# Globally Efficient Nonparametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting

K. C. G. Chan

*Department of Biostatistics, University of Washington*

S. C. P. Yam and Z. Zhang

*Department of Statistics, Chinese University of Hong Kong*

**Summary**. The estimation of average treatment effects based on observational data is extremely important in practice and has been studied by generations of statisticians under different frameworks. Existing globally efficient estimators require non-parametric estimation of a propensity score function, an outcome regression function or both, but their performance can be poor in practical sample sizes. Without explicitly estimating either functions, we consider a wide class calibration weights constructed to attain an exact three-way balance of the moments of observed covariates among the treated, the control, and the combined group. The wide class includes exponential tilting, empirical likelihood and generalized regression as important special cases, and extends survey calibration estimators to different statistical problems and with important distinctions. Global semiparametric efficiency for the estimation of average treatment effects is established for this general class of calibration estimators. The results show that efficiency can be achieved by solely balancing the covariate distributions without resorting to direct estimation of propensity score or outcome regression function. We also propose a consistent estimator for the efficient asymptotic variance, which does not involve additional functional estimation of either the propensity score or the outcome regression functions. The proposed variance estimator outperforms existing estimators that require a direct approximation of the efficient influence function.

*Keywords*: Global semiparametric efficiency; Propensity score; Sieve estimator; Treatment effects.

## 1. Introduction

Studying the effect of an intervention or a treatment is central to experimental scientists. While a randomized trial is a gold standard to identify average treatment effects, it may be infeasible, or even unethical, to conduct in practice. Observational studies are common in econometrics, social science, and public health, where the participation of intervention is only observed rather than manipulated by scientists. A typical concern for inferring causality in an observational study is confounding, where individual characteristics such as demographic factors can be related to both the treatment selection and the outcome of interest. In these cases, a simple comparison of sample averages from the two intervention groups can lead to a seriously biased estimate of the population average treatment effects.

When the treatment selection process depends entirely on observable covariates, there are two broad classes of strategies for estimating average treatment effects, namely outcome regression and propensity score estimation. When a linear model is assumed for the outcome given covariates, the coefficient for treatment status provides an estimate of average treatment effects when all relevant confounders are controlled and when there is no effect modifiers. In general, more complex regression models can be used for predicting unobservable potential outcomes, while the average treatment effects can be estimated by averaging predicted outcomes (Oaxaca; 1973; Blinder; 1973). An alternative class of strategies is based on the propensity score, which is the probability of receiving treatment given covariates. Rosenbaum and Rubin (1983) showed that adjusting the true propensity score can remove all bias due to confounding. They also showed that the true propensity score balances the covariate distributions between the two treatment arms. Propensity score can be used for subclassification (Rosenbaum and Rubin; 1984; Rosenbaum; 1991), matching (Rosenbaum and Rubin; 1985; Abadie and Imbens; 2006), and weighting (Rosenbaum; 1987; Hirano et al.; 2003). However, propensity score based methods may not be efficient in general.

To study efficient estimation, semiparametric efficiency bounds were derived independently by Robins et al. (1994) and Hahn (1998). Interestingly, the efficient influence function for the average treatment effects involves both the propensity score and the outcome regression functions. This motivated subsequent development of methods involving a combination of propensity score and outcome regression modeling (Robins et al.; 1994; Hahn; 1998; Bang and Robins; 2005; Qin and Zhang; 2007; Cao et al.; 2009; Tan; 2010; Graham et al.; 2012; Vansteelandt et al.; 2012). Recently, Chan (2013), Han and Wang (2013), Chan and Yam (2014) and Han (2014) considered methods that can accommodate multiple non-nested models of the propensity score and outcome regression at the same time. Many recent methods focus on improving covariate balance within the propensity score and outcome regression frameworks (Qin and Zhang; 2007; Tan; 2010; Chan; 2012; Graham et al.; 2012; Vansteelandt et al.; 2012; Han and Wang; 2013; Chan and Yam; 2014). Imai and Ratkovic (2014) argue that the estimation of propensity score parameters with a specification of outcome model does not align well with the original spirit of propensity score methodology as discussed in Rubin (2007). They proposed a covariate balancing propensity score method for the estimation of propensity score parameters, which balances covariates for an overidentified moment restriction without assuming an outcome model.

All methods mentioned so far require specification of either a propensity score model, an outcome model, or both. Consistency of the estimators requires some underlying models to be correctly specified. Since the estimand of interest is the average treatment effects, and the propensity score or the outcome models are just intermediate steps, it is natural to question whether the correctness of the intermediate models were necessary for producing correct inference. Nonparametric estimators are developed to provide valid inference in large samples without relying on parametric assumptions in the intermediate steps of estimation. Hahn (1998), Hirano et al. (2003), Imbens et al. (2006), Chen et al. (2008) have considered various nonparametric estimators for the average treatment effects. Although the validity of estimation does not rely on any parametric assumption on the propensity score and outcome models, their methods require sieve approximations of those unknown conditional functions, as these functions appear explicitly in the semiparametric efficient inference functions (Robins et al.; 1994; Hahn; 1998). An important observation has been made by Hirano et al. (2003) that the celebrated inverse probability weighted estimator of Horvitz and Thompson (1952) is globally semiparametric efficient when a sieve maximum likelihood propensity score estimator is used. Global semiparametric efficiency is more desirable than local semiparametric efficiency which requires the correct specification of parametric models. An implication of Hirano et al. (2003) is that global efficiency can be achieved by solely estimating the propensity score nonparametrically, without requiring to estimate the outcome model, which also appears in the efficient influence function. An efficient estimator is adapted from the Horvitz-Thompson estimator, a simple estimator used by decades of statisticians. Alternatively, Imbens et al. (2006) and Chen et al. (2008) showed that globally efficient estimators can be constructed from nonparametric estimators of the outcome model only. A combination of nonparametric estimators of the propensity score and outcome models can also produce globally efficient estimators (Imbens et al.; 2006; Chen et al.; 2008).

The existing globally efficient estimators do not require correct specification of propensity score or outcome regression models in large samples, but the need to specify a nonparametric approximation of either or both functions is still present. It has been shown that estimators for average treatment effects can have substantial bias when either functions are poorly estimated (Kang and Schafer; 2007; Ridgeway and McCaffrey; 2007). It is natural to question whether nonparametric estimation of these functions is even necessary, and whether they can be replaced by an alternative simple balancing criterion which is inherited from the unknown propensity score function. Although estimators that improve balance of covariate distribution are discussed in recent papers (Qin and Zhang; 2007; Tan; 2010; Graham et al.; 2012; Vansteelandt et al.; 2012; Han and Wang; 2013; Imai and Ratkovic; 2014; Chan and Yam; 2014), their methods require parametric modeling of the propensity score model or the outcome model. Nonparametric methods for improving covariate balance have been studied widely in the survey sampling literature (Deming and Stephan; 1940; Deville and Särndal; 1992; Kim and Park; 2010; Hainmueller; 2012). The recent paper of Hainmueller (2012) focused on using the implied weights of the raking estimator of Deming and Stephan (1940) to preprocess data for estimating the treatment effects on the treated. Since he focused on preprocessing, he did not study statistical inference and estimation efficiency.

