploits the fact that a single model determines the treatment assignment mechanism and the covariate balancing weights. Here, we briefly review some closely related existing methods. First, Hainmueller (2012) proposed the entropy balancing method, which weights each observation to achieve optimal balance. Since identifying N weights by balancing the sample mean of K covariates between the treatment and control groups is an underdetermined problem when $K < N$, the entropy balancing method minimizes the Kullback–Leibler divergence from a set of baseline weights chosen by researchers. In contrast, the CBPS directly models the propensity score, which is simultaneously used to construct weights for observations. Although the weights that result from the entropy balancing method may imply the propensity score, the CBPS makes it easier for applied researchers to model the propensity score directly by using a familiar parametric model.

Second, Hainmueller (2008) briefly discussed the possibility of incorporating covariate balancing conditions into the estimation of the propensity score under the EL framework as a potential extension of his method (section IV C). We note that his set-up differs from the CBPS in that the weights are constructed separately from the propensity score. Similarly, Graham *et al.* (2012) proposed the EL method, which uses covariate balancing moment conditions to estimate the propensity score. Unlike these methods, however, the CBPS optimizes the covariate balance while modelling the treatment assignment. Moreover, the overidentified CBPS enables the specification test for the propensity score model.

Third, under the likelihood framework, Tan (2010) identified a set of constraints to generate observation-specific weights that enable desirable properties such as double robustness, local efficiency and sample boundedness. These weights may not fall between 0 and 1 and hence cannot be interpreted as the propensity score. In addition, like the method of Graham *et al.* (2012), Tan's method incorporates the outcome model in a creative manner when constructing weights. In contrast, the CBPS focuses on estimation of the propensity score without consulting the outcome data, which aligns with the original spirit of the propensity score methods (Rubin, 2007). As illustrated in Section 4, the direct connection between the CBPS and the propensity score widens the applicability of the CBPS. For the same reason, the CBPS can also be easily used in conjunction with existing propensity score methods such as matching and weighting.

All these covariate balancing methods resemble the weighting methods for analysing sample surveys when some auxiliary information about the population distribution is available (see Deming and Stephan (1940), Oh and Scheuren (1983), Little and Wu (1991), Hellerstein and Imbens (1999), Nevo (2003) and many others). For example, Nevo (2003) proposed a method where the availability of auxiliary data is assumed and the set of covariate balancing moment conditions is used to estimate a parametric sample selection model.

In addition, several methods have been proposed to improve the estimation of the propensity score. In particular, McCaffrey *et al.* (2004) proposed to use the generalized boosting model (GBM) for the estimation of the propensity score, which they reported works well in practice. The advantage of this method is that it is non-parametric, allowing the complex relationship between the treatment variable and a large number of covariates. In contrast, the CBPS represents a simple and yet powerful method to improve the commonly used parametric estimation of the propensity score. Although the parametric method has its own limitations, it is easier for applied researchers to use and interpret. Unlike non-parametric methods such as the GBM, the CBPS does not require tuning parameters.

Finally, some recent methods aim to balance the joint distribution of all covariates directly without estimating the propensity score. Iacus *et al.* (2011) proposed to conduct exact matching after making continuous covariates discrete and coarsening discrete covariates. Camillo and D'Attoma (2011) followed the same strategy of discretizing continuous covariates and constructed a global measure of imbalance based on the conditional multiple-correspondence analysis, which is a projection method that is related to factor analysis. Ratkovic (2012) developed a non-parametric subsetting method which asymptotically achieves joint independence between a general treatment regime and a vector of covariates.

3. Simulation and empirical studies

In this section, we apply the proposed methodology to prominent simulation and empirical studies where propensity score methods have been shown to fail. We show that the CBPS can dramatically improve the poor performance of the propensity score weighting and matching methods that have been reported in the previous studies. Before we describe the results, we note that the CBPS optimizes the balance measure that is directly relevant for weighting methods. For matching, the connection may not be as clear since matching can be done by using different balance measures. Here, we limit our investigation to the question of how the CBPS can improve the performance of a certain propensity score matching procedure by estimating the propensity score differently.

3.1. Improved performance of propensity score weighting methods

In a controversial paper, Kang and Schafer (2007) conducted a set of simulation studies to study the performance of propensity score weighting methods. They found that the misspecification of a propensity score model can negatively affect the performance of various weighting methods. In particular, they showed that, although the doubly robust estimator of Robins *et al.* (1994) provides a consistent estimate of the treatment effect if either the outcome model or the propensity score model is correct, the performance of the doubly robust estimator can deteriorate when both models are slightly misspecified. This finding led to a rebuttal by Robins *et al.* (2007), who criticized the simulation set-up and introduced alternative doubly robust estimators.

In this section, we replicate the simulation study of Kang and Schafer (2007) except that we

estimate the propensity score by using our proposed methodology. We then examine whether or not the CBPS can improve the empirical performance of propensity score weighting estimators. In particular, Kang and Schafer used the following data-generating process. There are four pretreatment covariates $X_i^* = (X_{i1}^*, X_{i2}^*, X_{i3}^*, X_{i4}^*)$, each of which is independently, identically distributed according to the standard normal distribution. The true outcome model is a linear regression with these covariates and the error term is an independently, identically distributed standard normal random variate such that the mean outcome of the treated observations equals 210, which is the quantity of interest to estimate. The true propensity score model is a logistic regression with X_i^* being the linear predictor such that the mean probability of receiving the treatment equals 0.5. Finally, only the non-linear transforms of covariates are observed and they are given by $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}) = \{\exp(X_{i1}^*/2), X_{i2}^*/\{1+\exp(X_{1i}^*)\}+10, (X_{i1}^* X_{i3}^*/25+0.6)^3, (X_{i1}^* + X_{i4}^* + 20)^2\}$.

