Advance Access publication October 16, 2011

Political Analysis (2012) 20:25–46
doi:10.1093/pan/mpr025

Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies

Jens Hainmueller
Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139
e-mail: jhainm@mit.edu

Edited by R. Michael Alvarez

This paper proposes entropy balancing, a data preprocessing method to achieve covariate balance in observational studies with binary treatments. Entropy balancing relies on a maximum entropy reweighting scheme that calibrates unit weights so that the reweighted treatment and control group satisfy a potentially large set of prespecified balance conditions that incorporate information about known sample moments. Entropy balancing thereby exactly adjusts inequalities in representation with respect to the first, second, and possibly higher moments of the covariate distributions. These balance improvements can reduce model dependence for the subsequent estimation of treatment effects. The method assures that balance improves on all covariate moments included in the reweighting. It also obviates the need for continual balance checking and iterative searching over propensity score models that may stochastically balance the covariate moments. We demonstrate the use of entropy balancing with Monte Carlo simulations and empirical applications.

1 Introduction

Matching and propensity score methods are nowadays often used in observational studies in political science and other disciplines to preprocess the data prior to the estimation of binary treatment effects under the assumption of selection on observables (Ho et al. 2007; Sekhon 2009). The preprocessing step involves reweighting or simply discarding units to improve the covariate balance between the treatment and control group such that the treatment variable becomes closer to being independent of the background characteristics. This reduces model dependence for the subsequent estimation of treatment effects with regression or other standard estimators in the preprocessed data (Abadie and Imbens 2007; Ho et al. 2007).

Although preprocessing methods are gaining ground in applied work, there exists no scholarly consensus in the methodological literature about how the preprocessing step is best conducted. One important concern is that many commonly used preprocessing approaches do not directly focus on the goal of producing covariate balance. In the most widely used practice, researchers "manually" iterate between propensity score modeling, matching, and balance checking until they attain a satisfactory balancing solution. The hope is that an accurately estimated propensity score will stochastically balance the covariates, but this requires finding the correct model specification and often fairly large samples. As a result of this

Author's note: The author is also affiliated with Harvard's Institute for Quantitative Social Science (IQSS). I thank Alberto Abadie, Alexis Diamond, Andy Eggers, Adam Glynn, Don Green, Justin Grimmer, Dominik Hangartner, Dan Hopkins, Kosuke Imai, Guido Imbens, Gary King, Gabe Lenz, Jasjeet Sekhon, and Yiqing Xu for very helpful comments. I would especially like to thank Alan Zaslavsky, who inspired this project. I would also like to thank the editor R. Michael Alvarez and our three anonymous reviewers for their excellent suggestions. The usual disclaimer applies.

Companion software (for Stata and R) that implements the methods proposed in this paper is provided on the author's webpage. Replication materials are available in the Political Analysis Dataverse at http://dvn.iq.harvard.edu/dvn/dv/pan. Supplementary materials for this article are available on the *Political Analysis* Web site.

intricate search process, low balance levels prevail in many studies and the user experience can be tedious. Even worse, matching may counteract bias reduction for the subsequent treatment effect estimation when improving balance on some covariates decreases balance on other covariates (see Diamond and Sekhon 2006, Ho et al. 2007, and Iacus, King, and Porro 2009 for similar critiques).

In this study, we propose entropy balancing as a preprocessing technique for researchers to achieve covariate balance in observational studies with a binary treatment. In contrast to most other preprocessing methods, entropy balancing involves a reweighting scheme that directly incorporates covariate balance into the weight function that is applied to the sample units. The researcher begins by imposing a potentially large set of balance constraints, which imply that the covariate distributions of the treatment and control group in the preprocessed data match exactly on all prespecified moments. After the researcher has prespecified her desired level of covariate balance, entropy balancing searches for the set of weights that satisfies the balance constraints but remains as close as possible (in an entropy sense) to a set of uniform base weights to retain information. This recalibration of the unit weights effectively adjusts for systematic and random inequalities in representation.

This procedure has several attractive features. Most importantly, entropy balancing allows the researcher to obtain a high degree of covariate balance by imposing a potentially large set of balance constraints that involve the first, second, and possibly higher moments of the covariate distributions as well as interactions. Entropy balancing always (at least weakly) improves upon the balance that can be obtained by conventional preprocessing adjustments with respect to the specified balance constraints. This is because the reweighting scheme directly incorporates the researcher's knowledge about the known sample moments and balances them exactly in finite samples (analogous to similar reweighting procedures in survey research that improve inferences about unknown population features by adjusting the sample to some known population features). This obviates the need for balance checking in the conventional sense, at least for the characteristics that are included in the specified balance constraints.

A second advantage of entropy balancing is that it retains valuable information in the preprocessed data by allowing the unit weights to vary smoothly across units. In contrast to other preprocessing methods such as nearest neighbor matching where units are either discarded or matched (weights of zero or one)[1] , the reweighting scheme in entropy balancing is more flexible: It reweights units appropriately to achieve balance, but at the same time keeps the weights as close as possible to the base weights to prevent loss of information and thereby retains efficiency for the subsequent analysis. In this regard, entropy balancing provides a generalization of the propensity score weighting approach (Hirano, Imbens, and Ridder 2003) where the researcher first estimates the propensity score weights with a logistic regression and then computes balance checks to see if the estimated weights equalize the covariate distributions. In practice, such estimated propensity score weights can fail to balance the covariate moments in finite samples. Entropy balancing in contrast directly adjusts the weights to the known sample moments and thereby obviates the need for continual balance checking and iterative searching over propensity score models that may stochastically balance the prespecified covariates.

A third advantage of entropy balancing is that the approach is fairly versatile. The weights that result from entropy balancing can be passed to almost any standard estimator for the subsequent estimation of treatment effects. This may include a simple (weighted) difference in means, a weighted least squares regression of the outcome on the treatment variable and possibly additional covariates that are not included as part of the reweighting, or whatever other standard statistical model the researcher would have applied in the absence of any preprocessing. Since entropy balancing orthogonalizes the treatment indicator with respect to the covariates that are included in the balance constraints, the resulting estimates in the preprocessed data can exhibit lower model dependency compared to estimates from the unadjusted data.

Lastly, entropy balancing is also computationally attractive since the optimization problem to find the unit weights is well behaved and globally convex; the algorithm attains the weighting solution within seconds even for moderately large data sets that may be encountered in political science applications (assuming that the balance constraints are feasible).

We show three Monte Carlo simulations that demonstrate the desirable finite sample properties of entropy balancing in several benchmark settings where the method improves in root mean squared error

[1]In practice, the weights may sometimes differ from zero or one in the case of ties or for controls units that are matched several times when matching with replacement.

(MSE) upon a variety of widely used preprocessing adjustments (including Mahalanobis distance matching, genetic matching, and matching or weighting on a logistic propensity score). We also illustrate the use of entropy balancing in two empirical settings including a validation exercise in the LaLonde (1986) data set and a reanalysis of the data used by Ladd and Lenz (2009) to examine the effect of newspaper endorsements on vote choice in the 1997 British general election. Two additional applications that consider the impact of media bias on voting (DellaVigna and Kaplan 2007) and the financial returns to political office (Eggers and Hainmueller 2009) are provided in a web appendix.[2] Entropy balancing yields high levels of covariate balance (as measured by standard metrics) in all four data sets and reduces model dependency for the subsequent estimation of the treatment effects.

Although entropy balancing provides a reweighting scheme for the context of causal inference in observational studies with a binary treatment (where the goal is to equate the covariate distributions across the treatment and the control group), important links exist between the reweighting scheme employed in entropy balancing and various strands of literatures in econometrics and statistics. In particular, the method heavily borrows from the survey literature that contains several reweighting schemes which are used to adjust sampling weights so that sample totals match population totals known from auxiliary data (see Särndal and Lundström 2006 for a recent review and earlier work by Deming and Stephan 1940, Ireland and Kullback 1968, Oh and Scheuren 1978, and Zaslavsky 1988 who proposed a similar log-linear reweighting scheme to adjust for undercount in census data). More broadly, similar reweighting schemes are also widely used in the literature on methods of moments estimation, empirical likelihood, exponential tilting, and missing data (Hansen 1982; Qin and Lawless 1994; Kitamura and Stutzer 1997; Imbens 1997; Imbens, Spady, and Johnson 1998; Hellerstein and Imbens 1999; Owen 2001; Schennach 2007; Qin, Zhang, and Leung 2009; Graham, Pinto, and Egel 2010).

2 Observational Studies with Binary Treatments

2.1 *Framework*

We consider a random sample of $n = n_1 + n_0$ units drawn from a population of size $N = N_1 + N_0$, where $n \leqslant N$ and N_1 and N_0 refer to the size of the target population of treated units and the source population of control units, respectively. Each unit i is exposed to a binary treatment $D_i \in \{1, 0\}$; $D_i = 1$ if unit i received the active treatment and $D_i = 0$ if unit i received the control treatment. In the sample, we have n_1 treated units and n_0 control units. Let X be a matrix of J exogenous pretreatment characteristics; entry X_{ij} refers to the value of the jth characteristic for unit i so that $X_i = [X_{i1}, X_{i2}, \ldots, X_{iJ}]$ is the row vector of characteristics for unit i and X_j is the column vector that captures the jth characteristic across units accordingly. Let $f_{X|D=1}$ and $f_{X|D=0}$ denote the densities of these covariates in the treatment and control population, respectively. Finally, let $Y_i(D_i)$ denote the pair of potential outcomes that individual i attains if it is exposed to the active treatment or the control treatment. Observed outcomes for each individual are realized as $Y_i = Y_i(1)\, D_i + (1 - D_i) Y_i(0)$ so that we never observe both potential outcomes simultaneously but the triple (D_i, Y_i, X_i).

The treatment effect for each unit is defined as $\tau_i = Y_i(1) - Y_i(0)$. Many causal quantities of interest are defined as functions of τ_i for different subsets of units.[3] Most common are the sample (SATE) and population (PATE) average treatment effects given by $\text{SATE} = n^{-1} \sum_i^n \tau_i$ and $\text{PATE} = N^{-1} \sum_i^N \tau_i$ and the sample (SATT) and population (PATT) average treatment effect on the treated given by $\text{SATT} = n_1^{-1} \sum_{\{i|D=1\}}^n \tau_i$ and $\text{PATT} = N_1^{-1} \sum_{\{i|D=1\}}^N \tau_i$. Notice that $\mathbb{E}[\text{SATE}] = \text{PATE} = \mathbb{E}[Y(1) - Y(0)]$ and similarly $\mathbb{E}[\text{SATT}] = \text{PATT} = \mathbb{E}[Y(1) - Y(0)|D = 1]$ since we consider random samples. Following the preprocessing literature, we focus on the PATT as our quantity of interest. The entropy balancing methods described below are also applicable to estimate the PATE and other commonly used quantities of interest analogously.[4]

The PATT is given by $\tau = \mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 1]$. The first expectation is easily estimable from the treatment group data. The second expectation, $\mathbb{E}[Y(0)|D = 1]$, is counterfactual and

[2]This web appendix is available on the authors webpage at http://www.mit.edu/ jhainm/research.htm.