The class of survey calibration estimators is simple to implement, conceptually appealing, and has been used by survey statisticians for decades for different applications than the estimation of average treatment effects. Survey calibration weights typically minimize the distance from indetermined weights to a set of pre-specified design weights, subject to moment conditions. For missing data applications, there is no known design weights, and calibration usually requires estimating inverse probability weights, see Chan and Yam (2014) for a recent review. The connections between the survey calibration estimators and several non-survey applications are recently rediscovered (Breslow et al.; 2009; Lumley et al.; 2011; Hainmueller; 2012; Saegusa and Wellner; 2013; Chan and Yam; 2014), but important theoretical properties have not been fully understood. In particular, it is unclear whether achieving covariate balance alone without an estimation of propensity score or outcome regression models can lead to globally semiparametric efficient estimators. This is an important theoretical property that has been established for estimators that involve nonparametric estimation of propensity score or outcome models (Hirano et al.; 2003; Imbens et al.; 2006; Chen et al.; 2008), but it is also well-known that exact matching with replacement and with a fixed number of matches can attain covariate balance but is inefficient in general (Abadie and Imbens; 2006).

There are two substantial gaps in the literature of nonparametric inference for average treatment effects that we aim to fill in this article. First, we show that a broad class of calibration estimators which solely targets on covariate balancing can attain semiparametric efficiency bound without explicitly estimating the propensity score or outcome regression functions. Compared to the seminal paper of Hirano et al. (2003) who showed that globally efficient estimation can be achieved by a nonparametric adaptation of a simple estimator by Horvitz and Thompson (1952), which has been used by statisticians for decades, we show that a globally efficient estimator can be adapted from another class of simple estimators pioneered by Deming and Stephan (1940). However, our work contains three very different conceptual aspects compared to existing survey calibration methods. The first important difference is that the proposed weights minimize a distance measure from a set of misspecified, uniform weights, whereas the original survey calibration estimators minimizes distance from the design weights, which are the unknown inverse propensity score weights for the evaluation problem. Therefore, our formulation does not involve the estimation of the unknown propensity score function. By minimizing the distance to uniform baseline weights, we improve robustness by avoiding extreme weights that typically ruin the performance of Horvitz-Thompson estimators with maximum likelihood estimated weights. The uniform baseline weights are misspecified unless the treatment is randomized, and the usual theory for survey calibration that requires a correctly specified baseline weights are therefore inapplicable. The mathematical proofs for the theorems are therefore very different from the existing results. Second, we reformulate the problem as the dual of the original calibration problem, which is a separable programming problem subject to linear constraints. The dual, as discussed in the optimization literature, is an unconstrained convex optimization problem. This reformulation allows us to provide a simple and stable algorithm for practical usage and streamlines the mathematical proofs. Third, we consider a growing number of moment conditions as opposed to a fixed number of moment conditions for survey calibration. The growing number of moment conditions is necessary for removing asymptotic bias associated with misspecified design weights while at the same time attaining global efficiency.

An equally important contribution of our paper is a novel nonparametric variance estimator for interval estimation and hypothesis testing. While there are plenty of point estimators for estimating average treatment effects, the problem of nonparametric estimation of efficient variance has received little attention because it is difficult. A consistent plugged-in estimator proposed by Hirano et al. (2003) involves the squared inverse of estimated propensity score function and can perform extremely poorly in small samples as shown in the simulation studies in Section 5. In a local semiparametric efficiency framework, consistent variance estimation often requires both the propensity score and outcome regression models to be correctly modeled despite point estimators that are often doubly robust. Due to these difficulties, many authors proposed novel point estimators while leaving the variance estimation unattended. Bootstrapping may be used but is typically computationally intensive and may not be practical to implement for large data sets. Others have suggested that the estimated weights shall be treated as fixed weights (see for example, Section 3.4 of Hainmueller, 2012). However, statistical inference can be very misleading. The variability of the estimated weights can be substantial, and in fact we illustrate using a real example in Section 6.2 to show that the standard error of the treatment effects can be underestimated by more than five fold if the weights are treated as fixed. Our proposed variance estimator is both novel and important for statistical inference in practice. It does not require direct nonparametric estimation of either the propensity score or outcome regression models. This is in contrast to Hirano et al. (2003), whose point estimator does not require nonparametric estimation of outcome regression function but that additional functional estimation is required for interval estimation. We show that the proposed estimator is consistent to the semiparametric variance bound and its validity does not depend on any parametric models; it outperforms existing estimators which require direct approximation of the efficient influence function.

The paper is organized as follows. In Section 2, we shall introduce the notations and a class of the calibration estimators, discuss related estimators that have been proposed in the literature, explain philosophical and practical differences between calibration and propensity score modeling, and study the large sample properties of the calibration estimators. A consistent asymptotic variance estimator is proposed in Section 3. In Section 4, we study three extensions of the problem: the estimation of weighted average treatment effects, treatment effects on the treated, and the estimation for multiple comparison groups. Simulation results will be presented in Section 5 and analyses of the National Health and Nutrition Examination Survey and the famous Lalonde (1986) data for the effect of job training on income are presented in Section 6. Some final remarks are given in Section 7.

The proposed methods can be implemented through an open-source R package ATE available from the Comprehensive R Archive Network (`http://cran.r-project.org/package=ATE`).

## 2. Point Estimation

### 2.1. Notations and basic framework

Let $T$ be a binary treatment indicator. We define $Y(1)$ and $Y(0)$ to be the potential outcomes when an individual is assigned to the treatment or control group respectively. The population average treatment effects is defined as $\tau \triangleq \mathbb{E}(Y(1)-Y(0))$. The estimation of $\tau$ is complicated by the fact that $Y(1)$ and $Y(0)$ cannot be observed jointly. The potential outcome $Y(1)$ is only observed when $T=1$, and $Y(0)$ is only observed when $T=0$. The observed outcome can be represented as $Y=TY(1)+(1-T)Y(0)$. In addition to $(T,Y)$, we assume that a vector of covariates $\mathbf{X}$ is observed for everyone, and $T$ is typically dependent on $(Y(1),Y(0))$ through $\mathbf{X}$. We assume the full data $\{(T_i,Y_i(1),Y_i(0),\mathbf{X}_i), i=1,\ldots,N\}$ are independent and identically distributed, and the observed data is $\{(T_i,Y_i,\mathbf{X}_i), i=1,\ldots,N\}$. The following assumption is often made for the identification of $\tau$:

Assumption 1. *(Unconfounded Treatment Assignment) Given* $\mathbf{X}$, $T$ *is independent of* $(Y(1),Y(0))$.