Kang and Schafer (2007) studied four propensity score weighting estimators. The propensity score model that they used is a logistic regression with X_i as the linear predictor. This is a misspecified model because the true propensity score is a logistic regression with X_i^* as the linear predictor (but non-linear in the observed covariates X_i). The weighting estimators that they examined are the Horvitz–Thompson estimator HT (Horvitz and Thompson, 1952), the inverse propensity score weighting estimator IPW (Hirano *et al.*, 2003), the weighted least squares regression estimator WLS (Robins *et al.*, 2000; Freedman and Berk, 2008) and the doubly robust estimator DR (Robins *et al.*, 1994):

$$\hat{\mu}_{\mathrm{HT}} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\pi_{\hat{\beta}}(X_i)},$$

$$\hat{\mu}_{\mathrm{IPW}} = \sum_{i=1}^{n}\frac{T_i Y_i}{\pi_{\hat{\beta}}(X_i)} \Big/ \sum_{i=1}^{n}\frac{T_i}{\pi_{\hat{\beta}}(X_i)},$$

$$\hat{\mu}_{\mathrm{WLS}} = \frac{1}{n}\sum_{i=1}^{n} X_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{WLS}}, \qquad \hat{\gamma}_{\mathrm{WLS}} = \left\{\sum_{i=1}^{n}\frac{T_i X_i X_i^{\mathrm{T}}}{\pi_{\hat{\beta}}(X_i)}\right\}^{-1}\sum_{i=1}^{n}\frac{T_i X_i Y_i}{\pi_{\hat{\beta}}(X_i)},$$

$$\hat{\mu}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left\{X_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{OLS}} + \frac{T_i(Y_i - X_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{OLS}})}{\pi_{\hat{\beta}}(X_i)}\right\}, \qquad \hat{\gamma}_{\mathrm{OLS}} = \left(\sum_{i=1}^{n} T_i X_i X_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n} T_i X_i Y_i.$$

Both the weighted least squares and the ordinary least squares regressions are misspecified because the true model is linear in X_i^* rather than X_i.

To estimate the propensity score, Kang and Schafer (2007) used the logistic regression with X_i being the linear predictor, i.e. $\pi_\beta(X_i) = \mathrm{logit}^{-1}(X_i^{\mathrm{T}}\beta)$, which is misspecified because the true propensity score model is the logistic regression with X_i^* being the linear predictor. Our simulation study uses the same propensity and outcome model specifications but we investigate whether the CBPS improves the empirical performance of the weighting estimators. To estimate the CBPS, we use the same logistic regression specification but use the covariate balancing moment conditions by setting $\tilde{X}_i = X_i$ under the GMM framework of the methodology proposed outlined in Section 2. Thus, our simulation study examines how replacing the standard logistic regression propensity score with the CBPS will improve the empirical performance of the four commonly used weighting estimators.

As in the original study, we conduct simulations under four scenarios:

(a) both propensity score and outcome models are correctly specified,
(b) only the propensity score model is correct,

(c) only the outcome model is correct and
(d) both the propensity score and the outcome models are misspecified.

For each scenario, two sample sizes, 200 and 1000, are used and we conduct 10000 Monte Carlo simulations and calculate the bias and root-mean-squared error (RMSE) for each estimator.

The results of our simulation study are presented in Table 1. For a given scenario, we examine the bias and RMSE of each weighting estimator on the basis of four different propensity score methods:

(a) the standard logistic regression with X_i being the linear predictor as in the original simulation study ('GLM'),
(b) the just-identified CBPS estimation with the covariate balancing moment conditions with respect to X_i and without the score condition ('CBPS1'),
(c) the overidentified CBPS estimation with both covariate balancing and score conditions ('CBPS2') and
(d) the true propensity score ('True'), i.e. $\pi_\beta(X_i) = \text{logit}^{-1}(X_i^{*\mathrm{T}}\beta)$.

In the first scenario where both models are correct, all four weighting methods have relatively low bias regardless of what propensity score estimation method we use (though the overidentified CBPS introduces some bias for estimators HT and IPW). However, the estimator HT has a large variance and hence a large RMSE when either the standard logistic regression or the true propensity score is used. In contrast, when used with the CBPS, the same estimator has a much lower RMSE. There is a bias–variance trade-off where, relative to method GLM, the CBPS dramatically reduces variance at the expense of some increase in bias. A similar observation can be made for estimator IPW whereas estimators WLS and DR are not sensitive to the choice of propensity score estimation methods.

The second simulation scenario shows the performance of various estimators when the propensity score model is correct but the outcome model is misspecified. As expected, the results are quite similar to those of the first scenario. When the propensity score model is correctly specified, the four weighting estimators have low bias. However, the CBPS significantly reduces the variance of estimator HT even when compared with the true propensity score. This confirms the theoretical result in the literature that the estimated propensity score leads to a more efficient estimator of the average treatment effect (e.g. Hahn (1998) and Hirano *et al.* (2003)).

The third scenario examines an interesting situation where the propensity score model is misspecified whereas the outcome models for estimators WLS and DR are correct. As expected, we find that estimators HT and IPW, which solely rely on the propensity score, have large bias and RMSE when used with the standard logistic regression model. However, the CBPS significantly reduces their bias and RMSE regardless of whether it incorporates the score equation from the logistic likelihood. In contrast, the bias and RMSE of estimators WLS and DR remain low and essentially identical across the propensity score estimation methods. Together with the results under the second scenario, these results confirm the double-robustness property where estimator DR performs well so long as either the propensity score or outcome model is correctly specified.