[3]Note that some other causal quantities of interest are not defined in this way (e.g., causal mediation or necessary causation).

[4]For example, the treatment group can be reweighted to match the control group. An important caveat is that it may be more difficult to estimate the PATE or SATE due to limited overlap in the covariate distributions.

thus unobserved even in the target population. The only information available about $Y(0)$ is in the sample from the source population not exposed to the treatment (the control group). In experimental studies, where treatment assignment is forced to be independent of the potential outcomes, $Y(1), Y(0) \perp D$, we can simply use $\mathbb{E}[Y(0)|D = 0]$ as our estimate of $\mathbb{E}[Y(0)|D = 1]$. In observational studies, however, selection into treatment usually renders the latter two quantities unequal. The conventional solution to this problem is to assume ignorable treatment assignment and overlap (Rosenbaum and Rubin 1983), which implies that $Y(0) \perp D|X$ and that $\Pr(D = 1|X = x) < 1$ for all x in the support of $f_{X|D=1}$. Therefore, conditional on all confounding covariates X, the potential outcomes are stochastically independent of D and the PATT is identified as

$$\tau = \mathbb{E}[Y|D = 1] - \int \mathbb{E}[Y|X = x, D = 0] f_{X|D=1}(x) \mathrm{d}x,$$

where the integral is taken over the support of X in the source population. Notice that the last term in this expression is equal to the covariate adjusted mean, that is, the estimated mean of Y in the source population if its covariates were distributed as in the target population (Frölich 2007).

To see why covariate balance is key for the estimation of the PATT, notice that the potential outcomes for the treated units can be written as $Y_i(D_i) = l(X_i)$, where $l()$ is an unknown function. For simplicity, suppose that the treatment effect is estimated by the difference in means. The treatment effect can then be decomposed into the estimated treatment effect and the average estimation error:

$$\mathrm{PATT} = \widehat{\mathrm{PATT}} + N_1^{-1} \sum_{\{i|D=1\}} (l_0(X_{\{i|D=0\}}) - l_0(X_{\{i|D=1\}})),$$

where $l_0(X_{\{i|D=0\}}) - l_0(X_{\{i|D=1\}}) = \hat{Y}_i(0) - Y_i(0)$ is the unit level treatment error (see Iacus, King, and Porro 2009). The estimation error has two components: (1) the unknown function $l()$, which determines the importance of the variables, and (2) the imbalance, which is defined as the difference between the empirical covariate distributions of the treatment $f_{X|D=1}$ and the control group $f_{X|D=0}$.

Data preprocessing procedures such as matching and related approaches involve reweighting or simply discarding units to reduce the imbalance in the covariate distributions to decrease the error and model dependency for the subsequent estimation of the treatment effect. As Ho et al. (2007, 209) put it, "the goal of matching is to achieve the best balance for a large number of observations, using any method of matching that is a function of X, so long as we do not consult Y."[5] A variety of such preprocessing procedures have been proposed (Imbens 2004; Rubin 2006; Ho et al. 2007; Sekhon 2009). If X is low dimensional, the units can simply be matched exactly on the covariates. However, selection on observables is often only plausible after conditioning on many confounders and if X is fairly high dimensional then the curse of dimensionality can render exact matching infeasible. However, as shown by Rosenbaum and Rubin (1983), the preprocessing problem may be reduced to a single dimension given that the counterfactual mean can also be identified as

$$\mathbb{E}[Y(0)|D = 1] = \int \mathbb{E}[Y|p(X) = \rho, D = 0] f_{p|D=1}(\rho) \mathrm{d}\rho$$

where $f_{p|D=1}$ is the distribution of the propensity score $p(x) = \Pr(D = 1|X = x)$ in the target population. This follows from their result that under selection on observables $Y(0) \perp D|X$ is equal to

[5]Notice that there is some debate about how to assess covariate balance in practice. Theoretically, we would like the two empirical distributions to be equal so that the density in the preprocessed control group $f^*_{X|D=0}$ mirrors the density in the treatment group $f_{X|D=1}$. Comparing the joint empirical distributions of all covariates X is difficult when X is high dimensional and therefore lower dimensional balance metrics are commonly used (but see Iacus, King, and Porro 2009 who propose a multidimensional metric). Opinions differ on what metric is most appropriate. The most commonly used metric is the standardized difference in means (Rosenbaum and Rubin 1983) and t-tests for differences in means. Diamond and Sekhon (2006) argue that paired t-test and bootstrapped Kolmogorov–Smirnov (KS) tests should be used instead and that commonly used p value cutoffs such as .1 or .05 are too lenient to obtain reliable causal inferences. Rubin (2006) also considers variance ratios and tests for residuals that are orthogonalized to the propensity score. Imai, King, and Stuart (2008) criticize the use of t-tests and stopping rules and argue that all balance measures should be maximized without limit. They advocate QQ plot summary statistics as better alternatives than t-tests or KS tests. Sekhon (2006) comes to the opposite conclusion. Hansen and Bowers (2008) advocate the use of Fisher's randomization inference for balance checking.

$Y(0) \perp D|p(x)$ and this implies that balance on all covariates can be achieved by matching or weighting on the propensity score alone.

The procedure of particular interest here involves weighting on the propensity score as suggested by Hirano and Imbens (2001) and Hirano, Imbens, and Ridder (2003). In this method, the researcher first estimates a propensity score (usually by a logit or probit regression of the treatment indicator on the covariates) and then the units are weighted by the inverse of this estimated score for the subsequent analysis. For example, the counterfactual mean in the preprocessed data may be estimated using

$$\mathbb{E}[\widehat{Y(0)|D} = 1] = \frac{\sum_{\{i|D=0\}} Y_i \, d_i}{\sum_{\{i|D=0\}} d_i},$$

where every control unit receives a weight given by $d_i = \frac{\hat{p}(x_i)}{1-\hat{p}(x_i)}$. If the assignment probabilities are correctly estimated by the propensity score model, then the control observations will form a balanced sample with the treated observations in the reweighted data.[6] The idea is similar to the classic Horvitz–Thompson adjustment used in the survey literature where units are weighted by the inverse of the inclusion probabilities that result from the sampling design (Horvitz and Thompson 1952). This similarity between survey sampling weights and propensity score weights provides the entry point for the reweighting methods proposed below.

2.2 *Achieving Balance with Matching and Propensity Score Methods*

In principle, propensity score weighting has some attractive theoretical features compared to other adjustment techniques such as pair matching or propensity score matching. Hirano, Imbens, and Ridder (2003) show that weighting on the estimated propensity score achieves the semiparametric efficiency bound for the estimation of average causal effects as derived in Hahn (1998). This result requires sufficiently large samples and a propensity score that is sufficiently flexibly estimated to approximate the true propensity score.

However, in practice, this procedure suffers from the same drawbacks that plague all propensity score methods: the true propensity score is valuable because it is a "balancing score" that stochastically equalizes the distributions of all covariates between the two groups, but the true score is usually unknown and often difficult to estimate accurately enough to actually produce the desired covariate balance.[7] Several studies have demonstrated that misspecified propensity scores can lead to substantial bias for the subsequent estimation of treatment effects (Drake 1993; Smith and Todd 2001; Diamond and Sekhon 2006) because misspecified propensity scores can fail to balance the covariates distributions.

When estimating propensity scores in practice it is often difficult to jointly balance all covariates, especially in high-dimensional data with possibly complex assignment mechanisms. Applied researchers almost always rely on simple logit or probit models to estimate the propensity score and try to avoid misspecification by "manually" iterating between matching or weighting, propensity score modeling, and balance checking until a satisfactory balancing solution is reached. In other words, the resulting balance provides the appropriate yardstick to assess the accuracy of a propensity score model. Some researchers have criticized this cyclical process as the "propensity score tautology" (Imai, King, and Stuart 2008). The iterative process of tweaking the propensity score model and balance checking can be tedious and frequently results in low balance levels. Even worse, as Diamond and Sekhon (2006, 8) observe, a "significant shortcoming of common matching methods such as Mahalanobis distance and propensity score matching is that they may (and in practice, frequently do) make balance worse across measured potential confounders." Unless the distributions of the covariates are ellipsoidally symmetric or are mixtures of

[6] Formally propensity score reweighting exploits the following equalities: $\mathbb{E}\left[\frac{DY}{p(x)}\right] = \mathbb{E}\left[\frac{DY(1)}{p(x)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{DY(1)}{p(x)}|X\right]\right] = \mathbb{E}\left[\frac{p(x)Y(1)}{p(x)}\right] = \mathbb{E}[Y(1)]$ which uses the ignorability assumption in the second to last equality (Hirano and Imbens 2001; Hirano, Imbens, and Ridder 2003).

[7] Hirano, Imbens, and Ridder (2003) derive their result for a case where the propensity score is estimated using a nonparametric sieve estimator that approximates the true propensity score by a power series in all variables. Asymptotically, this series will converge to the true propensity score function if the powers increase with the sample size, but no results exist about the finite sample properties of this estimator. By the authors' own admission, this approach is computationally not very attractive.

proportional ellipsoidally symmetric distributions, there is no guarantee that the matching techniques will be equally percent bias reducing (EPBR). Therefore, the bias of some linear functions of X may be increased, whereas all univariate covariate means are closer after the preprocessing.[8] Also notice that even with a good propensity score model, imbalances often remain because stochastic balancing occurs only asymptotically. Chance imbalances may remain in finite samples and in these cases one may still improve the balance by enforcing balance constraints on the specified moments.

One way to improve the search for a better balancing score is to replace the logistic regression with a better estimation techniques for the assignment mechanism such as boosted regression (McCaffrey, Ridgeway, and Morral 2004) or kernel regression (Frölich 2007). Entropy balancing takes a different approach and directly focuses on covariate balance.

3 Entropy Balancing

Entropy balancing is a preprocessing procedure that allows researchers to create balanced samples for the subsequent estimation of treatment effects. The preprocessing consists of a reweighting scheme that assigns a scalar weight to each sample unit such that the reweighted groups satisfy a set of balance constraints that are imposed on the sample moments of the covariate distributions. The balance constraints ensure that the reweighted groups match exactly on the specified moments. The weights that result from entropy balancing can be passed to any standard model that the researcher may want to use to model the outcomes in the reweighted data—the subsequent effect analysis proceeds just like with survey sampling weights or weights that are estimated from a logistic propensity score covariate model. The preprocessing step can reduce the model dependence for the subsequent analysis since entropy balancing orthogonalizes the treatment indicator with respect to the covariate moments that are included in the reweighting.