Based on Assumption 1, the semiparametric efficiency bound for estimating $\tau$ has been developed by Robins et al. (1994) and Hahn (1998). Let $\pi(x) \triangleq \mathbb{P}(T=1|\mathbf{X}=x)$ be the non-missing probability, also known as the propensity score, and $m_1(x) \triangleq \mathbb{E}[Y(1)|\mathbf{X}=x]$, $m_0(x) \triangleq \mathbb{E}[Y(0)|\mathbf{X}=x]$ are the conditional mean functions. Conventional statistical methods for estimating $\tau$ involves modeling of $\pi(\mathbf{X})$, $(m_1(\mathbf{X}), m_0(\mathbf{X}))$ or both, based on different representations of $\tau$:

$$\begin{aligned}
\tau &= \mathbb{E}\left[\frac{TY}{\pi(\mathbf{X})} - \frac{(1-T)Y}{1-\pi(\mathbf{X})}\right] \quad (1)\\
&= \mathbb{E}[m_1(\mathbf{X}) - m_0(\mathbf{X})] \quad (2)\\
&= \mathbb{E}\left[\frac{TY}{\pi(\mathbf{X})} - \frac{T-\pi(\mathbf{X})}{\pi(\mathbf{X})}m_1(\mathbf{X}) - \frac{(1-T)Y}{1-\pi(\mathbf{X})} - \frac{T-\pi(\mathbf{X})}{1-\pi(\mathbf{X})}m_0(\mathbf{X})\right]\ . \quad (3)
\end{aligned}$$

The inverse probability weighted estimators (Horvitz and Thompson; 1952; Hirano et al.; 2003) have been constructed based on (1); the regression prediction estimators (Oaxaca; 1973; Imbens et al.; 2006) have been proposed based on (2); and the augmented inverse probability weighted estimators (Robins et al.; 1994; Bang and Robins; 2005; Cao et al.; 2009) have been proposed based on (3).

Based on Assumption 1, another important feature for the propensity score $\pi(\mathbf{X})$ is that

$$\mathbb{E}\left[\frac{Tu(\mathbf{X})}{\pi(\mathbf{X})}\right] = \mathbb{E}\left[\frac{(1-T)u(\mathbf{X})}{1-\pi(\mathbf{X})}\right] = \mathbb{E}[u(\mathbf{X})] \ . \tag{4}$$

Recently, many authors have proposed estimators by combining (1) and (4) in various creative manners under the propensity score framework, see Qin and Zhang (2007), Tan (2010), Chan (2012), Graham et al. (2012), Vansteelandt et al. (2012), Han and Wang (2013), Imai and Ratkovic (2014) and Chan and Yam (2014). Since (4) often defines an overidentifying set of moment restrictions, estimation is generally done within the generalized method of moments or the empirical likelihood framework. These methods require modeling and estimation of the propensity score but the proposed method does not.

### 2.2. A general class of calibration estimators

Let $D(v, v_0)$ be a distance measure, for a fixed $v_0 \in \mathbb{R}$, that is continuously differentiable in $v \in \mathbb{R}$, non-negative, strictly convex in $v$ and $D(v_0, v_0) = 0$. The general idea of calibration as in Deville and Särndal (1992) is to minimize the aggregate distance between the final weights $w = (w_1, \ldots, w_N)$ to a given vector of design weights $d = (d_1, \ldots, d_N)$ subject to moment constraints. The minimum distance estimation is closely related to generalized empirical likelihood as discussed in Newey and Smith (2004). In survey applications, the design weights are known inverse probability weights. In the estimation of average treatment effects, the inverse probability weights are $d_i = \pi(\mathbf{X}_i)^{-1}$ if $T_i = 1$ or $d_i = (1 - \pi(\mathbf{X}_i))^{-1}$ if $T_i = 0$, for $i = 1, \ldots, N$, which are unknown and need to be estimated. A recent paper by Chan and Yam (2014) discussed the calibration methods with design weights estimated by maximum likelihood approach. Here we consider a different formulation. To circumvent the need to estimate the design weights, we consider a vector of misspecified uniform design weights $d^* = (1, 1, \ldots, 1)$, and construct weights $w$ by solving the following constrained minimization problem:

$$\text{Minimize} \quad \sum_{i=1}^{N} D(w_i, 1) \ ,$$

subject to the empirical counterparts of (4), that are

$$\frac{1}{N}\sum_{i=1}^{N} T_i w_i u(\mathbf{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u(\mathbf{X}_i) \quad \text{and} \quad \frac{1}{N}\sum_{i=1}^{N}(1-T_i) w_i u(\mathbf{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u(\mathbf{X}_i) \ .$$

The choice of uniform design weights is based on a few observations. First, if the counterfactual variables are observable for everyone, we can estimate $\tau$ by the sample mean of $Y(1) - Y(0)$ which assigns equal weights. Also, the need for estimating $\pi(x)$ is not needed when the design weights are assumed to be uniform. Moreover, by minimizing the aggregate distance from constant weights, the dispersion of final weights is controlled and extreme weights are less likely to obtain. In contrast, extreme weights cause instability in Horvitz-Thompson estimators with maximum likelihood weights under model misspecification. However, the choice of uniform design weights also poses unique challenges. When the number of matching conditions is fixed, Hellerstein and Imbens (1999) showed that an empirical likelihood calibration estimator with misspecified design weights yields inconsistent estimators in general. To circumvent this theoretical difficulty, we consider matching $u_K$ which is a $K(N)$-dimensional function of $\mathbf{X}$, $K(N)$ increases to infinity when $N$ goes to infinity yet with $K(N) = o(N)$.

The constrained optimization problem stated above is equivalent to two separate constrained optimization problems:

$$\text{Minimize} \quad \sum_{i=1}^{N} T_i D(Np_i, 1) \quad \text{subject to} \quad \sum_{i=1}^{N} T_i p_i u_K(\mathbf{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u_K(\mathbf{X}_i) \ , \tag{5}$$

and

$$\text{Minimize} \quad \sum_{i=1}^{N}(1-T_i)D(Nq_i,1) \quad \text{subject to} \quad \sum_{i=1}^{N}(1-T_i)q_i u_K(\mathbf{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u_K(\mathbf{X}_i) \ . \tag{6}$$

Furthermore, to efficiently implement the method, we consider the dual problems of (5) and (6). The reason is that the primal problems (5) and (6) are convex separable programming with linear constraints, and Tseng and Bertsekas (1987) showed that the dual problems are unconstrained convex maximization problems that can be solved by efficient and stable numerical algorithms.