The final simulation scenario illustrates the most important point made by Kang and Schafer (2007) that the performance of estimator DR can deteriorate when both the propensity score and the outcome models are misspecified. Under this scenario, all models suffer from some degree of bias when used with the standard logistic regression model. The bias and RMSE are

Table 1. Relative performance of the four different propensity score weighting estimators based on different propensity score estimation methods under the simulation setting of Kang and Schafer (2007)†

Sample size n	*Estimator*	*Bias for the following methods:*				*RMSE for the following methods:*			
		GLM	*CBPS1*	*CBPS2*	*True*	*GLM*	*CBPS1*	*CBPS2*	*True*
(1) Both models correct									
200	HT	0.33	2.06	−4.74	1.19	12.61	4.68	9.33	23.93
	IPW	−0.13	0.05	−1.12	−0.13	3.98	3.22	3.50	5.03
	WLS	−0.04	−0.04	−0.04	−0.04	2.58	2.58	2.58	2.58
	DR	−0.04	−0.04	−0.04	−0.04	2.58	2.58	2.58	2.58
1000	HT	0.01	0.44	−1.59	−0.18	4.92	1.76	4.18	10.47
	IPW	0.01	0.03	−0.32	−0.05	1.75	1.44	1.60	2.22
	WLS	0.01	0.01	0.01	0.01	1.14	1.14	1.14	1.14
	DR	0.01	0.01	0.01	0.01	1.14	1.14	1.14	1.14
(2) Propensity score model correct									
200	HT	−0.05	1.99	−4.94	−0.14	14.39	4.57	9.39	24.28
	IPW	−0.13	0.02	−1.13	−0.18	4.08	3.22	3.55	4.97
	WLS	0.04	0.04	0.04	0.04	2.51	2.51	2.51	2.51
	DR	0.04	0.04	0.04	0.04	2.51	2.51	2.51	2.51
1000	HT	−0.02	0.44	−1.67	0.29	4.85	1.77	4.22	10.62
	IPW	0.02	0.05	−0.31	−0.03	1.75	1.45	1.61	2.27
	WLS	0.04	0.04	0.04	0.04	1.14	1.14	1.14	1.14
	DR	0.04	0.04	0.04	0.04	1.14	1.14	1.14	1.14
(3) Outcome model correct									
200	HT	24.25	1.09	−5.42	−0.18	194.58	5.04	10.71	23.24
	IPW	1.70	−1.37	−2.84	−0.26	9.75	3.42	4.74	4.93
	WLS	−2.29	−2.37	−2.19	0.41	4.03	4.06	3.96	3.31
	DR	−0.08	−0.10	−0.10	−0.10	2.67	2.58	2.58	2.58
1000	HT	41.14	−2.02	2.08	−0.23	238.14	2.97	6.65	10.42
	IPW	4.93	−1.39	−0.82	−0.02	11.44	2.01	2.26	2.21
	WLS	−2.94	−2.99	−2.95	0.20	3.29	3.37	3.33	1.47
	DR	0.02	0.01	0.01	0.01	1.89	1.13	1.13	1.13
(4) Both models incorrect									
200	HT	30.32	1.27	−5.31	−0.38	266.30	5.20	10.62	23.86
	IPW	1.93	−1.26	−2.77	−0.09	10.50	3.37	4.67	5.08
	WLS	−2.13	−2.20	−2.04	0.55	3.87	3.91	3.81	3.29
	DR	−7.46	−2.59	−2.13	0.37	50.30	4.27	3.99	3.74
1000	HT	101.47	−2.05	1.90	0.01	2371.18	3.02	6.75	10.53
	IPW	5.16	−1.44	−0.92	0.02	12.71	2.06	2.39	2.25
	WLS	−2.95	−3.01	−2.98	0.19	3.30	3.40	3.36	1.47
	DR	−48.66	−3.59	−3.79	0.08	1370.91	4.02	4.25	1.81

†The bias and RMSE are computed for the Horvitz–Thompson HT, the inverse propensity score weighting IPW, the inverse propensity score weighted least squares WLS and the doubly robust least squares DR estimators. The performance of the just-identified CBPS (CBPS1) and the overidentified CBPS (CBPS2) is compared with that of the standard logistic regression (GLM) and the true propensity score (True). We consider four scenarios where the outcome and/or propensity score models are misspecified. The sample sizes are $n = 200$ and $n = 1000$. The number of simulations is 10000. Across the four weighting estimators, the CBPS dramatically improves the performance of GLM when the model is misspecified.

the largest for estimator HT but estimator DR also exhibits a significant amount of bias and variance. However, the CBPS with or without the score equation can substantially improve the performance of estimator DR. Specifically, when the sample size is 1000, the bias and RMSE are dramatically reduced. The CBPS also significantly improves the performance of estimators HT and IPW even when compared with the true propensity score. In sum, even when both the

outcome and the propensity score models are misspecified, the CBPS can yield robust estimates of treatment effects.

Throughout this simulation study, the just-identified CBPS without the score equation significantly outperforms the overidentified CBPS for estimator HT. For other estimators, the results are similar. The overidentifying restriction test that was described in Section 2.3 mostly fails to detect model misspecification, indicating that this simulation study is well designed.

How can the CBPS dramatically improve the performance of the GLM? Fig. 1 shows that the overidentified CBPS sacrifices likelihood to improve balance under two of the four simulation scenarios. We use the following multivariate version of the 'standardized bias' (Rosenbaum and Rubin, 1985) to measure the overall covariate imbalance:

$$\text{Imbalance} = \left\{ \left(\frac{1}{N} \sum_{i=1}^{N} w_{\hat{\beta}}(T_i, X_i) X_i \right)^{\mathrm{T}} \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i^{\mathrm{T}} \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} w_{\hat{\beta}}(T_i, X_i) X_i \right) \right\}^{1/2}. \tag{20}$$

Fig. 1 shows that, when both the outcome and propensity score models are correctly specified (Figs 1(a)–1(c)), the CBPS achieves better covariate balance (Fig. 1(b)) without sacrificing much likelihood (Fig. 1(a)). In contrast, when both models are misspecified (Figs 1(d)–1(f)), the CBPS significantly improves covariate balance (Fig. 1(e)) at the cost of some likelihood (Fig. 1(d)). Figs 1(c) and 1(f) show the CBPS's trade-off between likelihood and covariate balance where a larger improvement in covariate balance is associated with a larger loss of likelihood.

Finally, as an additional comparison, we have applied the GBM of propensity score that was proposed by McCaffrey *et al.* (2004) (we use their `twang` R package with its default options). This non-parametric method has been reported to work well in the Kang and Schafer (2007) simulation setting (Ridgeway and McCaffrey, 2007). We briefly summarize the results of this comparison. We find that the performance of the GBM is significantly worse than that of the CBPS for estimators HT and IPW. In particular, for estimator HT, the GBM often exhibits greater bias and RMSE than does GLM. In contrast, the GBM performs best for estimators WLS and DR, outperforming the CBPS.