3.1 *Entropy Balancing Scheme*

For convenience, we motivate entropy balancing for the simplest scenario where the researcher's goal is to reweight the control group to match the moments of the treatment group in order to subsequently estimate the PATT $\tau = \mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 1]$ using the difference in mean outcomes between the treatment group and the reweighted control group. In this case, the counterfactual mean may be estimated by

$$\mathbb{E}[\widehat{Y(0)|D} = 1] = \frac{\sum_{\{i|D=0\}} Y_i \, w_i}{\sum_{\{i|D=0\}} w_i}, \tag{1}$$

where w_i is a weight chosen for each control unit. The weights are chosen by the following reweighting scheme:

$$\min_{w_i} H(w) = \sum_{\{i|D=0\}} h(w_i) \tag{2}$$

subject to balance and normalizing constraints

$$\sum_{\{i|D=0\}} w_i \, c_{ri}(X_i) = m_r \quad \text{with} \quad r \in 1, \ldots, R \quad \text{and} \tag{3}$$

$$\sum_{\{i|D=0\}} w_i = 1 \quad \text{and} \tag{4}$$

$$w_i \geqslant 0 \quad \text{for all} \quad i \quad \text{such that} \quad D = 0, \tag{5}$$

where $h(\cdot)$ is a distance metric and $c_{ri}(X_i) = m_r$ describes a set of R balance constraints imposed on the covariate moments of the reweighted control group as discussed below.

[8] Ellipsoidal symmetry fails if X includes binary, categorical, and or skewed continuous variables.

The reweighting scheme consists of three features. First, the loss function $h(\cdot)$ is a distance metric chosen from the general class of empirical minimum discrepancy estimators defined by the Cressie–Read (CR) divergence (Read and Cressie 1988). We prefer to use the directed Kullback (1959) entropy divergence defined by $h(w_i) = w_i\, log(w_i/q_i)$ with estimated weight w_i and base weight q_i.[9] The loss function measures the distance between the distribution of estimated control weights defined by the vector $W = [w_i, \ldots, w_{n_0}]'$ and the distribution of the base weights specified by the vector $Q = [q_i, \ldots, q_{n_0}]'$ with $q_i \geqslant 0$ for all i such that $D = 0$ and $\sum_{\{i|D=0\}} q_i = 1$. Notice that the loss function is nonnegative and decreases the closer W is to Q; the loss equals zero if $W = Q$. We usually use the set of uniform weights with $q_i = 1/n_0$ as our base weights.

The second feature of the scheme involves the balance constraints defined in equation (3). They are imposed by the researcher to equalize the moments of the covariate distributions between the treatment and the reweighted control group (we assume that the relevant moments exist). A typical balance constraint is formulated with m_r containing the rth order moment of a given variable X_j from the target population (i.e., the treatment group), whereas the moment functions are specified for the source population (i.e., the control group) as $c_{ri}(X_{ij}) = X_{ij}^r$ or $c_{ri}(X_{ij}) = (X_{ij} - \mu_j)^r$ with mean μ_j.

The third feature are the two normalization constraints in equations (4–5). The first condition implies that the weights sum to the normalization constant of one. This choice is arbitrary and other constants can be used by the researcher.[10] The second condition implies a nonnegativity constraint because the distance metric is not defined for negative weight values. Below we see that this constraint is not binding and can be safely ignored.

The entropy balancing scheme can be understood as a generalization of the conventional propensity score weighting approach where the researcher first estimates the unit weights with a logistic regression and then computes balance checks to see if the estimated weights indeed equalize the covariate distributions. Entropy balancing tackles the adjustment problem from the reverse and estimates the weights directly from the imposed balance constraints. Instead of hoping that an accurately estimated logistic score will balance the covariates stochastically, the researcher directly exploits her knowledge about the sample moments and starts by prespecifying a potentially large set of balance constraints that imply that the sample moments in the reweighted control group exactly match the corresponding moments in the treatment group. The entropy balancing scheme then searches for a set of weights that are adjusted far enough to satisfy the balance constraints, but at the same time kept as close as possible (in an entropy sense) to the set of uniform base weights in order to retain efficiency for the subsequent analysis. This procedure has the key advantage that it directly adjusts the unit weights to the known sample moments such that exact moment matching is obtained in finite samples. Balance checking in the conventional sense is therefore no longer necessary, at least for the moments included in the balance constraints.

In the case of a large randomized experiment where the distributions are (asymptotically) balanced before the reweighting, the specified balance constraints in equation (3) are nonbinding (assuming no chance imbalances) and the counterfactual mean is simply estimated as a weighted average of the outcomes with every control unit weighted equally. The higher the level of imbalance in the covariate distributions, the further the weights have to be adjusted to meet the balance constraints. The number of moment conditions may vary depending on the dimensionality of the covariate space, the shapes of the covariate densities in the two groups, the sample sizes, and the desired balance level. At a minimum, the researcher would want to adjust at least the first moments of the marginal distributions of all confounders in X, but variances can be similarly adjusted (see the empirical examples below). In many empirical cases,

[9] The CR divergence family is described by $h(w) \equiv CR(\gamma) = \frac{w^{\gamma+1}-1}{\gamma(\gamma+1)}$, where γ indexes the family and limits are defined by continuity so that $\lim_{\gamma\to 0} CR(\gamma) = \lim_{\gamma\to 0} \frac{w^{\gamma+1}-1}{\gamma} = \lim_{\gamma\to 0} w\, log(w)$ and $\lim_{\gamma\to -1} CR(\gamma) = \lim_{\gamma\to -1} \frac{w^{\gamma+1}-1}{\gamma} = \lim_{\gamma\to -1} -log(w_i)$ where the last equalities follow from l'Hospital respectively. Notice that $h(w) = w\, log(w)$ represents the Shannon entropy metric which is (up to a constant) equivalent to the Kullback entropy divergence when uniform weights q_i are used for the null distribution. Another choice with good properties is $\gamma = -1$ which results in an empirical likelihood (EL) scheme. We prefer the entropy loss because it is more robust under misspecification (Imbens, Spady, and Johnson 1998; Schennach 2007) and constrains the weights to be non-negative.

[10] For example, the sum of the control weights could be normalized to equal the number of treated units; that is identical to setting the normalization constraint to n_1.

we would expect the bulk of the confounding to depend on the first and second moments. If, however, the researcher is concerned about dependencies in higher moments, these can be similarly adjusted by including higher moments in the condition vector. Interactions can be similarly included. The number of moment constraints can be increased at a constant rate with a growing sample size.

Notice that this reweighting scheme is analogous to reweighting adjustments that are sometimes used in the survey literature to correct sampling weights for bias due to nonresponse, frame undercoverage, response biases, or integrate auxiliary information to improve precision of estimates. The idea is that by introducing auxiliary information about known characteristics of the target population (e.g., population totals known from the census), one can improve estimates about unknown characteristics of the target population by adjusting the sampling design weights so that the sample moments match (at least) the known population moments. These adjustments include a wide variety of methods such as poststratification, raking, and calibration estimators (see, e.g., Deming and Stephan 1940, Oh and Scheuren 1978, or Särndal and Lundström 2006 for a recent review). Zaslavsky (1988) proposes a similar log-linear reweighting scheme with an entropy divergence to adjust for undercount in census data. Ireland and Kullback (1968) develop a minimum discrimination estimator that fits the cell probabilities of a (multidimensional) contingency table based on fixed marginal probabilities by minimizing the directed entropy divergence (starting from equal weights). They show that minimizing the entropy from uniform base weights provides an estimator that is consistent as well as asymptotically normal and efficient.

In contrast to most applications of reweighting in a survey context, where the vector of auxiliary information is commonly limited to a few known totals, in the case of entropy balancing, the data from the treatment group allows us to create a very large set of moment conditions. This can force the density of X in the reweighted control group to look very close to that in the treatment group. Moreover, by including balance constraints for the moments of all confounders, the researcher can rule out the possibility that balance decreases on any of the specified moments. This is an important advantage over conventional propensity score weighting where the weights are not directly adjusted to the know sample moments.

3.2 *Implementation*

To fit the entropy balancing weights, we need to minimize the loss function $H(w)$ subject to the balance and normalization constraints given in equations (3–5). Using the Lagrange multiplier, we obtain the primal optimization problem:

$$\min_{W,\lambda_0,Z} L^p = \sum_{\{i|D=0\}} w_i \log(w_i/q_i) + \sum_{r=1}^{R} \lambda_r \left(\sum_{\{i|D=0\}} w_i c_{ri}(X_i) - m_r \right) + (\lambda_0 - 1) \left(\sum_{\{i|D=0\}} w_i - 1 \right), \tag{6}$$

where $Z = \{\lambda_1, \ldots, \lambda_R\}'$ is a vector or Lagrange multipliers for the balance constraints and $\lambda_0 - 1$, the Lagrange multiplier for the normalization constraints. This system of equations is computationally inconvenient given its dimensionality of $n_0 + R + 1$. However, we can exploit several structural features that make this problem very susceptible to solution. First, the loss function is (strictly) convex since $\frac{\partial^2 h}{\partial W^2} > 0$ for $w_i \geqslant 0$, so that every local solution W^* is a global solution and any global solution is unique if the constraints are consistent. Second, as was recognized by Erlander (1977), duality holds and we can substitute out the constraints.[11] The first order condition of $\frac{\partial L^p}{\partial w_i} = 0$ yields that the solution for each weight is attained by

$$w_i^* = \frac{q_i \exp(-\sum_{r=1}^{R} \lambda_r c_{ri}(X_i))}{\sum_{\{i|D=0\}} q_i \exp\left(-\sum_{r=1}^{R} \lambda_r c_{ri}(X_i)\right)}. \tag{7}$$

[11] Also see Kapur and Kevsavan (1992) or Mattos and Veiga (2004) for detailed treatments and similar algorithms for entropy optimization.

The expression makes clear that the weights are estimated as a log-linear function of the covariates specified in the moment conditions.[12] Plugging this expression back into L^p eliminates the constraints and leads to an unrestricted dual problem given by

$$\min_Z L^d = \log\left(\sum_{\{i|D=0\}} q_i \exp\left(-\sum_{r=1}^{R} \lambda_r\, c_{ri}(X_i)\right)\right) + \sum_{r=1}^{R} \lambda_r\, m_r. \tag{8}$$

The solution to the dual problem Z^* solves the primal problem and the weights W^* can be recovered via equation (7). This dual problem is much more tractable because it is unconstrained and dimensionality is reduced to a system of nonlinear equations in the R Lagrange multipliers. Moreover, if a solution exists, it will be unique since L^d is strictly convex.

We use a Levenberg–Marquardt scheme to find Z^* for this dual problem. We rewrite the constraints in matrix form by defining the $(R \times n_0)$ constraint matrix $C = [c_1(X_i), \ldots, c_R(X_i)]'$ and the moment vector $M = [m_1, \ldots, m_R]'$. The balance constraints are given by $CW = M$, where C' must be full column rank, otherwise the constraints are not linearly independent and the system has no feasible solution. The rewritten problem is

$$\min_Z L^d = \log(Q'\exp(-C'Z)) + M'Z \text{ with solution } W^* = \frac{Q \cdot \exp(-C'Z))}{Q'\exp(-C'Z)}. \tag{9}$$

The gradient and Hessian are $\frac{\partial L^d}{\partial Z} = M - CW$ and $\frac{\partial^2 L^d}{\partial Z^2} = C[D(W) - WW']C'$, where $D(W)$ is a n_0-dimensional diagonal matrix with W in the diagonal. We exploit this second-order information by iterating

$$Z^{new} = Z^{old} - l\, \nabla_Z^2 L^{d\,-1} \nabla_Z L^d, \tag{10}$$

where l is a scalar that denotes the step length. In each iteration, we either take the full Newton step or otherwise l is chosen by backtracking in the Newton direction to the optimal step length using line search that combines a golden section search and successive quadratic approximation. $Z^0 = (CC')^{-1}M$ provides a starting guess. This iterative algorithm is globally convergent if the problem is feasible, and the solution is usually obtained within seconds even in moderately large data sets.