Let $D(v) = D(v,1), f(v) = D(1-v)$ and its derivative be $f'(v)$, $\forall v \in \mathbb{R}$. The dual solutions are given as follows. For $i$ such that $T_i = 1$,

$$\hat{p}_K(\mathbf{X}_i) \triangleq \frac{1}{N}\rho'(\hat{\lambda}_K^T u_K(\mathbf{X}_i)) \ ,$$

where $\rho'$ is the first derivative of a strictly concave function

$$\rho(v) = f((f')^{-1}(v)) + v - v \cdot (f')^{-1}(v) \tag{7}$$

and $\hat{\lambda}_K \in \mathbb{R}^K$ maximizes the following objective function

$$\hat{G}_K(\lambda) \triangleq \frac{1}{N}\sum_{i=1}^{N}\left[T_i\rho(\lambda^T u_K(\mathbf{X}_i)) - \lambda^T u_K(\mathbf{X}_i)\right] \ . \tag{8}$$

Similarly, for $i$ such that $T_i = 0$,

$$\hat{q}_K(\mathbf{X}_i) \triangleq \frac{1}{N}\rho'(\hat{\beta}_K^T u_K(\mathbf{X}_i)) \ ,$$

and $\hat{\beta}_K \in \mathbb{R}^K$ maximizes the following objective function

$$\hat{H}_K(\beta) \triangleq \frac{1}{N}\sum_{i=1}^{N}\left[(1-T_i)\rho(\beta^T u_K(\mathbf{X}_i)) - \beta^T u_K(\mathbf{X}_i)\right] . \tag{9}$$

According to the first order conditions of the maximizations of (8) and (9), we can easily check that the linear constraints in (5) and (6) are satisfied. We define the proposed empirical balancing estimator for $\tau$ to be

$$\hat{\tau}_K \triangleq \sum_{i=1}^{N}\{T_i\hat{p}_K(\mathbf{X}_i)Y_i - (1-T_i)\hat{q}_K(\mathbf{X}_i)Y_i\}.$$

The relationship (7) between $\rho(v)$ and $f(v) = D(1-v)$ is shown in Appendix B, where we also show that strict convexity of $D$ is equivalent to strict concavity of $\rho$. Since the dual formulation is equivalent to the primal problem and will simplify the following discussions, we shall express the estimator in terms of $\rho(v)$ in the rest of the discussions. When $\rho(v) = -\exp(-v)$, the weights are equivalent to the implied weights of exponential tilting (Kitamura and Stutzer; 1997; Imbens et al.; 1998). When $\rho(v) = \log(1+v)$, the weights correspond to empirical likelihood (Owen; 1988; Qin and Lawless; 1994). When $\rho(v) = -(1-v)^2/2$, the weights are the implied weights of the continuous updating estimator of generalized method of moments (Hansen; 1982; Hansen et al.; 1996), and also minimizes the squared distance function. When $\rho(v) = v - \exp(-v)$, the weights are equivalent to the inverse of a logistic function.

Despite the close connections with generalized empirical likelihood, the calibration estimator has several important differences compared to the generalized empirical likelihood literature. In econometrics, generalized empirical likelihood is often employed for estimating a $p$-dimensional parameter by specifying a $q$-dimensional estimating equation, where $q > p \geq 1$. However, we are not estimating the target parameter $\tau$ by directly solving an overidentified estimating equation. The calibration conditions in (5) and (6) can be regarded as a $K$-dimensional moment restriction with a degenerate parameter of interest, and (8) and (9) are essentially degenerate cases of generalized empirical likelihood with only the auxiliary parameters $\lambda$

and $\beta$ appearing but not the target parameter $\tau$. Even though the generalized empirical likelihood estimation problem is undefined because the moment restrictions are not functions of target parameters, implied weights can still be constructed. In econometrics, the generalized empirical likelihood estimators are usually solutions to saddlepoint problems and can be difficult to compute. In our case, $\hat{\lambda}$ and $\hat{\beta}$ are solutions to unconstrained convex maximization problems rather than a saddlepoint problem and can be computed by a fast and stable Newton-type algorithm. Moreover, the generalized empirical likelihood literature mainly deals with a fixed number of moment restrictions, but the dimension $K$ of moment restrictions increases with $N$ in our present consideration. Furthermore, the moment restrictions are misspecified for finite $K$ in our case, but most theoretical results for generalized empirical likelihood are derived under a correct model specification and are therefore inapplicable.

### *2.3. Related estimators*

Existing globally efficient estimators following the expressions (1), (2) and (3) have been proposed so far without considering (4). For instance, it follows from (1) that we can estimate $\tau$ provided that a nonparametric estimator for $\pi(\mathbf{X})$ is available. Denote $\hat{\pi}(\mathbf{X})$ a series logit estimator (Geman and Hwang; 1982; Newey; 1994, 1997). Hirano et al. (2003) suggested that $\tau$ can be estimated by

$$\hat{\tau}_{HIR} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{T_iY_i}{\hat{\pi}(\mathbf{X}_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(\mathbf{X}_i)}\right] .$$

Alternatively, it follows from (2) that the average treatment effects $\tau$ can be estimated provided that a nonparametric estimator for $m_1(\mathbf{X})$ and $m_0(\mathbf{X})$ are available. Under Assumption 1, $m_1(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}, T = 1]$ and $m_0(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}, T = 0]$. Therefore, $m_1(x)$ and $m_0(x)$ can be estimated using data from the treated group and the control group respectively by polynomial series estimators (Newey; 1994, 1997), denoted by $\hat{m}_1(x)$ and $\hat{m}_0(x)$. Imbens et al. (2006) suggested the estimator

$$\hat{\tau}_{INR} \triangleq \frac{1}{N}\sum_{i=1}^{N}[\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)] ,$$

where the average is taken over the full sample, including the observations for which $Y_i(1)$ and $Y_i(0)$ are not observed, and as a result $\hat{\tau}_{INR}$ is called an imputation estimator in Imbens et al. (2006). Chen et al. (2008) considered a more general setting for M-estimation under moment restrictions and proposed an estimator by first projecting estimating equations onto basis functions of a series estimator, which is similar to $\hat{\tau}_{INR}$ estimating treatment effects. Hahn (1998) proposed a different estimator by first estimating the three conditional expectations $\mathbb{E}[YT|\mathbf{X} = x], \mathbb{E}[Y(1-T)|\mathbf{X} = x]$ and $\pi(x)$, and used these estimated conditional expectations to estimate the two regression functions $m_1(x)$ and $m_0(x)$. Since $m_0(x) = \mathbb{E}[Y(1-T)|\mathbf{X} = x]/(1-\pi(x))$ and $m_1(x) = \mathbb{E}[YT|\mathbf{X} = x]/\pi(x)$, Hahn (1998) suggested the estimator

$$\hat{\tau}_H \triangleq \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{\mathbb{E}}[YT|\mathbf{X} = \mathbf{X}_i]}{\hat{\pi}(\mathbf{X}_i)} - \frac{\hat{\mathbb{E}}[Y(1-T)|\mathbf{X} = \mathbf{X}_i]}{1-\hat{\pi}(\mathbf{X}_i)} .$$

Imbens et al. (2006) further suggested a modified estimator $\hat{\tau}_{mod}$, that combines the features of $\hat{\tau}_{INR}$ and $\hat{\tau}_{HIR}$, defined by

$$\hat{\tau}_{mod} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{T_i\hat{m}_1(\mathbf{X}_i)}{\hat{\pi}(\mathbf{X}_i)} - \frac{(1-T_i)\hat{m}_0(\mathbf{X}_i)}{1-\hat{\pi}(\mathbf{X}_i)}\right] .$$