3.2. Improved performance of propensity score matching methods

In an influential paper, LaLonde (1986) empirically evaluated the ability of various estimators to obtain an unbiased estimate of the average treatment effect in the absence of randomized treatment assignment. From a randomized study of a job training programme (the 'National supported work demonstration') where an unbiased estimate of the average treatment effect is available, LaLonde constructed an 'observational study' by replacing the control group of these experimental data with untreated observations from non-experimental data sets such as the Current Population Survey and the Panel Study of Income Dynamics. LaLonde showed that the estimators he evaluated failed to replicate the experimental benchmark, and this finding led to increasing interest in experimental evaluation among social scientists.

More than a decade later, Dehejia and Wahba (1999) revisited LaLonde's study and showed that the propensity score matching estimator could closely replicate the experimental benchmark. Dehejia and Wahba estimated the propensity score by using the logistic regression and matched a treated observation with a control observation on the basis of the estimated propensity score. This finding, however, came under intense criticism by Smith and Todd (2005). They argued that the impressive performance of the propensity score matching estimator that had been reported by Dehejia and Wahba critically hinges on a particular subsample of the original LaLonde data that they analysed. This subsample excludes most of the high income workers

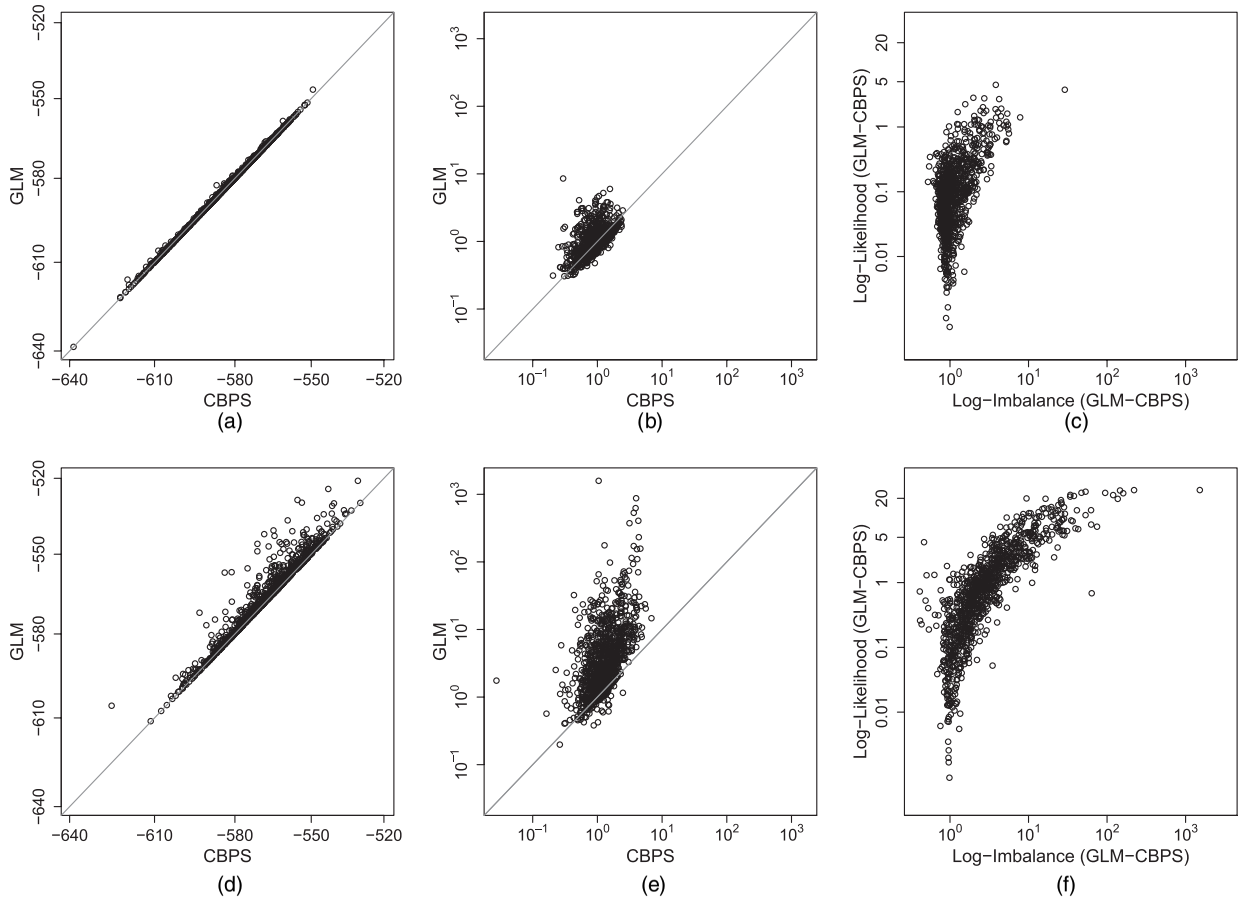

Fig. 1. Comparison of likelihood and imbalance between the standard logistic regression (GLM) and overidentified CBPS: the two scenarios from the simulation results reported in Table 1 are shown for the sample size $n = 1000$; the propensity score model is correctly specified ((a)–(c)), and both models are misspecified ((d)–(f)); each dot in a plot represents one Monte Carlo simulation draw; when the model is correct, the CBPS reduces the multivariate standardized bias, which is a summary measure of imbalance across all covariates ((b)) without sacrificing much likelihood, which is a measure of model fit ((a)); in contrast, when the model is misspecified, the CBPS significantly improves balance (((e)) while sacrificing some degree of model fit) ((d)); (c) and (f) demonstrate this likelihood–balance trade-off

from the non-experimental comparison sets, thereby making the selection bias much easier to overcome. Indeed, Smith and Todd showed that, if one focuses on the Dehejia and Wahba sample, other conventional estimators do as well as the propensity score matching estimator. Moreover, the propensity score matching estimator cannot replicate the experimental benchmark when applied to the original LaLonde data and is quite sensitive to the propensity score model specification.