3.3 *Alternative Base Weights*

Instead of minimizing the distance from uniform weights $q_i = 1/n_0$, the entropy balancing adjustment may be started from alternative base weights. In the survey context, the base weights usually come from the sampling design and the goal is to adjust the sample to some known features of the target population while moving the design weights as little as possible (Oh and Scheuren 1978; Zaslavsky 1988; Särndal and Lundström 2006). In our context, a base weight can be similarly drawn from preexisting sampling weights or weights that are constructed from a balancing score that is initially estimated with a logistic regression of the treatment indicator on the covariates. These base weights can provide a first step toward balancing the covariates, but for various reasons discussed above imbalances may remain on several covariates. Entropy balancing can then "overhaul" the weights to fix these remaining imbalances for the specified moments.

3.4 *Estimation in the Preprocessed Data*

As indicated above, the entropy balancing weights can be easily combined with almost any standard estimator that the researcher may want to use to model the outcome in the preprocessed data. In particular, the entropy balancing weights are easily passed to regression models that may further address the correlation between the outcome and covariates in the reweighted data and also provide variance estimates for the treatment effects (which treat the weights as fixed). Such regression models may include covariates or

[12] Evidently, the inequality bounds $w_i \geqslant 0$ are inactive and can be safely ignored.

interactions that are not directly included in the reweighting to remove bias that may arise from remaining differences between the treatment and the reweighted control group. The outcome model may also increase precision if the (additional) variables in the outcome model account for residual variation in the outcome of interest (Robins, Rotnitzky, and Zhao 1995; Hirano and Imbens 2001). Notice that because the entropy balancing weights orthogonalize the treatment variable with respect to the covariates that are included in the reweighting, adding these covariates to the outcome regression has no effect on the point estimate of the treatment indicator (see the empirical applications below).

3.5 *Entropy Balancing and Other Preprocessing Methods*

As described above, entropy balancing may be seen as a generalization of conventional propensity score weighting approach where the unit weights are directly estimated from the balance constraints. Among other commonly used preprocessing methods, entropy balancing shares a similarity with genetic matching as described in Diamond and Sekhon (2006) insofar as it directly focuses on covariate balance. However, it differs from genetic matching in several important aspects. Genetic matching finds nearest neighbors based on a generalized distance metric that assigns weights to each covariate included in the matching. These covariate weights are chosen by a genetic algorithm in order to find a matching that maximizes covariate balance as measured by the minimum p value across a set of balance tests. In contrast, entropy balancing directly searches for a set of unit weights that balances the covariate distributions with respect to the specified moments. This obviates the need for balance checking altogether, at least with respect to the moments included in the balance constraints. Moreover, by freeing the weights to vary smoothly across units, entropy balancing also gains efficiency as it dispenses with the weight constraints that require that a unit is either matched or discarded. Entropy balancing is also computationally less demanding. The optimization problem in genetic matching is usually very difficult and irregular.

Entropy balancing is also related to coarsened exact matching (CEM) as recently proposed in Iacus, King, and Porro (2009) insofar as covariate balance is specified before the preprocessing adjustment, but entropy balancing also differs from CEM in important ways. CEM involves coarsening the covariates in order to match units exactly on the coarsened scale; treated and control units that cannot be matched exactly are discarded. Since exact matching is difficult in high-dimensional data, CEM often involves dropping some treated units (depending on the coarsening) and thereby changes the estimand from the PATT or SATT to a more local treatment effect for the remaining treated units (see Iacus, King, and Porro [2009] for reasons about why this can be beneficial). This differs from entropy balancing and other preprocessing approaches like genetic matching that traditionally do not involve the discarding of treated units in order to leave the estimand unchanged.[13] In principle, entropy balancing can be easily combined with other matching methods. For example, the researcher could first run CEM to trim the data and then apply entropy balancing to the remaining units. This may be useful when the researcher is not concerned about changing the estimand, perhaps because there are a small number of very unusual treated units that may be discarded to gain overlap. We leave it for further research to more closely investigate such a combined approach.

3.6 *Potential Limitations*

For any method, it is important to understand its potential limitations. There are at least three particular instances when entropy balancing may run into problems. First, no weighting solution exists if the balance constraints are inconsistent. For example, the researcher cannot specify a constraint which implies that the control group has a higher fraction of both males and females. This is easily avoided.

A second and more important issue can arise when the balance constraints are consistent, but there exists no set of positive weights to actually satisfy the constraints. This may occur if a user with limited data specifies extreme balance constraints that are very far from the control group data (e.g., imagine a treatment group with only 1% males and a control group with 99% males). This challenge of finding good matches with limited overlap is shared by all matching methods of course. The user has to be realistic about how much balance she asks for given the available data and overlap therein. If there simply are not

[13]There are exceptions to this rule (e.g., when calipers are used).

enough controls that look anything like the treated units, then the existing data do not contain sufficient information to reliably infer the counterfactual of interest.

Third, there may be a scenario where a solution exists, but due to limited overlap, the solution involves an extreme adjustment to the weights of some control units. In particular, if there are only very few "good" control units that are similar to the treated units then these controls may receive large weights because they contribute most information about the counterfactual of interest. Large weights increase the variance for the subsequent analysis and the user may also be uncomfortable with relying too heavily on a small number of highly weighted controls. A similar problem is shared by many preprocessing methods when matching with replacement reuses the good controls several times. In these cases, a weight refinement may be used to trim weights that are considered too large (see below). The researcher should also apply commonly used model diagnostics to check if the results for the subsequent analysis are possibly sensitive to some extreme weights. With limited overlap, the results will necessarily be more model dependent.

In general, the severity of these issues depends on the specific application (size of the data set, dimensionality, and degree of overlap). Below, we provide extensive simulations and several empirical applications that suggest that the method performs well in scenarios that may be typical of problems that are commonly encountered in political science.

3.7 *Weight Refinements*

Once a weighting solution is obtained that satisfies the balance constraints, the weights may be further refined by trimming large weights to lower the variance of the weights and thus the variance for the subsequent analysis. The weight refinement is easily implemented by iteratively calling the search algorithm described above. In each iteration, the set of solution weights w^* from the previous call are trimmed from above and or below at user specified thresholds and passed as the vector of starting weights q for the subsequent call. This augmented search is iterated until the weights meet the weight thresholds. Alteratively, the refinement can be fully automated by iterating until the variance of the weights can be no further reduced while still satisfying the balance constraints.

4 Monte Carlo Simulations

In this section, we conduct Monte Carlo experiments in order to evaluate the performance of entropy balancing in a variety of commonly used benchmark settings.[14] We compare the following commonly used matching and weighting procedures: difference in means (Raw), propensity score matching (PSM), Mahalanobis distance matching (MD), genetic matching (GM), combined propensity score and Mahalanobis distance matching (PSMD), propensity score weighting (PSW), and entropy balancing as described above. All matching is one-to-one matching with replacement. For the propensity score adjustments, the score is estimated with a logit or probit regression (following common practice in applied work). The web appendix provides a detailed description of the different preprocessing methods. In all cases, the counterfactual mean is computed as the average outcome of the control units in the preprocessed (matched or reweighted) data.

4.1 *Design*

We conduct three different simulations overall. The first two simulations follow the designs presented in Diamond and Sekhon (2006) and are described in detail in the web appendix. The first experiment

[14]There is a growing literature that uses simulation to assess the properties of matching procedures (partially reviewed in Imbens 2004). Frölich (2004) presents an extensive simulation study that considers various matching methods across a wide variety of sample designs, but his study is limited to a single covariate and true propensity scores. Zhao (2004) investigates the finite sample properties of pair matching and propensity score matching and finds no clear winner among these techniques. Although including different sample sizes, his study does not vary the controls to treated ratio and is also limited to true propensity scores. Brookhart et al. (2006) simulate the effect of including or excluding irrelevant variables in propensity score matching. Abadie and Imbens (2007) present a matching simulation using data from the Panel Study of Income Dynamics data and find that their bias corrected matching estimator outperforms linear regression adjustment. Diamond and Sekhon (2006) provide two Monte Carlo experiments, one with multivariate normal data and three covariates and a second using data from the Lalonde data set. They find that their genetic matching outperforms other matching techniques. Further simulations using multivariate normal data are presented in Gu and Rosenbaum (1993) and several of the papers collected in Rubin (2006). Drake (1993) finds that misspecified propensity scores often result in substantial bias in simulations with two normally distributed covariates.

involves three covariates and is based on conditions that are necessary for matching to achieve the EPBR property. We consider three cases: equal variances, unequal variances, and one scenario where we adjust for irrelevant covariates. We find that entropy balancing achieves the lowest root MSE compared to the other methods across all three cases (see Table I in the online appendix). The second experiment is based on the LaLonde (1986) data where the covariates are not ellipsoidally distributed and thus the EPBR conditions do not hold. Again, entropy balancing achieves the lowest MSE across all methods which suggests that the procedure retains fairly good finite sample properties even in this scenario where the EPBR conditions do not hold (see Table II in the online appendix).

Here we focus on the third, most comprehensive simulation. It follows the design developed in Frölich (2007) who to our knowledge provides the most extensive investigation of the finite sample properties of propensity score adjustments to date. We extend his design and consider a mixture of continuous and binary variables and we also examine additional factors such as the ratio of treated to controls and the degree of misspecification for the propensity score model. The idea is to mirror a range of typical scenarios that may be encountered in empirical settings in political science. We use six covariates X_j with $j \in (1, 2, \ldots, 6)$: X_1, X_2, and X_3 are multivariate normal with means zero, variances of $(2, 1, 1)$ and covariances of $(1, -1, -0.5)$ respectively; X_4 is distributed uniform on $[-3, 3]$; X_5 is distributed χ^2_1; X_6 is Bernoulli with mean 0.5. The treatment and control group are formed using

$$D = \mathbf{1}[X_1 + 2X_2 - 2X_3 - X_4 - 0.5X_5 + X_6 + \epsilon > 0].$$

Notice that the covariates are weighted unequally as is reasonable in many empirical settings. We consider three designs for the error term ϵ, which relate to different distributions for the true propensity score: Sample Design 1: $\epsilon \sim N(0, 30)$; Sample Design 2: $\epsilon \sim N(0, 100)$; Sample Design 3: $\epsilon \sim \chi^2_5$ and scaled to mean 0.5 and variance 67.6. Figure 1 visualizes the densities of the true propensity score in the three designs. The first design shows the strongest separation between the treatment and control group and provides a fairly difficult case for preprocessing. The second design has weaker separation so that the adjustments are expected to be more precise. The third design provides a middle ground as the variance lies between the first and the second design. However, the error term is leptokurtic such that the probit estimator for the estimated propensity score is misspecified.