It has been shown by Imbens et al. (2006) that $\hat{\tau}_{HIR}, \hat{\tau}_{INR}, \hat{\tau}_{mod}$ and $\hat{\tau}_H$ attains the semiparametric efficiency bound for the estimation of $\tau$ under Assumption 1 and some regularity conditions, and they all admit the following asymptotic expansion

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{T_iY_i}{\pi(X_i)} - \frac{T_i - \pi(\mathbf{X}_i)}{\pi(\mathbf{X}_i)}m_1(\mathbf{X}_i) - \frac{(1-T_i)Y_i}{1-\pi(\mathbf{X}_i)} - \frac{T_i - \pi(\mathbf{X}_i)}{1-\pi(\mathbf{X}_i)}m_0(\mathbf{X}_i)\right] + o_p(N^{-1/2}) .$$

The estimators mentioned are all globally semiparametric efficient, which holds for arbitrary $\pi(x)$, $m_0(x)$ and $m_1(x)$ subject to some mild smoothness conditions. It is different from local semiparametric efficient estimators, where semiparametric efficiency holds only for a restricted submodel, usually under the correct specification of parametric models for $\pi(x)$, $m_0(x)$ and $m_1(x)$.

Locally efficient estimators have been proposed that incorporates (4) in the estimation under the propensity score or outcome regression framework. Qin and Zhang (2007) proposed an empirical likelihood estimator that matches the sample moment of covariates and the propensity score, where the propensity score parameters are estimated by maximum likelihood approach. Since the propensity score is estimated, the empirical likelihood estimator does not guarantee efficiency improvement even when additional moment restrictions are added. Chan (2012) proposed a modified empirical likelihood estimator that guarantees the asymptotic variance to be non-increasing whenever additional moment conditions are included. Han and Wang (2013) and Han (2014) showed that the empirical likelihood estimator of Qin and Zhang (2007) is multiply robust. A related but different class of generalized empirical likelihood estimators is shown to be multiply robust by Chan and Yam (2014), where the propensity score parameters are estimated by maximizing the underlying likelihood. Tan (2010) and Graham et al. (2012) over-parameterize the propensity score model by including contributions of the outcome regression model as predictors in their extended propensity score models. Estimation of the parameters in those extended propensity score models is done by the method of moments or empirical likelihood, based on estimating equations defined by different moment balancing constraints. Imai and Ratkovic (2014) noted that the inclusion of outcome model in propensity score modeling does not align with the original spirit of the propensity score methods (Rubin; 2007). Instead of over-parameterizing the propensity score model, they constructed an over-identified set of estimating equations which combines the likelihood score equation of the propensity score with additional moment balancing conditions.

Survey calibration improves a given set of design weights by calibration weights that satisfy certain moment constraints. Many survey calibration estimators correspond to special cases of $\rho(v)$. The exponential tilting calibration estimator is often known as raking, which dates back to Deming and Stephan (1940) and Deville et al. (1993). The empirical likelihood calibration estimator is discussed in Chen and Sitter (1999), Wu and Sitter (2001) and Chen et al. (2002), among others. The quadratic $\rho(v)$ gives rise to the generalized regression estimator of Cassel et al. (1976). A general formulation of calibration estimators is given in Deville and Särndal (1992) and Kim and Park (2010). These estimators are usually used to improve finite sample estimates when population totals of auxiliary variables are known. The connections between the survey calibration estimators and some non-survey applications were recently rediscovered (Breslow et al.; 2009; Lumley et al.; 2011; Hainmueller; 2012; Saegusa and Wellner; 2013; Chan and Yam; 2014). Breslow et al. (2009) and Saegusa and Wellner (2013) studied calibration for outcome dependent sampling in epidemiology. Lumley et al. (2011) studied connections with genetic epidemiology and measurement error problems. Hainmueller (2012) used calibration for pre-processing data for program evaluation. Chan and Yam (2014) reviewed the connections among many calibration-type estimators proposed independently in survey sampling, biostatistics and econometrics literature for a missing response problem, in which the missing data mechanism is modeled parametrically.

### 2.4. Philosophical differences and practical implications

Although the calibration weights in Section 2.2 are also constructed from moment balancing conditions as in certain propensity score methodologies, there is a fundamental difference in the modeling philosophy that leads to important practical implications in the estimation of average treatment effects. Philosophically, the calibration weights are constructed without any reference to a propensity score model. It ignores the explicit link between the weights in the treated and the control groups that are present in propensity score modeling. Practically, it leads to an exact three-way balance between the treated, the controls and the combined group for finite samples as well as asymptotically, whereas finite-sample exact three-way balance is not guaranteed for propensity score modeling in general. Furthermore, calibration can be viewed as a unification of the existing globally efficient estimations that are constructed from different modeling strategies as discussed in Section 2.3.

To illustrate the first idea, we consider a general class of weighting estimators $N^{-1}\sum_{i=1}^{N}\{T_i w_1(\mathbf{X}_i)Y_i - (1-T_i)w_0(\mathbf{X}_i)Y_i\}$. If the true propensity score $\pi(\mathbf{X})$ is known, we set $w_1(\mathbf{X}) = (\pi(\mathbf{X}))^{-1}$ and $w_0(\mathbf{X}) = (1-\pi(\mathbf{X}))^{-1}$, so that the corresponding weighting estimator is an unbiased estimator for $\tau$ based on (1). The propensity score setting confines the weight functions $w_1(x)$ and $w_0(x)$ in the following relationship:

$$w_0(x) = (1-(w_1(x))^{-1})^{-1} \; . \tag{10}$$

When a propensity score model $\pi(x;\gamma)$ is assumed, and $\hat{\gamma}$ is an estimate of $\gamma$, the Horvitz-Thompson estimator sets $w_1(x) = \pi^{-1}(x;\hat{\gamma})$ and $w_0(x) = (1-\pi(x;\hat{\gamma}))^{-1}$. It follows that (10) is satisfied. Under model misspecification, $C_1 = N^{-1}\times\sum_{i=1}^{N} T_i\pi^{-1}(\mathbf{X}_i;\hat{\gamma})$ and $C_0 = N^{-1}\times\sum_{i=1}^{N}(1-T_i)(1-\pi(\mathbf{X}_i;\hat{\gamma}))^{-1}$ can be very different from 1, therefore it has been suggested that $w_1(x) = [C_1\pi(x;\hat{\gamma})]^{-1}$ and $w_0(x) = [C_0(1-\pi(x;\hat{\gamma}))]^{-1}$ which are ratio-type inverse probability weighting estimators. However, these weight functions do not satisfy (10) unless $C_1 = C_0 = 1$. In fact, the ratio-type inverse probability weighting estimator is a special calibration estimator that requires propensity score modeling to be discussed in Section 7. The class of calibration estimators in Section 2.2 does not rely on propensity score modeling in the first place, and the weights $w_1(x) = \rho'(\hat{\lambda}^T u_K(x))$ and $w_0(x) = \rho'(\hat{\beta}^T u_K(x))$ do not satisfy (10) in general. We note that one of the two weights can correspond to the inverse probability weight from a propensity score model, but it is generally impossible to have both sets of weights to be consistent with a single propensity score model. Therefore, for any fixed $K$, the calibration estimator for the average treatment effects cannot be locally efficient. Despite this seemingly undesirable property, we shall show that $\hat{\tau}$ is globally semiparametric efficient when $K$ is allowed to increase with $N$. Existing globally efficient estimators are all locally efficient for any fixed dimension of approximation, but the calibration estimator sacrifices local efficiency by ignoring the link (10), while achieving exact three-way balance in finite samples. Note that the true propensity score attains the exact three-way balance in (4), but the estimated propensity scores typically do not achieve exact three-way balance in finite samples. While the Horvitz-Thompson estimator with maximum likelihood weights is globally efficient and three-way balance should hold for extremely large samples, the balance can be quite poor for practical sample sizes. In our current proposal, the exact three-way balance holds for finite samples as well as asymptotically, which is a reason why our asymptotic results for the point and variance estimators hold well even for finite samples.

To further illustrate the balancing properties of estimators, suppose that $\pi(\gamma^T x)$ is a propensity score model. The expression (4) leads to

$$\mathbb{E}(b_1(T,\mathbf{X};\gamma)) \triangleq \mathbb{E}\left[\left(\frac{T}{\pi(\gamma^T\mathbf{X})} - 1\right)\mathbf{X}\right] = 0 \; , \tag{11}$$

$$\mathbb{E}(b_2(T,\mathbf{X};\gamma)) \triangleq \mathbb{E}\left[\left(\frac{1-T}{1-\pi(\gamma^T\mathbf{X})} - 1\right)\mathbf{X}\right] = 0 \; , \tag{12}$$

$$\mathbb{E}(b_3(T,\mathbf{X};\gamma)) \triangleq \mathbb{E}\left[\left(\frac{T}{\pi(\gamma^T\mathbf{X})} - \frac{1-T}{1-\pi(\gamma^T\mathbf{X})}\right)\mathbf{X}\right] = 0 \; . \tag{13}$$

Suppose we estimate $\gamma$ by solving each of the just-identified system of estimating equations defined by moment conditions (11), (12) and (13), and denote the corresponding estimators to be $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_3$; that is, $\hat{\gamma}_j$ satisfies $N^{-1}\times\sum_{i=1}^{N} b_j(T_i,\mathbf{X}_i;\hat{\gamma}_j) = 0$, for $j = 1,2,3$. Note that, however, $N^{-1}\times\sum_{i=1}^{N} b_j(T_i,\mathbf{X}_i;\hat{\gamma}_{j'}) \neq 0$ for $j \neq j'$. The covariate balancing propensity score of Imai and Ratkovic (2014) shares the same spirit of $\hat{\gamma}_3$, and the inverse probability tilting method of Graham et al. (2012) shares the same spirit of $\hat{\gamma}_1$. In general, matching one set of moment conditions creates two-way balance, but does not guarantee three-way balance between the treated, the controls and the combined group. However, a lack of three-way balance can adversely affect the quality of the final estimate since the average treatment effects is defined for the combined population, and the data for $Y(1)$ and $Y(0)$ are only available for the treated and controls respectively. We shall further illustrate this point by the simulation studies in Section 5. On the other hand, four-or-more-way balance is not necessary because asymptotic efficiency is attained by three-way balance as shown in Theorem 1 in Section 2.5.

Since $b_3 = b_1 - b_2$, balancing any two systems out of (11), (12) and (13) can lead to a balance of the remaining system. Therefore, by considering an overidentified combined system of estimating functions $b_1$ and $b_2$, one can estimate $\gamma$ by using generalized method of moments or empirical likelihood, and we denote the corresponding estimator by $\hat{\gamma}_{12}$. However, it is typical that

$$\frac{1}{N}\sum_{i=1}^{N} b_j(T_i, \mathbf{X}_i, \hat{\gamma}_{12}) \neq 0, \qquad j = 1, 2, 3,$$

because the generalized method of moment estimator does not solve that corresponding overidentified system exactly, and therefore the exact three-way balance is typically not achieved.

For the calibration estimator, however,

$$\sum_{i=1}^{N} T_i \hat{p}_K(\mathbf{X}_i) u_K(\mathbf{X}_i) = \sum_{i=1}^{N} (1 - T_i) \hat{q}_K(\mathbf{X}_i) u_K(\mathbf{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u_K(\mathbf{X}_i) \ ,$$

by construction, and exact three-way balance can naturally be achieved.

Several remarks are in order. First, when the propensity score model is misspecified, which is likely in practice, $\hat{\gamma}_j$ converges in probability to $\gamma_j^*$, $j = 1, 2, 3$, which are different in general. Therefore, covariate balancing based on one of (11), (12) or (13) can lead to very different results. For balancing an over-identified system of equations using the empirical likelihood, there is no guarantee that the $\gamma$ estimate is $\sqrt{N}$-consistent (Schennach; 2007) under a misspecified propensity score model. Calibration is similar to using $\hat{\gamma}_1$ for reweighting the treated and $\hat{\gamma}_2$ for reweighting the controls when the propensity score model is misspecified, but our proposed non-parametric calibration method does not involve propensity score estimation.

Despite the dissimilarities in the weighting aspects compared to propensity score methodologies, calibration can be viewed as a unification of the existing globally efficient estimation that are constructed from two very different strategies: weighting and prediction. As discussed in Section 2.3, $\hat{\tau}_{HIR}$ is a weighting estimator, $\hat{\tau}_{INR}$ is based on prediction, $\hat{\tau}_H$ and $\hat{\tau}_{mod}$ are based on a combination of the two strategies. Let $\tilde{m}_1(\mathbf{X})$ and $\tilde{m}_0(\mathbf{X})$ be weighted least square estimators for $Y$ against $u_K(\mathbf{X})$ among the treated and controls with weights $\hat{p}_K(\mathbf{X})$ and $\hat{q}_K(\mathbf{X})$ respectively. It follows that

$$\begin{aligned}
& \sum_{i=1}^{N} T_i \hat{p}_K(\mathbf{X}_i) Y_i - \sum_{i=1}^{N} (1 - T_i) \hat{q}_K(\mathbf{X}_i) Y_i \\
= & \sum_{i=1}^{N} T_i \hat{p}_K(\mathbf{X}_i) \tilde{m}_1(\mathbf{X}_i) - \sum_{i=1}^{N} (1 - T_i) \hat{q}_K(\mathbf{X}_i) \tilde{m}_0(\mathbf{X}_i) \\
= & \sum_{i=1}^{N} \tilde{m}_1(\mathbf{X}_i) - \tilde{m}_0(\mathbf{X}_i) \ .
\end{aligned}$$

The first equality holds from the score equation of weighted least squares, and the second equality holds because of the exact three-way balance. Therefore, calibration unifies $\hat{\tau}_{HIR}$, $\hat{\tau}_{INR}$, $\hat{\tau}_H$ and $\hat{\tau}_{mod}$ by a rather unexpected way of relaxing the propensity score and outcome regression estimations.