In what follows, we investigate whether the CBPS can improve the poor performance of the propensity score matching estimator that was reported by Smith and Todd (2005). Specifically, we analyse the original LaLonde experimental sample (297 treated and 425 untreated observations) and use the Panel Study of Income Dynamics as the comparison data (2490 observations). The pretreatment covariates in the data include age, education, race (white, black or Hispanic), marriage status, high school degree, earnings in 1974 and earnings in 1975. The outcome variable of interest is earnings in 1978. In this sample, the experimental benchmark for the average treatment effect is \$886 with a standard error of \$488.

Following Smith and Todd's (2005) analysis, we first estimate the 'evaluation bias', which is defined as the average effect of being in the experimental sample on 1978 earnings. Specifically, we estimate the conditional probability of being in the experimental sample given the pretreatment covariates. On the basis of this estimated propensity score, we match the control observations in the experimental sample with observations in the non-experimental sample. Since neither group of workers received the job training programme, the true average treatment effect is zero. We conduct 1-to-1 nearest neighbour matching with replacement where matching is done on the log-odds of the estimated propensity score. Whereas Smith and Todd trimmed the data to improve the credibility of the overlap assumption, we avoid doing so in this illustration for simplicity and transparency. In addition to the 1-to-1 matching, we also conduct the optimal 1-to-N nearest neighbour matching with replacement where N is chosen on the basis of the value of the J-statistic in equation (19) calculated after refitting the overidentified CBPS to each matched subset, which is obtained on the basis of the initial fit of the overidentified CBPS to the entire data, with N ranging from 1 to 10.

Finally, to examine the sensitivity to the propensity score model specification, we follow Smith and Todd's (2005) analysis and fit three different logistic regression models—a linear specification, a quadratic specification that includes the squares of non-binary covariates and the specification that was used by Smith and Todd which is based on Dehejia and Wahba's (1999) variable selection procedure and adds an interaction term between Hispanic and zero earnings in 1974 to the quadratic specification. Our standard errors are based on Abadie and Imbens (2006) rather than the bootstrap that was used by Smith and Todd.

Table 2 presents the estimated evaluation bias of 1-to-1 and optimal 1-to-N nearest neighbour propensity score matching with replacement across three different propensity score model specifications described above. The results are compared across the propensity score estimation methods—the standard logistic regression (GLM), the just-identified CBPS with balance condition only (CBPS1) and the overidentified CBPS that combines the likelihood and balance conditions (CBPS2). Across all specifications and matching methods, the CBPS substantially improves the performance of the propensity score matching estimator based on the standard logistic regression. In particular, the overidentified CBPS has the smallest evaluation bias across most propensity score model specifications. The just-identified CBPS also has a smaller bias than the standard logistic regression.

To see where the dramatic improvement of the CBPS comes from, we examine the covariate balance across the estimation methods and model specifications. Table 3 presents the log-likelihood as well as the multivariate standardized bias statistics for the treated observations

Table 2. Estimated evaluation bias of 1-to-1 and optimal 1-to-*N* nearest neighbour propensity score matching estimators with replacement†

Model specification	*Results for 1-to-1 nearest neighbour estimator*			*Results for optimal 1-to-N nearest neighbour estimator*		
	GLM	*CBPS1*	*CBPS2*	*GLM*	*CBPS1*	*CBPS2*
Linear	−1209.15	−654.79	−505.15	−1209.15	−654.79	−130.84
	(1426.44)	(1247.55)	(1335.47)	(1426.44)	(1247.55)	(1335.47)
Quadratic	−1439.14	−955.30	−216.73	−1234.33	−175.92	−658.61
	(1299.05)	(1496.27)	(1285.28)	(1074.88)	(943.34)	(1041.47)
Smith and Todd (2005)	−1437.69	−820.89	−640.99	−1229.81	−826.53	−464.06
	(1256.84)	(1229.63)	(1757.09)	(1044.15)	(1179.73)	(1130.73)

†The results (standard errors are in parentheses) represent the estimated average effect of being in the experimental sample (i.e. the estimated evaluation bias) on the 1978 earnings where the experimental control group is compared with the matched subset of the untreated non-experimental sample. If a matching estimator is successful, then its estimated effect should be close to the true effect, which is zero. The propensity score is estimated in three different ways using the logistic regression—standard logistic regression (GLM), the just-identified CBPS with the balance conditions alone (CBPS1) and the overidentified CBPS which combines the score equation and the balance conditions (CBPS2). Three different logistic propensity score model specifications are considered—the linear, and the quadratic function of covariates, and the model specification that was used by Smith and Todd (2005). Across model specifications and matching estimators, the CBPS substantially improves on the GLM. In addition, CBPS2 with the optimal 1-to-*N* matching generally has the smallest bias.

(as opposed to the log-likelihood for the entire sample that was given in equation (20)) defined as

$$\text{imbalance} = \left\{ \left(\frac{1}{N} \sum_{i=1}^{N} w_{\hat{\beta}}(T_i, X_i) X_i \right)^{\mathrm{T}} \left(\frac{1}{N_1} \sum_{i=1}^{N} T_i X_i X_i^{\mathrm{T}} \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} w_{\hat{\beta}}(T_i, X_i) X_i \right) \right\}^{1/2} \quad (21)$$

for the entire sample (the top panel) and the optimal 1-to-*N* matched sample (the bottom panel). In addition to these statistics, we also present the average Mahalanobis distance for matched pairs and the L_1 imbalance statistic (Iacus *et al.*, 2011) for the matched sample.