We consider three sample sizes $n \in (300, 600, 1500)$ and also vary the ratio of control to treated units $r = n_0/n_1$ with $r \in (1, 2, 5)$ by sampling the specified numbers of treated and control units. For the estimators that rely on the estimated propensity score, we use three different probit specifications with the following mean functions:

- PS Design 1: $\widehat{p(x)} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6$
- PS Design 2: $\widehat{p(x)} = \alpha_0 + \alpha_1 X_1^2 + \alpha_2 X_2^2 + \alpha_3 X_3 + \alpha_4 X_4^2 + \alpha_5 X_5^2 + \alpha_6 X_6$
- PS Design 3: $\widehat{p(x)} = \alpha_0 + \alpha_1 X_1 X_3 + \alpha_2 X_2^2 + \alpha_3 X_4 + \alpha_4 X_5 + \alpha_5 X_6$.

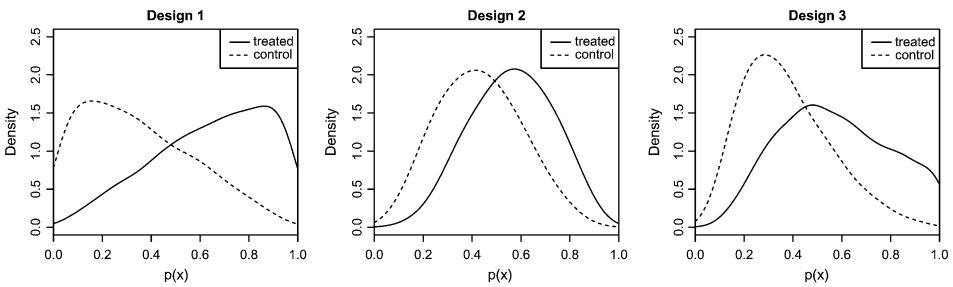

Fig. 1 Sample designs for Monte Carlo experiment: Density of true propensity score in treatment and control group. Left graph refers to Sample Design 1 with $\epsilon \sim N(0, 30)$ (strong separation and normal errors); middle graph refers to Sample Design 2 with $\epsilon \sim N(0, 100)$ (weaker separation and normal errors); right graph refers to Sample Design 3 with $\epsilon \sim \chi^2_5$ and scaled to mean 0.5 and variance 67.6 (medium separation and leptokurtic errors).

These functions are designed to yield various degrees of misspecification of the propensity score model. For normal ϵ (Sample Designs 1 and 2), the first model is correct, the second model is slightly misspecified, and the third model is heavily misspecified. The correlations between the true and the estimated propensity scores are 1, 0.8, and 0.3, respectively. For nonnormal ϵ (Sample Design 3), all three are misspecified, again with increasing levels of misspecification. Finally, we consider three outcome designs:

- Outcome Design 1: $Y = X_1 + X_2 + X_3 - X_4 + X_5 + X_6 + \eta$
- Outcome Design 2: $Y = X_1 + X_2 + 0.2\, X_3\, X_4 - \sqrt{X_5} + \eta$
- Outcome Design 3: $Y = (X_1 + X_2 + X_5)^2 + \eta$

with $\eta \sim N(0, 1)$. These regression functions are increasing in the degrees of nonlinearity in the mapping of the covariates to the outcome. The true treatment effect is fixed at zero for all units. The different outcomes also exhibit different correlations with the true propensity score decreasing from 0.8, 0.54, to 0.16 from sample design 1 to 3, respectively. We run 1000 simulations and report the bias and root MSE.

4.2 *Results*

The full results for $N = 300$ are presented in Table 1. To facilitate the interpretation, Figure 2 also presents a graphical summary of the sampling distributions for the case of the 1:5 treated to control ratio. Full results for $N = 600$ and $N = 1500$ are reported in the web appendix. The results are fairly similar across sample sizes.

Overall, the results suggest that entropy balancing outperforms the other adjustment techniques in terms of MSE. This result is robust for all three sample designs, the three outcome specifications, the three ratios of controls to treated, and the three propensity score equations. The gains in MSE are often substantial. For example, in the most difficult case of sample design 1 (strong separation), $N = 300$, and the highly nonlinear outcome design 3, the MSE from entropy balancing is about 2.6 times lower than that of genetic matching, 3.4 times lower than pair matching on a propensity score that is estimated with the correctly specified probit regression, 3.9 times lower than Mahalanobis distance matching, and 4.6 times lower than weighting on the estimated propensity score. As expected, we find that weighting or matching on misspecified propensity scores (PS designs 2 and 3) results in much higher MSE even in large samples.

Entropy balancing also outperforms the other matching techniques in terms of bias, except in larger samples where matching and weighting on the propensity scores from the correctly specified probit models yield equally good bias performance as one would expect given that stochastic balancing of the covariates improves. Yet, in these cases, entropy balancing retains lower MSE even at a sample size of $N = 1500$. This demonstrates the efficiency gains in finite samples that can be derived from adjusting the weights directly to the known sample moments.

5 Empirical Applications

In this section, we illustrate the use of entropy balancing in two real data settings. The first illustration reanalyzes data from a randomized evaluation of a large scale job training program. The second illustration applies the methods to a typical political science data set provided by Ladd and Lenz (2009) who study the effect of newspaper endorsements on vote choice in the 1997 British general election. Additional illustrations are provided in the web appendix.

5.1 *The LaLonde Data*

As a validation exercise, we first apply entropy balancing to the LaLonde (1986) data set, a canonical benchmark in the causal inference literature (see Diamond and Sekhon (2006) for the extensive debate surrounding this data set).[15] The LaLonde data consist of two parts. The first data set comes from a randomized evaluation of a large scale job training program, the National Supported Work Demonstration (NSW). This experimental data provide a benchmark estimate for the effect of the program. Using

[15]Notice that we focus on the Dehejia and Wahba subset of the LaLonde data.

Table 1 Results for Monte Carlo experiment ($N = 300$)

Sample Design 1: Strong separation and normal errors													
MSE	RAW	MD	GM	PSM1	PSM2	PSM3	PSMD1	PSMD2	PSMD3	PSW1	PSW2	PSW3	EB
Ratio CtoT 1 Y1	332	35	29	27	384	193	68	385	200	25	370	188	4
Ratio CtoT 1 Y2	502	24	15	14	163	532	71	193	442	26	161	537	5
Ratio CtoT 1 Y3	1196	1355	898	1186	1676	1835	2266	1995	2377	1590	1486	1633	346
Ratio CtoT 3 Y1	326	29	21	23	369	191	56	370	196	17	362	184	4
Ratio CtoT 3 Y2	495	20	10	11	155	523	57	175	445	18	157	528	4
Ratio CtoT 3 Y3	1269	1197	723	1108	1586	1849	2054	1777	2316	1074	1356	1666	291
Ratio CtoT 5 Y1	341	31	19	24	400	210	53	383	210	14	385	195	4
Ratio CtoT 5 Y2	512	20	12	13	165	550	52	172	473	15	166	547	5
Ratio CtoT 5 Y3	1471	1154	693	1184	1622	2041	2028	1892	2460	942	1378	1723	325
BIAS	RAW	MD	GM	PSM1	PSM2	PSM3	PSMD1	PSMD2	PSMD3	PSW1	PSW2	PSW3	EB
Ratio CtoT 1 Y1	179	52	42	13	189	133	79	192	138	13	188	135	1
Ratio CtoT 1 Y2	222	43	29	16	122	227	81	134	208	15	122	230	3
Ratio CtoT 1 Y3	304	350	215	125	359	404	461	428	472	98	347	389	102
Ratio CtoT 3 Y1	177	47	33	7	184	132	70	187	136	12	186	133	-1
Ratio CtoT 3 Y2	221	38	22	11	118	225	71	127	208	16	121	227	1
Ratio CtoT 3 Y3	302	322	206	97	332	400	434	398	461	102	334	389	85
Ratio CtoT 5 Y1	179	45	27	9	189	135	65	187	138	14	191	134	0
Ratio CtoT 5 Y2	223	35	19	10	119	227	65	122	211	17	123	229	2
Ratio CtoT 5 Y3	301	302	182	75	315	400	416	396	460	90	325	384	55
Sample Design 2: Weaker separation and normal errors													
MSE	RAW	MD	GM	PSM1	PSM2	PSM3	PSMD1	PSMD2	PSMD3	PSW1	PSW2	PSW3	EB
Ratio CtoT 1 Y1	148	15	8	12	137	91	28	146	97	7	129	83	2
Ratio CtoT 1 Y2	229	8	4	6	59	231	21	58	192	8	54	230	2
Ratio CtoT 1 Y3	654	661	364	590	655	994	1156	1016	1383	573	482	825	196
Ratio CtoT 3 Y1	151	14	7	12	138	89	24	140	99	5	130	82	2
Ratio CtoT 3 Y2	225	7	4	7	60	225	18	52	196	6	54	225	3
Ratio CtoT 3 Y3	777	660	360	575	673	1063	1114	984	1421	428	517	902	208
Ratio CtoT 5 Y1	162	16	7	19	159	103	26	154	104	5	144	88	3
Ratio CtoT 5 Y2	236	9	6	10	66	248	20	56	214	6	58	238	3
Ratio CtoT 5 Y3	963	642	349	822	872	1233	1080	1013	1507	482	527	943	288
BIAS	RAW	MD	GM	PSM1	PSM2	PSM3	PSMD1	PSMD2	PSMD3	PSW1	PSW2	PSW3	EB
Ratio CtoT 1 Y1	117	32	18	4	107	88	48	114	94	1	108	87	1
Ratio CtoT 1 Y2	149	19	9	5	69	147	40	70	134	2	68	148	2
Ratio CtoT 1 Y3	194	238	149	31	180	281	321	297	353	4	177	266	27
Ratio CtoT 3 Y1	117	29	14	3	106	86	43	110	94	3	108	87	-1
Ratio CtoT 3 Y2	147	16	6	4	68	144	36	64	135	3	68	146	.0
Ratio CtoT 3 Y3	209	231	140	47	177	281	309	285	353	39	185	273	33
Ratio CtoT 5 Y1	118	27	10	3	109	87	41	111	92	2	111	87	-0
Ratio CtoT 5 Y2	149	16	5	5	67	147	33	62	136	3	68	148	0
Ratio CtoT 5 Y3	198	210	119	17	165	276	285	270	341	18	174	263	5
Sample Design 3: Medium separation and leptokurtic errors													
MSE	RAW	MD	GM	PSM1	PSM2	PSM3	PSMD1	PSMD2	PSMD3	PSW1	PSW2	PSW3	EB
Ratio CtoT 1 Y1	226	26	20	23	251	147	47	254	159	14	246	144	3
Ratio CtoT 1 Y2	350	18	11	11	116	404	46	131	337	18	122	416	3
Ratio CtoT 1 Y3	1213	1174	757	1069	1143	1534	1983	1680	2024	775	1098	1315	374
Ratio CtoT 3 Y1	221	23	16	20	253	153	41	250	156	12	246	142	3
Ratio CtoT 3 Y2	343	16	9	11	114	402	38	118	341	17	122	408	3
Ratio CtoT 3 Y3	1212	1046	645	1050	1011	1521	1769	1465	1934	739	951	1281	354
Ratio CtoT 5 Y1	239	25	15	24	287	161	38	268	169	12	273	150	3
Ratio CtoT 5 Y2	355	16	10	12	125	421	34	118	363	15	127	428	4
Ratio CtoT 5 Y3	1563	1122	701	1328	1253	1799	1828	1562	2124	888	1061	1465	440
BIAS	RAW	MD	GM	PSM1	PSM2	PSM3	PSMD1	PSMD2	PSMD3	PSW1	PSW2	PSW3	EB
Ratio CtoT 1 Y1	146	45	35	12	150	116	65	154	123	26	152	117	0
Ratio CtoT 1 Y2	185	36	23	16	101	197	64	110	180	35	107	201	4
Ratio CtoT 1 Y3	306	325	217	123	270	362	430	390	434	201	295	343	145
Ratio CtoT 3 Y1	144	41	29	9	150	117	59	152	120	25	152	116	1
Ratio CtoT 3 Y2	183	32	20	15	98	195	57	102	180	34	106	199	4
Ratio CtoT 3 Y3	295	300	202	91	239	354	400	358	419	186	272	336	130
Ratio CtoT 5 Y1	147	39	23	9	157	116	53	153	122	22	159	117	1
Ratio CtoT 5 Y2	185	29	16	14	101	197	50	97	182	29	107	201	4
Ratio CtoT 5 Y3	309	291	186	85	221	364	387	347	417	187	268	342	122