### 2.5. Large sample properties of calibration estimators

To show the large sample properties, we need the following technical assumptions in addition to Assumption 1.

Assumption 2. $\mathbb{E}[Y(1)^2] < \infty$ *and* $\mathbb{E}[Y(0)^2] < \infty$.

Assumption 3 (Distribution of $\mathbf{X}$). *The support $\mathcal{X}$ of $r$-dimensional covariate $\mathbf{X}$ is a Cartesian product of $r$ compact intervals.*

Assumption 4. *$\pi(x)$ is uniformly bounded away from 0 and 1, i.e. there exist some constants $\frac{1}{\eta_1} \triangleq \inf_{x\in\mathcal{X}} \pi(x)$, $\frac{1}{\eta_2} \triangleq \sup_{x\in\mathcal{X}} \pi(x)$ such that*

$$0 < \frac{1}{\eta_1} \leq \pi(x) \leq \frac{1}{\eta_2} < 1 \quad \forall x \in \mathcal{X}$$

*where $\mathcal{X}$ is the support of $\mathbf{X}$.*

Assumption 5. *$\pi(x)$ is $s$-times continuously differentiable, where $s > 13r$.*

Assumption 6. *$m_0(x)$ and $m_1(x)$ are $t$-times continuously differentiable, where $t > \frac{3r}{2}$.*

Assumption 7. *$K = O\left(N^{\nu}\right)$ and $\frac{1}{\frac{s}{r} - 2} < \nu < \frac{1}{11}$.*

Assumption 8. *$\rho$ is a strictly concave function defined on $\mathbb{R}$ i.e. $\rho''(\gamma) < 0$, $\forall\gamma \in \mathbb{R}$, and the range of $\rho'$ contains $[\eta_2, \eta_1]$ which is a subset of the positive real line.*

Assumptions 1-7 or their analogues also appeared in Hahn (1998), Hirano et al. (2003) and Imbens et al. (2006). Assumption 1 is required for the identification of the average treatment effects. Assumption 2 is required for the finiteness of asymptotic variance. Assumptions 3 and 4 are required for uniform boundedness of approximations. Assumption 4 is an overlap condition that is necessary for the nonparametric identification of the average treatment effects in the population. If there exists a region of $\mathbf{X}$ such that the probability of receiving treatment is 0 or 1, the treatment effects cannot be identified unless some extrapolatory modeling assumptions are imposed. In that case, one could define a subpopulation with a sufficient overlap so that the average treatment effects can be estimated nonparametrically within this subpopulation. Assumptions 5 and 6 are required for controlling the remainder of approximations with a given basis function. They are standard assumptions for multivariate smoothing where the order of smoothness required increases with the dimension of $\mathbf{X}$. There is usually no *a-priori* reason to believe that the $\pi(x)$, $m_1(x)$ and $m_0(x)$ are not smooth in $x$. Also the dimension of $\mathbf{X}$ is not restricted by the assumptions, unlike in the kernel estimation of $\pi(x)$ discussed in Chen et al. (2013), in which their assumptions imply that the dimension of $\mathbf{X}$ cannot be greater than 4. Assumption 7 is required for controlling the stochastic order of the residual terms, which is desirable in practice because $K$ grows very slowly with $N$ so that a relatively small number of moment conditions is sufficient for the proposed method to perform well. Assumption 8 is a mild assumption on $\rho$ which is chosen by the statisticians and includes all the important special cases considered in the literature. In contrast, the theoretical results for Hahn (1998), Hirano et al. (2003), Imbens et al. (2006) and Chen et al. (2008) were developed only for linear or logistic models for propensity score.

Define $\mu_0 \triangleq \mathbb{E}[Y(0)]$, $\mu_1 \triangleq \mathbb{E}[Y(1)]$, $\sigma_1^2(\mathbf{X}) = Var(Y(1)|\mathbf{X})$ and $\sigma_0^2(\mathbf{X}) = Var(Y(0)|\mathbf{X})$, which are finite by Assumption 2. We have the following theorem.

Theorem 1. *Under Assumptions 1-8, we have*

*(a)* $\sum_{i=1}^{N}\{T_i\hat{p}_K(\mathbf{X}_i)Y_i - (1-T_i)\hat{q}_K(\mathbf{X}_i)Y_i\} \xrightarrow{p} \tau$;

*(b)* $\sqrt{N}\left(\sum_{i=1}^{N}\{T_i\hat{p}_K(\mathbf{X}_i)Y_i - (1-T_i)\hat{q}_K(\mathbf{X}_i)Y_i\} - \tau\right) \xrightarrow{d} \mathcal{N}(0, V_{semi})$, *where*

*$V_{semi} \triangleq \mathbb{E}\left[(m_1(\mathbf{X}) - m_0(\mathbf{X}) - \tau)^2 + \frac{\sigma_1^2(\mathbf{X})}{\pi(\mathbf{X})} + \frac{\sigma_0^2(\mathbf{X})}{1-\pi(\mathbf{X})}\right]$ attains the semi-parametric efficiency bound as shown in Robins et al. (1994) and Hahn (1998).*

A detailed proof of Theorem 1 will be provided in the supplementary materials. We note that our global efficiency result is substantially more general than existing results in the literature, in which global efficiency for weighting estimators has only been established for two particular estimators for the propensity score: the series least square estimator (Hahn; 1998; Chen et al.; 2008), and the maximum likelihood series logit estimator (Hirano et al.; 2003; Imbens et al.; 2006). The proof of lemma B.2 in Chen et al. (2008) relies crucially on the least square property of the projection of $T$ on the approximation basis. The validity of the result in Hirano et al. (2003) requires a key fact about the least square projection of $-\mathbb{E}(Y(1)|\mathbf{X}=x)\pi^{-1}(x)\sqrt{\pi(x)(1-\pi(x))}$ onto a transformed approximation basis $\sqrt{\pi^*(x)(1-\pi^*(x))}u_K(x)$, where $\pi^*(x)$ is defined in terms of the limit of a maximum likelihood estimator under a logistic regression model. Their projection argument yields an asymptotically negligible residual term only when the series maximum likelihood logit estimator is used. We are able to establish the global efficiency results for any strictly concave $\rho$ satisfying Assumption 8 because we employed a different and more delicate projection argument. We studied a weighted least square projection of $-\mathbb{E}(Y(1)|\mathbf{X}=x)$ and $-\mathbb{E}(Y(0)|\mathbf{X}=x)$ onto the original approximation basis $u_K(x)$, where $\rho$ only enters the weights of the projection, but not the approximation basis. Our projection argument yields an asymptotically negligible residual term when the weights of the projection are bounded from above and below, which was established under our regularity conditions. Theorem 1 holds even when the $\rho$ functions used for computing $\hat{p}_K$ and $\hat{q}_K$ are different. However, we do not recommend this in practice, because each $\rho(v)$ corresponds to a measure of distance $D(v)$ from the unit weight, and usually there is not any justifiable reason for choosing a different measure for the two treatment groups.