As seen in Section 3.1, the CBPS significantly improves the covariate balance when compared with the standard logistic regression (GLM) while sacrificing some degree of likelihood. This is true across model specifications. This gain is achieved without sacrificing covariate balance measured by using the Mahalanobis distance or the L_1 imbalance statistic. We note that the overidentifying restriction test that was described in Section 2.3 does not reject the null hypothesis that the propensity score model specification is correct. The J-statistic is equal to 6.8 (22 degrees of freedom), 7.7 (30 degrees of freedom) and 8.1 (32 degrees of freedom) for the linear, quadratic and the Smith and Todd (2005) model specifications respectively. This may explain the fact that the estimates based on the CBPS, although still biased by several hundred dollars, are fairly stable across model specifications and offer a noticeable improvement when compared with the standard logistic regression estimates.

Finally, we compare the estimated average treatment effect based on the propensity score matching estimators with its experimental estimate. Note that this estimate is not the true treatment effect. For this reason, Smith and Todd (2005) focused on the estimation of evaluation bias, but others, including LaLonde (1986) and Dehejia and Wahba (1999), studied the differences between the experimental estimates and the estimates based on various statistical methods. As explained by Smith and Todd, the LaLonde experimental sample combined with the Panel

Table 3. Covariate balance across different propensity score estimation methods and model specifications†

Measure	*Results for linear model*			*Results for quadratic model*			*Results for Smith and Todd (2005) model*		
	GLM	*CBPS1*	*CBPS2*	*GLM*	*CBPS1*	*CBPS2*	*GLM*	*CBPS1*	*CBPS2*
Entire sample									
Log-likelihood	−547	−564	−558	−521	−551	−534	−520	−550	−534
Overall imbalance	5.106	0.000	0.538	4.338	0.000	1.109	4.359	0.000	1.137
Matched sample									
Overall imbalance	0.563	0.490	0.359	0.767	0.166	0.075	0.216	0.162	0.056
Mahalanobis	0.020	0.020	0.020	0.074	0.011	0.042	0.039	0.016	0.039
L_1-imbalance	0.936	0.929	0.925	0.928	0.925	0.929	0.938	0.933	0.926

†We compare three different ways of estimating the propensity score—standard logistic regression (GLM), the just-identified CBPS with the balance conditions alone (CBPS1) and the overidentified CBPS that combines the score equation with the balancing conditions (CBPS2). For each estimation method, we consider three logistic regression specifications—the linear, the quadratic function of covariates, and the model specification that was used by Smith and Todd (2005). Given each specification, we report the 'overall imbalance' which represents the multivariate standardized bias statistic defined in equation (21) for the entire sample (upper panel) and for the optimal 1-to-N nearest neighbour matched sample (bottom panel). In addition, the log-likelihood is presented for the entire sample, and Mahalanobis distance and L_1 imbalance statistics are shown for the matched sample. When compared with the GLM method the CBPS significantly reduces the overall imbalance without sacrificing covariance measured by the Mahalanobis and L_1-distance measures.

Table 4. Comparison between the estimated average treatment effects from matching estimators and the experimental estimate†

Model specification	*Results for 1-to-1 nearest neighbour*			*Results for optimal 1-to-N nearest neighbour*		
	GLM	*CBPS1*	*CBPS2*	*GLM*	*CBPS1*	*CBPS2*
Linear	−304.92	423.30	183.67	−211.07	423.30	138.20
	(1437.02)	(1295.19)	(1240.79)	(1201.49)	(1110.26)	(1161.91)
Quadratic	−922.16	239.46	1093.13	−715.54	307.51	185.57
	(1382.38)	(1284.13)	(1567.33)	(1145.82)	(1158.06)	(1247.99)
Smith and Todd (2005)	−734.49	−269.07	423.76	−439.54	−617.68	690.09
	(1424.57)	(1711.66)	(1404.15)	(1259.28)	(1438.86)	(1288.68)

†The experimental estimate is 886 with the standard error of 488. As in Table 2, we compare three different ways of estimating the propensity score—the standard logistic regression (GLM), the just-identified CBPS with the balance conditions alone (CBPS1) and the overidentified CBPS that combines the score equation with the balancing conditions (CBPS2). For each estimation method, we consider three logistic regression specifications—the linear, the quadratic function of covariates and the model specification that was used by Smith and Todd (2005). The standard errors are in parentheses. The CBPS generally yields the estimates that are much closer to the experimental estimate when compared with standard logistic regression.

Study of Income Dynamics comparison sample that we analyse present a particularly difficult selection bias problem to be overcome. In the literature, for example, Diamond and Sekhon (2012) analysed this same sample by using the genetic matching estimator and presented the estimate of −571 (with a standard error of 1130), which is somewhat far from the experimental estimate, which is 866 with a standard error of 488. In contrast, when applied to other samples, various methods including genetic matching yielded estimates that are generally much closer to

the experimental estimate (Dehejia and Wahba, 1999; Diamond and Sekhon, 2012; Hainmueller, 2012).

We base our matching estimates of the average treatment effects on the same propensity scores as those used to generate the estimates of evaluation bias given in Table 2. As before, we then conduct the 1-to-1 and optimal 1-to-N nearest neighbour matching with replacement based on the log-odds of the estimated propensity score except that we now match the treated experimental units with the (untreated) observational units. Table 4 presents the results. Although the standard errors are large, one clear pattern emerges from these results. The CBPS with or without the score equation generally yields the matching estimates that are much closer to the experimental estimate when compared with the standard logistic regression (GLM). This pattern is consistent with those observed in our analysis of evaluation bias (see Table 2). In sum, although substantial differences remain between the experimental and matching estimates, the CBPS generates effect estimates with substantively less bias than the standard logistic regression.

4. Extensions

We have shown that the CBPS can dramatically improve the performance of propensity score weighting and matching estimators when estimating the average treatment effects in observational studies. Another important advantage of the CBPS is that it can be extended to other important causal inference settings by directly incorporating various balancing conditions. In this section, we briefly describe a couple of potential extensions of the CBPS.