Note. MSE and BIAS across 1000 simulations. Experimental factors are three sample designs as in Fig. 1, three outcome designs (Y1 is linear, Y2 is somewhat nonlinear, Y3 is highly nonlinear), and three controls-to-treated ratios (Ratio CtoT 1, 3, and 5). Methods are Difference in means (RAW), Mahalanobis distance matching (MD), genetic matching (GM), entropy balancing (EB), matching or weighting on propensity score that is estimated with a probit regression (PSM and PSW), and Mahalanobis distance matching on the estimated propensity score and orthogonalized covariates (PSMD). For the propensity score adjustments the postfixes 1–3 indicate increasing degrees of misspecification for the propensity score estimation (see text for details). We use three specifications (labeled with a 1, 2, or 3 postfix) for all propensity-score-based methods (PSM, PSW, PSMD). The first propensity score model is correct for sample designs 1 and 2, and slightly misspecified for sample design 3. Propensity score models 2 and 3 are increasingly misspecified (as measured by the linear correlation between the true and the estimated score).

a simple difference in means, the program is estimated to increase postintervention earnings by $1794 with a 95% confidence interval of [551; 3038]. In the next step, we replace the experimental control group with a control group drawn from the Current Population Survey-Social Security Administration file (CPS-1) where we measure the same covariates as in the experimental data. The covariates include a set of measures that researchers would typically control for in observational studies on the impact of job training programs. Using this second data set, LaLonde found that many commonly used methods of covariate adjustment (such as regression) were not able to recover the results obtained from the randomized experiment.

Overall, there are 185 program participants from the experimental NSW evaluation (the treated units) and 15, 992 nonparticipants from the current population survey data (the control units). The outcome

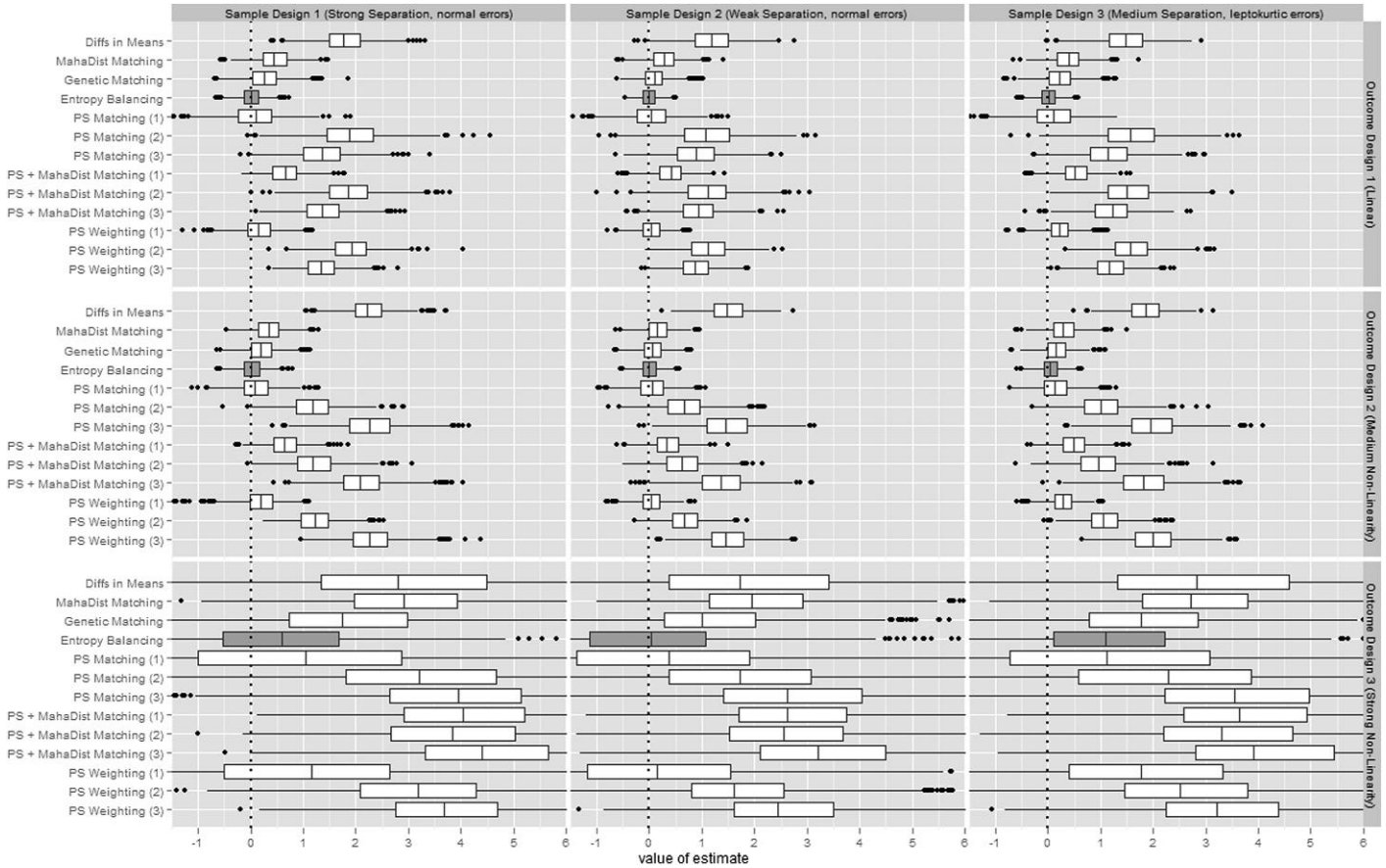

Fig. 2 Results of Monte Carlo experiment. Boxplots visualize the sampling distribution of 1000 Monte Carlo estimates from the various preprocessing methods for the three outcome designs and three sample designs (with a 1:5 treated to control ratio $N = 300$). The true treatment effect is zero (dashed vertical line). Methods are difference in means, Mahalanobis distance matching, genetic matching, entropy balancing, matching or weighting on propensity score (estimated with a probit regression), and Mahalanobis distance matching on the estimated propensity score and orthogonalized covariates. For the propensity score adjustments, the postfixes 1–3 indicate increasing degrees of misspecification for the propensity score estimation (see text for details).

of interest is postintervention earnings from the year 1978. The data contain 10 preintervention characteristics to control for the selection into the training program. These include earnings and employment status for two preintervention years (1974 and 1975), education (years of schooling and an indicator for completed high school degree), age, ethnicity (indicators for black and hispanic), and marital status. We conduct entropy balancing using the 10 raw variables, all their pairwise one-way interactions, as well as squared terms for the continuous variables age and years of education. Overall, this results in 52 covariate combinations.[16] We also apply Mahalanobis distance matching, genetic matching, propensity score matching, and propensity score weighting to the same data as a benchmark. The propensity score is estimated with a logistic regression of the treatment indicator on all 52 covariates. Notice that this biases the results in favor of the conventional propensity score adjustments because this extensive specification is considerably more flexible than models that are commonly used in applied work where researchers usually just include the raw covariates.[17]

For each of the 52 covariates, Figure 3 visualizes the covariate balance that we obtain from the different techniques as measured by two conventional balance statistics: the standardized difference in means between the treatment and control group (left panel) and the p value for a difference of means test (right panel). The open circles refer to the statistics for the unadjusted data. Not surprisingly, participants of the job training program differ in many respects from the general population so that almost all of these covariates are heavily imbalanced between the two groups. Standardized differences often exceed the |.1| threshold and almost all mean differences are significantly different from zero at conventional levels.[18] Due to this stark imbalance, the LaLonde data are generally regarded as a fairly difficult adjustment problem (the unadjusted difference in mean outcomes is far away from the experimental target at $-8506).

The black squares refer to the balance statistics obtained after the entropy balancing. As expected, balance is markedly improved such that the reweighted control group now has identical means compared to the treatment group on all covariates (the standardized means are zero and the p values are one). According to this metric, entropy balancing provides a much higher level of covariate balance than the other adjustment techniques including matching or weighting on the logistic propensity score that often leaves several covariates imbalanced (standardized differences often exceed |0.1|, and p values are low). Even worse, on a few variables the bias is actually increased after the logistic propensity score adjustment in the sense that the means are now further apart than in the unadjusted data.

Balance from these methods may be improved by tinkering with the propensity score specification. However, the current propensity score model is already fairly flexible. Moreover, given the skewed distributions of many covariates in this data, it seems very difficult to find a propensity score specification via trial and error that would jointly balance all covariates. This demonstrates the advantage of entropy balancing where balance is directly attained by construction of the moment conditions and never decreases for the moments that are included in the reweighting. Table V in the online appendix provides additional balance statistics which show that after entropy balancing the variance of the variables are also very similar on almost all covariates (note that variances are exactly adjusted for binary variables and all continuous variables for which squared terms are included). Figure I in the online appendix provides QQ plots which show that for the four continuous variables age, education, and the two years of pretreatment earnings, the distributions are fairly similar after the preprocessing.

Taken together, entropy balancing delivers a high degree of balance in this data set (according to standard metrics) despite the low computational cost (the weighting solution is obtained within seconds). The difference in means between the treatment group and the reweighted control group yields an average treatment effect on the treated of $1571 with a 95% confidence interval of [97, 3044], an estimate that is close to the experimental target and slightly more efficient than the final estimate of 1734 [−298; 3766] reported by Diamond and Sekhon (2006) for the best run from the genetic matching procedure (a linear regression in all covariates yields an effect estimate of $1159 [−52; 2371]).

[16]Notice that we exclude nonsensical interactions such as for example between high school degree and years of schooling. We also omit squared terms for pretreatment earnings and their interaction because due to their collinearity they are simply balanced by adjusting on the lower order terms. For example, their T-test p values in the reweighted data are .76, .83, .99, respectively.