4.1. Generalized propensity score for multivalued treatments

First, we extend the CBPS to causal inference with multivalued treatments. Suppose that we have a multivalued treatment where the treatment variable T_i takes one of the K integer values, i.e. $T_i \in \mathcal{T} = \{0, \ldots, K-1\}$ where $K \geqslant 2$. The binary treatment case that was considered in Section 2 is a special case with $K = 2$. Following Imbens (2000), we can define the generalized propensity score as the multinomial probabilities

$$\pi_\beta^k(X_i) = \Pr(T_i = k | X_i) \tag{22}$$

where all conditional probabilities sum to 1, i.e. $\Sigma_{k=0}^{K-1} \pi_\beta^k(X_i) = 1$. For example, we may use multinomial logistic regression to model this generalized propensity score. As in the binary case, we have the moment condition that is based on the score function under the likelihood framework:

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \left[\frac{\mathbf{1}\{T_i = k\}}{\pi_\beta^k(X_i)} \cdot \frac{\partial \pi_\beta^k(X_i)}{\partial \beta^{\mathrm{T}}} \right] = 0. \tag{23}$$

Regarding the balancing conditions, weighting covariates by their inverse will balance them across all treatment levels. This fact yields the following $K-1$ sets of moment conditions:

$$\frac{1}{N} \sum_{i=1}^{N} \left[\frac{\mathbf{1}\{T_i = k\} \tilde{X}_i}{\pi_\beta^k(X_i)} - \frac{\mathbf{1}\{T_i = k-1\} \tilde{X}_i}{\pi_\beta^{k-1}(X_i)} \right] = 0 \tag{24}$$

for each of $k = 1, \ldots, K-1$. As before, these moment conditions can be combined with that of equation (23) under the GMM or EL framework.

4.2. Generalizing experimental results

Another possible extension of the CBPS concerns the generalization of experimental results to a target population when an experimental sample is not representative. Cole and Stuart (2010) and Stuart *et al.* (2011) used the propensity score to generalize experimental results (see also Imai and Ratkovic (2013)). Suppose that we have an experimental sample of N_e units where the binary treatment variable T_i is completely randomized. Let S_i represent the sampling indicator where $S_i = 1$ if unit i is in the experimental sample and $S_i = 0$ otherwise. In this context, the 'propensity score' is defined as the conditional probability of being in the experimental sample given the pretreatment characteristics

$$\pi_\beta(X_i) = \Pr(S_i = 1 | X_i). \tag{25}$$

In addition to the experimental sample, we assume that a random sample that is representative of the target population $\mathcal{P}$ is available and its sample size is N_{ne}. Without loss of generality, we assume that the first N_e units belong to the experimental sample $S_i = 1$ for $i = 1, \ldots, N_e$, and the last N_{ne} units belong to the non-experimental sample $S_i = 0$ for $i = N_e + 1, \ldots, N$ where $N = N_e + N_{ne}$ represents the total sample size. The assumption that makes the generalization of experimental results possible is $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp S_i | X_i$ and $0 < \pi_\beta(X_i) < 1$, which together imply that the sample selection bias can be eliminated by conditioning on X_i. Under this assumption, the propensity score $\pi_\beta(X_i)$ is estimated by fitting, for example, a logistic regression with S_i as the response variable. Similar to equation (7), the moment condition from this model is the score function

$$\frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{S_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1 - S_i) \pi'_\beta(X_i)}{1 - \pi_\beta(X_i)} \right\} = 0. \tag{26}$$

In addition, if the propensity score is correct, then appropriately weighting the covariates in the experimental sample will make their distribution similar to that of the weighted non-experimental sample,

$$\frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{S_i \tilde{X}_i}{\pi_\beta(X_i)} - \frac{(1 - S_i) \tilde{X}_i}{1 - \pi_\beta(X_i)} \right\} = 0, \tag{27}$$

where $\tilde{X}_i = f(X_i)$ is an M-dimensional vector-valued function of covariates.

Finally, the moment conditions given in equations (26) and (27) can be combined under the GMM or EL framework as done in Section 2 to estimate the propensity score.

5. Concluding remarks

Propensity score matching and weighting methods have become popular tools for applied researchers in various disciplines who conduct causal inference in observational studies. The propensity score methodology also has been extended to various other settings including longitudinal data, non-binary treatment regimes and the generalization of experimental results. Despite this development, little attention has been paid to the question of how the propensity score should be estimated (see McCaffrey *et al.* (2004) for an exception). This is unfortunate because a slight misspecification of the propensity score model can result in substantial bias of estimated treatment effects.

The CBPS methodology proposed enables the robust and efficient parametric estimation of the propensity score by directly incorporating the key covariate balancing property of propensity

scores. We exploit this property and estimate the propensity score within the familiar framework of GMMs or EL.

Although we provide some empirical evidence that the CBPS can dramatically improve the performance of propensity score weighting and matching methods, some remaining issues merit future research. First, although the CBPS is relatively robust to model misspecification, its successful application requires scientists to identify a complete set of confounders. Thus, model selection methods should be developed for propensity score estimation. Second, there are various potential extensions beyond those described in this paper. For example, the CBPS can be applied to improve propensity score estimation for marginal structural models in the longitudinal setting (Robins *et al.*, 2000). Another potential application is the estimation of average treatment effects from the instrumental variable estimate (Angrist and Fernandez-Val, 2010). These and other potential extensions are currently being explored by us.

Acknowledgements

Financial support from the National Science Foundation (grants SES–0550873; SES–0752050) is acknowledged. Jeffrey Smith and Petra Todd generously provided the data that are analysed in this paper. We thank Jens Hainmueller, Mark Handcock, Gary King, Dylan Small, Richard Wyss and seminar participants at the Inter-American Development Bank, the Joint Statistical Meetings, University of California at Los Angeles, and Princeton University for helpful suggestions.

References

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

Abadie, A. and Imbens, G. W. (2008) On the failure of the bootstrap for matching estimators. *Econometrica*, **76**, 1537–1557.

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

Angrist, J. and Fernandez-Val, I. (2010) ExtrapoLATE-ing: external validity and overidentification in the LATE framework. *Working Paper 16566*. National Bureau of Economic Research, Cambridge.

Camillo, F. and D'Attoma, I. (2011) A multivariate strategy to measure and test global imbalance. *Exprt Syst. Applic.*, **38**, 3451–3460.

Chaudhuri, S., Handcock, M. S. and Rendall, M. S. (2008) Generalized linear models incorporating population level information: an empirical-likelihood-based approach. *J. R. Statist. Soc.* B, **70**, 311–328.

Cole, S. R. and Stuart, E. A. (2010) Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am. J. Epidem.*, **172**, 107–115.

Dehejia, R. H. and Wahba, S. (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J. Am. Statist. Ass.*, **94**, 1053–1062.

Deming, W. E. and Stephan, F. F. (1940) On a least squares adjustment of a sampled frequency when the expected marginal tables are known. *Ann. Math. Statist.*, **11**, 427–444.

Diamond, A. and Sekhon, J. (2012) Genetic matching for estimating causal effects: a new method of achieving balance in observational studies. *Rev. Econ. Statist.*, to be published.

Freedman, D. A. and Berk, R. A. (2008) Weighting regressions by propensity scores. *Evaln Rev.*, **32**, 392–409.

Graham, B. S., Campos de Xavier Pinto, C. and Egel, D. (2012) Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.*, **79**, 1053–1079.

Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, **66**, 315–331.

Hainmueller, J. (2008) Synthetic matching for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Technical Report*. Department of Government, Harvard University, Cambridge.

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

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

Hansen, L. P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.
Hansen, L. P., Heaton, J. and Yaron, A. (1996) Finite-sample properties of some alternative GMM estimators. *J. Bus. Econ. Statist.*, **14**, 262–280.
Hayashi, F. (2000) *Econometrics*. Princeton: Princeton University Press.
Heckman, J. J., Ichimura, H. and Todd, P. (1998) Matching as an econometric evaluation estimator. *Rev. Econ. Stud.*, **65**, 261–294.
Hellerstein, J. K. and Imbens, G. W. (1999) Imposing moment restrictions from auxiliary data by weighting. *Rev. Econ. Statist.*, **81**, 1–14.
Hirano, K., Imbens, G. and Ridder, G. (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **71**, 1307–1338.
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.
Horvitz, D. and Thompson, D. (1952) A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Ass.*, **47**, 663–685.
Iacus, S., King, G. and Porro, G. (2011) Multivariate matching methods that are monotonic imbalance bounding. *J. Am. Statist. Ass.*, **106**, 345–361.
Imai, K. and van Dyk, D. A. (2004) Causal inference with general treatment regimes: generalizing the propensity score. *J. Am. Statist. Ass.*, **99**, 854–866.
Imai, K., King, G. and Stuart, E. A. (2008) Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Statist. Soc.* A, **171**, 481–502.
Imai, K. and Ratkovic, M. (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Statist.*, **7**, 443–470.
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. Statist.*, **86**, 4–29.
Kang, J. D. and Schafer, J. L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussions). *Statist. Sci.*, **22**, 523–539.
La Londe, R. J. (1986) Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.*, **76**, 604–620.
Little, R. J. A. and Wu, M. M. (1991) Models for contingency tables with known margins when target and sampled populations differ. *J. Am. Statist. Ass.*, **86**, 87–95.
Lunceford, J. K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statist. Med.*, **23**, 2937–2960.
McCaffrey, D. F., Ridgeway, G. and Morral, A. R. (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Meth.*, **9**, 403–425.
Nevo, A. (2003) Using weights to adjust for sample selection when auxiliary information is available. *J. Bus. Econ. Statist.*, **21**, 43–52.
Newey, W. and McFadden, D. (1994) Large sample estimation and hypothesis testing. In *Handbook of Econometrics* (eds R.F. Engle and D. L. McFadden), vol. IV, pp. 2111–2245. Amsterdam: North-Holland.
Oh, H. and Scheuren, F. (1983) Weighting adjustments for unit non-response. In *Incomplete Data in Sample Surveys*, vol. II, *Theory and Annotated Bibliography* (eds W. Madow, I. Olkin and D. Rubin). New York: Academic Press.
Owen, A. B. (2001) *Empirical Likelihood*. Boca Raton: Chapman and Hall–CRC.
Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.
Ratkovic, M. (2012) Achieving optimal covariate balance under general treatment regimes. *Working Paper*. Department of Politics, Princeton University, Princeton.
Ratkovic, M., Imai, K. and Fong, C. (2012) Cbps: R package for covariate balancing propensity score. (Available from `http://CRAN.R-project.org/package=CBPS`.)
Ridgeway, G. and McCaffrey, D. F. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, **22**, 540–543.
Robins, J. M., Hernán, M. A. and Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.
Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–866.
Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Ass.*, **90**, 106–121.
Robins, J., Sued, M., Lei-Gomez, Q. and Rotnitzky, A. (2007) Performance of double-robust estimators when 'inverse probability' weights are highly variable. *Statist. Sci.*, **22**, 544–559.
Rosenbaum, P. R. (1987) Model-based direct adjustment. *J. Am. Statist. Ass.*, **82**, 387–394.
Rosenbaum, P. R. (1989) Optimal matching for observational studies. *J. Am. Statist. Ass.*, **84**, 1024–1032.

Rosenbaum, P. R. (1991) A characterization of optimal designs for observational studies. *J. R. Statist. Soc.* B, **53**, 597–610.
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. Am. Statist. Ass.*, **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. *Am. Statistn*, **39**, 33–38.
Rubin, D. B. (2007) The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statist. Med.*, **26**, 20–36.
Smith, J. A. and Todd, P. E. (2005) Does matching overcome LaLonde's critique of nonexperimental estimators? *J. Econmetr.*, **125**, 305–353.
Stuart, E. A. (2010) Matching methods for causal inference: a review and a look forward. *Statist. Sci.*, **25**, 1–21.
Stuart, E. A., Cole, S. R., Bradshaw, C. P. and Leaf, P. J. (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Statist. Soc.* A, **174**, 369–386.
Tan, Z. (2010) Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, **97**, 661–682.