[17]In the online appendix, we show another example where weighting on a logistic propensity score that is estimated without any. squared terms leads to a strong decrease in balance over the raw data for many covariates (see Figs. 5 and 6 in the online appendix).

[18]Notice that we use p values as a measure of balance, and not to conduct hypothesis tests in the conventional sense (see Imai, King, and Stuart 2008).

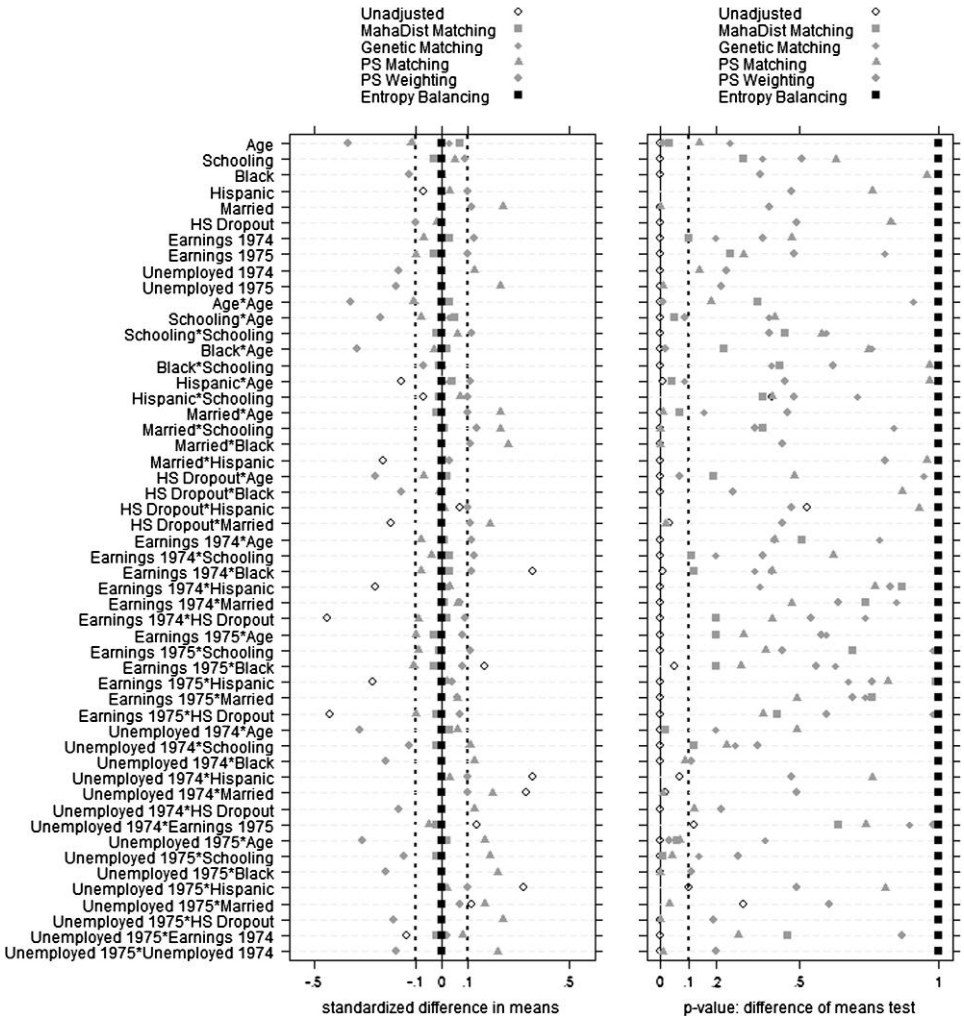

Fig. 3 Covariate balance in the LaLonde data. Left panel shows plot of covariate-by-covariate standardized bias in the unadjusted data and after the various preprocessing methods. The standardized bias measures the difference in means between the treatment and control group (scaled by the standard deviation). Zero bias indicates identical means, dots to the right (left) of zero indicate a higher mean among the treatment (control) group. The right panel shows the p value for a covariate-by-covariate t-test for the differences in means after the unadjusted data and after the various preprocessing methods.

In order to investigate the reduction in model dependency, we follow Ho et al. (2007) and examine the sensitivity of the effect estimates in both the unadjusted and the preprocessed data across a wide range of possible specifications of the outcome model. In particular, we fit one million regressions of the outcome on the treatment variable and a subset of covariates that we randomly draw from the set of all possible subsets of the 52 covariates.[19] We fit each regression twice, once with the unadjusted data (unweighted) and once with the preprocessed data (regressions are weighted by the entropy balancing weights). Figure 4 provides the densities of the estimates. The results are extremely model dependent in the unadjusted data with effect sizes ranging from $−8500 to over $4000. In the preprocessed data, however, all regressions yield the exact same estimate that is expected because the weights orthogonalize the treatment indicator with respect to all 52 covariate combinations that are included in the reweighting. This suggests that model dependency is reduced after entropy balancing.

[19]Notice that there are over 4.5 quadrillion possible subsets of the 52 covariates $\left(\sum_{i=1}^{52}\binom{52}{i}\right)$ so we cannot run all possible regressions.

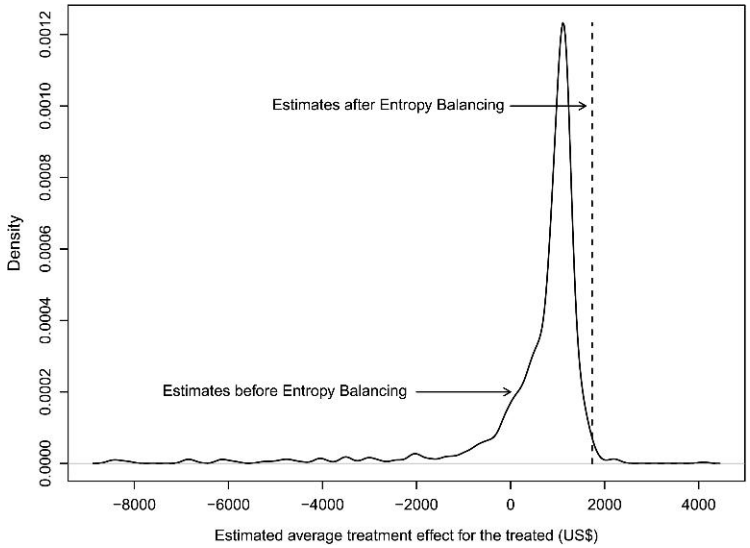

Fig. 4 Model dependency in the LaLonde data. Density of estimated treatment effects across one million randomly sampled model specifications in the unadjusted data (solid) and the data preprocessed with entropy balancing (dashed line).

5.2 *News Media Persuasion*

In this section, we apply entropy balancing to a typical political science survey data set by reanalyzing data from Ladd and Lenz (2009) who examine how shifts in the partisan endorsements of British newspapers affected major party vote choice in the 1997 general election.[20] The authors' identification strategy exploits the fact that on the second day of the official election campaign, the *Sun* (which had the largest circulation in Great Britain) and several other British newspapers ended their long-standing support for the ruling Conservative party and switched their endorsement to the Labour candidate Tony Blair. Ladd and Lenz (2009) draw upon data from several waves of the British Election Panel Study 1992–1997, where the same voters are being interviewed before the endorsement shifts (in 1992, 1994, 1995, and 1996) and once following the 1997 election. The main comparison involves 211 "treated" respondents who in 1996 (the last wave before the shifts in endorsements) report that they read one of the newspapers that eventually switched their endorsement to Labour prior to the 1997 election. These treated voters are compared to 1382 "control" respondents who either read papers whose partisan endorsements remained constant or who report that they did not read a paper. The outcome variable is vote choice in the 1997 election as reported in the postelection survey.

The authors control for a battery of pretreatment variables to account for the nonrandom selection into readership of switching newspapers. The control variables include various measures for a respondent's prior evaluation of the Labour Party (such as prior party support, prior labour vote, etc.), prior ideology, socioeconomic status, authoritarianism, gender, age, region, and occupation.[21] There are 39 covariates overall; most of them are binary or ordinal. In their analysis, the authors rescale all variables to vary from zero to one, match on a subset of eight of the most important covariates, and finally include the additional controls in a subsequent regression of the outcome on the treatment indicator and all control variables in the preprocessed data. We conduct entropy balancing by imposing moment conditions on the means of all covariates directly. Since most variables are binary, exactly adjusting the means also exactly adjusts the variances. We also apply the other adjustment methods to the same data; the propensity score is estimated with a logistic regression in all 39 covariates.

Figure 5 displays the balance results from the various preprocessing methods as measured by the standardized differences in means (left panel) and the p values of the difference in means tests (right panel).

[20] I am grateful to the authors for sharing their data.

[21] Notice that variables that are labeled as "prior" are measured in the 1992–1996 survey waves. See the authors' web appendix for a detailed explanation of the variable definitions.

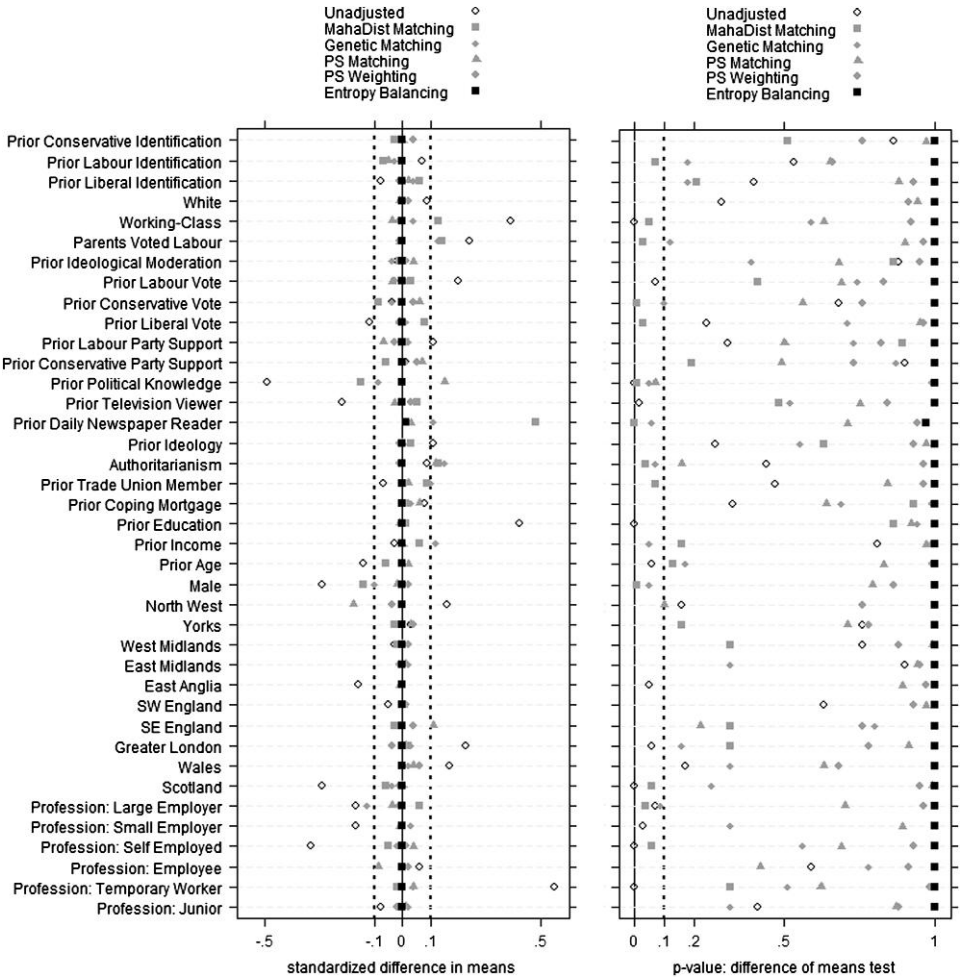

Fig. 5 Covariate balance in the news media persuasion data. Left panel shows plot of covariate-by-covariate standardized bias in the unadjusted data and after the various preprocessing methods. The standardized bias measures the difference in means between the treatment and control group (scaled by the standard deviation). Zero bias indicates identical means, dots to the right (left) of zero indicate a higher mean among the treatment (control) group. The right panel shows the p value for a covariate-by-covariate t-test for the differences in means after the unadjusted data and after the various preprocessing methods.

As indicated by the open circles, the unadjusted data are imbalanced on several important covariates. For example, readers of newspapers that switched their endorsements to Labour were more likely to be female, younger, prior Labour voters, members of the working class, in temporary employment, and less politically informed. The standardized differences exceed the |0.1| threshold for 19 covariates, and the p values are below conventional levels of significance in 16 cases. As indicated by the black squares, entropy balancing removes these differences effectively and exactly adjusts the means for all 39 covariates; the variances are also adjusted on almost all covariates (see Table VI in the online appendix that provides additional balance statistics). This exceeds the balance level reported by the authors in their balance table (Ladd and Lenz, 2009, 401) despite the fact that additional variables are included in the matching. Entropy balancing also improves these balance metrics compared to the other preprocessing methods some of which leave several covariates imbalanced or even decrease the balance in a few instances (especially Mahalanobis distance and propensity score matching).

The difference in mean outcomes in the reweighted data suggests that the endorsement switch increased the reported probability of voting for Labour by 0.12 with a 95% confidence interval of [0.20, 0.04]—a magnitude close to the authors' original estimate. Figure 6 examines the reduction in model dependence where we display the effect estimates across one million regression specifications with randomly drawn

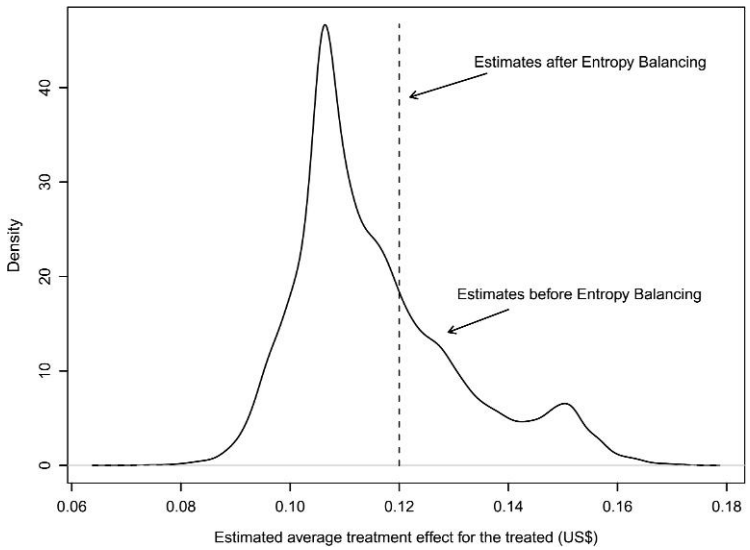

Fig. 6 Model dependency in the news media persuasion data. Density of estimated treatment effects across one million randomly sampled model specifications in the unadjusted data (solid) and the data preprocessed with entropy balancing (dashed line).

covariate subsets in the raw and the preprocessed data. The effect estimates remain stable in the preprocessed data since the entropy balancing weights orthogonalize the treatment indicator to all covariates. In the raw data, the effects vary from about 0.06 to 0.18 across the different regressions suggesting that model dependence is considerably lower after the preprocessing.

6 Conclusion

The goal of preprocessing is to generate well-balanced samples, but commonly used methods often make it difficult for applied researchers to achieve high balance targets. One reason for this is that many commonly used methods fail to focus on covariate balance directly, but instead rely on an intricate and often ineffective process of "manually" iterating between propensity score modeling, matching, and balance checking to search for a suitable balancing solution. In the worst case, these techniques may increase bias for the subsequent estimation of treatment effects when balance improvements in some covariates are accompanied with decreased balance for other important covariates.

We propose entropy balancing as a processing technique to create balanced samples. In entropy balancing, the researcher starts by imposing a potentially large set of balance conditions which imply that the treatment and reweighted control group match exactly on a possibly large set of the prespecified moments. Entropy balancing then directly adjusts the unit weights to the specified sample moments while moving the weights as little as possible to retain information. This makes it easier for the user to find unit weights that balance the moments between the treatment and control group and obviates the need for continual balance checking for the moments that are included in the reweighting.

The entropy balancing weights can be paired with standard estimators that the researcher may want to use to subsequently model the outcome in the preprocessed data. The balance improvements that result from entropy balancing can translate into lower approximation error and reduced model dependency in finite samples as demonstrated through the extensive Monte Carlo simulations and several empirical applications. Future research may consider the combination of entropy balancing and other preprocessing methods.

While entropy balancing simplifies the search for covariate balance for practitioners, it is important to notice that other problems that are commonly associated with preprocessing methods (and covariate adjustment more generally) still apply. For example, entropy balancing provides no safeguard against bias from unmeasured confounders that are often a vexing problem in observational studies.

References

Abadie, A., and G. Imbens. 2007. Simple and bias-corrected matching estimators for average treatment effects. Working Paper. Harvard University.

Brookhart, M., S. Schneeweiss, K. Rothman, R. Glynn, J. Avorn, and T. Sturmer. 2006. Variable selection for propensity score models. *American Journal of Epidemiology* 163:1149–56.

DellaVigna, S., and E. Kaplan. 2007. The Fox News effect: Media bias and voting. *Quarterly Journal of Economics* 122:1187–34.

Deming, W., and F. Stephan. 1940. On the least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics* 11:427–44.

Diamond, A. J., and J. Sekhon. 2006. Genetic matching for causal effects: A general multivariate matching method for achieving balance in observational studies. Unpublished manuscript, Department of Political Science, UC Berkeley.

Drake, C. 1993. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 49:1231–36.

Eggers, A., and J. Hainmueller. 2009. MPs for sale? Returns to office in postwar British politics. *American Political Science Review* 103:513–33.

Erlander, S. 1977. *Entropy in linear programs–an approach to planning*. Report No. LiTH-MAT-R-77-3. Department of Mathematics, Linköping University, Sweden.

———. 2004. Finite sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics* 86:77–90.

Frölich, M. 2007. Propensity score matching without conditional independence assumption with an application to the gender wage gap in the United Kingdom. *The Econometrics Journal* 10:359–407.

Graham, B. S., C. Pinto, and D. Egel. 2010. Inverse probability tilting for moment condition models with missing data. Working paper. New York University.

Gu, X., and P. Rosenbaum. 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 2:405–20.

Hahn, J. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66:315–31.

Hansen, B. B., and J. Bowers. 2008. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science* 23:219–36.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–54.

Hellerstein, J., and G. Imbens. 1999. Imposing moment restrictions from auxiliary data by weighting. *The Review of Economics and Statistics* 81:1–14.

Hirano, K., and G. Imbens. 2001. Estimation of causal effects using propensity score weighting: An application of data on right hear catherization. *Health Services and Outcomes Research Methodology* 2:259–78.

Hirano, K., G. Imbens, and G. Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71:1161–89.

Ho, D., K. Imai, G. King, and E. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15:199–236.

Horvitz, D., and D. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47:663–85.

Iacus, S., G. King, and G. Porro. 2009. *Causal inference without balance checking: Coarsened exact matching*. Mimeo Harvard University.

Imai, K., G. King, and E. Stuart. 2008. Misunderstandings among experimentalists and observationalists: Balance test fallacies in causal inference. *Journal of the Royal Statistical Society, Series A* 171:481–502.

Imbens, G. 1997. One-step estimators for over-identified generalized method of moments models. *The Review of Economic Studies* 64:359–83.

———. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86:4–29.

Imbens, G., R. Spady, and P. Johnson. 1998. Information theoretic approaches to inference in moment condition models. *Econometrica* 66:333–57.

Ireland, C., and S. Kullback. 1968. Contingency tables with given marginals. *Biometrika* 55:179–88.

Kapur, J., and H. Kevsavan. 1992. *Entropy optimization principles with applications*. London: Academic Press.

Kitamura, Y., and M. Stutzer. 1997. An information-theoretic alternative to generalized method of moments estimation. *Econometrica* 65:861–74.

Kullback, S. 1959. *Information theory and statistics*. New York: Wiley.

Ladd, J., and G. Lenz. 2009. Exploiting a rare communication shift to document the persuasive power of the news media. *American Journal of Political Science* 53:394–10.

LaLonde, R. J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76:604–20.

Mattos, R., and A. Veiga. 2004. Entropy optimization: Computer implementation of the maxent and minexent principles. Working Paper. Universidade Federal de Juiz de Fora, Brazil.

McCaffrey, D., G. Ridgeway, and A. Morral. 2004. Propensity score estimation with boosted regression for evaluating adolescent substance abuse treatment. *Psychological Methods* 9:403–25.

Oh, H. L., and F. J. Scheuren. 1978. Multivariate ratio raking estimation in the 1973 exact match study. *Proceedings of the Section on Survey Research Methods* XXV:716–22.

Owen, A. 2001. *Empirical likelihood*. Boca Raton, FL: Chapman & Hall.

Qin, J., and J. Lawless. 1994. Empirical likelihood and general estimating equations. *Annals of Statistics* 22:300–25.
Qin, J., B. Zhang, and D. Leung. 2009. Empirical likelihood in missing data problems. *Journal of the American Statistical Association* 104:1492–503.
Read, T., and N. Cressie. 1988. *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer.
Robins, J., A. Rotnitzky, and L. Zhao. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 90:106–21.
Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.
Rubin, D. 2006. *Matched sampling for causal effects*. Cambridge: Cambridge University Press.
Särndal, C. E., and S. Lundström. 2006. *Estimation in surveys with nonresponse*. New York: John Wiley & Sons, Ltd.
Schennach, S. 2007. Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics* 35:634–72.
Sekhon, J. 2006. Alternative balance metrics for bias reduction in matching methods for causal inference. Unpublished manuscript, Department of of Political Science, UC Berkeley.
Sekhon, J. S. 2009. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science* 12:487-08.
Smith, J., and P. Todd. 2001. Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review* 91:112–18.
Zaslavsky, A. 1988. Representing local reweighting area adjustments by of households. *Survey Methodology* 14:265–88.
Zhao, Z. 2004. Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics* 86:91–107.