## 3. A nonparametric variance estimator

By Theorem 1, the proposed estimator attains the semiparametric efficiency bound with the following asymptotic variance: $V_{semi} \triangleq \mathbb{E}\left[(m_1(\mathbf{X}) - m_0(\mathbf{X}) - \tau)^2 + \frac{\sigma_1^2(\mathbf{X})}{\pi(\mathbf{X})} + \frac{\sigma_0^2(\mathbf{X})}{1-\pi(\mathbf{X})}\right]$. Since the variance involves unknown functions $\pi(x), m_1(x)$ and $m_0(x)$, plug-in estimators typically involve nonparametric estimation of functions other than $\pi(x)$, as in Hirano et al. (2003). One of the advantages of our proposed point estimator is that we do not need to directly estimate those functions, and it would be nice to have a variance estimator that also does not involve any additional estimates of nonparametric functions. Moreover, existing nonparametric estimators that require estimation of $\pi(x)$ often fail in practice, as illustrated in Table 4 of Section 5. A particular thorny issue is that the asymptotic variance estimators often depend on the squared inverse of estimated propensity score, and the instability caused by extreme inverse weights is even magnified. In this section, we shall study a consistent asymptotic variance estimator that can be computed easily from the point estimator and avoids the problem of extreme weights.

Define

$$\begin{aligned}
&g_{K1}(T,\mathbf{X};\lambda) \triangleq T\rho'(\lambda^T u_K(\mathbf{X}))u_K(\mathbf{X}) - u_K(\mathbf{X}) \ ,\\
&g_{K2}(T,\mathbf{X};\beta) \triangleq (1-T)\rho'(\beta^T u_K(\mathbf{X}))u_K(\mathbf{X}) - u_K(\mathbf{X}) \ ,\\
&g_{K3}(T,\mathbf{X},Y;\theta) \triangleq T\rho'(\lambda^T u_K(\mathbf{X}))Y - (1-T)\rho'(\beta^T u_K(\mathbf{X}))Y - \tau\\
&g_K(T,\mathbf{X},Y;\theta) \triangleq \begin{pmatrix} g_{K1}(T,\mathbf{X};\lambda) \\ g_{K2}(T,\mathbf{X};\beta) \\ g_{K3}(T,\mathbf{X},Y;\theta) \end{pmatrix} \ ,
\end{aligned}$$

where $\theta \triangleq (\lambda,\beta,\tau)^T$. Also define $\hat{\theta}_K \triangleq (\hat{\lambda}_K, \hat{\beta}_K, \hat{\tau}_K)^T$, $\theta_K^* \triangleq (\lambda_K^*, \beta_K^*, \tau_K^*)^T$ and $\tau_K^* \triangleq \mathbb{E}[TNp_K^*(\mathbf{X})Y - (1-T)Nq_K^*(\mathbf{X})Y]$.

Note that $\hat{\theta}_K$ satisfies

$$\frac{1}{N}\sum_{i=1}^{N} g_K(T_i, \mathbf{X}_i, Y_i; \hat{\theta}_K) = 0.$$

Taylor series expansion on the left hand side at $\theta_K^*$ yields

$$0 = \frac{1}{N}\sum_{i=1}^{N} g_K(T_i, \mathbf{X}_i, Y_i; \theta_K^*) + \frac{1}{N}\sum_{i=1}^{N} \frac{\partial g_K(T_i, \mathbf{X}_i, Y_i; \tilde{\theta}_K)}{\partial \theta}(\hat{\theta}_K - \theta_K^*) \ , \tag{14}$$

where $\tilde{\theta}_K = (\tilde{\lambda}_K, \tilde{\beta}_K, \tilde{\tau}_K)^T$ lies on the line joining $\hat{\theta}_K$ and $\theta_K^*$. We shall show in the supplementary material that

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial g_K(T_i, \mathbf{X}_i, Y_i; \tilde{\theta}_K)}{\partial \theta} = \mathbb{E}\left[\frac{\partial g_K(T, \mathbf{X}, Y; \theta_K^*)}{\partial \theta}\right] + o_p(1) \ , \tag{15}$$

where

$$\mathbb{E}\left[\frac{\partial g_K(T, \mathbf{X}, Y; \theta_K^*)}{\partial \theta}\right] = \begin{pmatrix} A_{2K\times 2K}, & B_{2K\times 1} \\ C_{1\times 2K}, & D_{1\times 1} \end{pmatrix},$$

and

$$\begin{aligned}
&A_{2K\times 2K} \triangleq \begin{pmatrix} \mathbb{E}[T\rho''((\lambda_K^*)^T u_K(\mathbf{X}))u_K(\mathbf{X})u_K^T(\mathbf{X})], & 0_{K\times K} \\ 0_{K\times K}, & \mathbb{E}[(1-T)\rho''((\beta_K^*)^T u_K(\mathbf{X}))u_K(\mathbf{X})u_K^T(\mathbf{X})] \end{pmatrix} \ , \\
&B_{2K\times 1} \triangleq 0_{2K\times 1} \ , \\
&C_{1\times 2K} \triangleq (\mathbb{E}[T\rho''((\lambda_K^*)^T u_K(\mathbf{X}))Yu_K^T(\mathbf{X})], -\mathbb{E}[(1-T)\rho''((\beta_K^*)^T u_K(\mathbf{X}))Yu_K^T(\mathbf{X})]) \ , \\
&D_{1\times 1} \triangleq -1 \ .
\end{aligned}$$

However, note that we are only interested in the limiting behavior of $Var(\sqrt{N}(\hat{\tau}_K - \tau_K^*))$, which is the last element of $Var(\sqrt{N}(\hat{\theta}_K - \theta_K^*))$ when $N \uparrow \infty$, this leads us to consider the last row of $\mathbb{E}\left[\frac{\partial g_K(T,\mathbf{X},Y;\theta_K^*)}{\partial \theta}\right]^{-1}$ which is

$$\left(C_{1\times 2K} \cdot A_{2K\times 2K}^{-1}, -1\right) = (L_K, R_K, -1) \ ,$$

where

$$\begin{aligned}
&L_K \triangleq \mathbb{E}[T\rho''((\lambda_K^*)^T u_K(\mathbf{X}))Yu_K^T(\mathbf{X})] \cdot \mathbb{E}[T\rho''((\lambda_K^*)^T u_K(\mathbf{X}))u_K(\mathbf{X})u_K^T(\mathbf{X})]^{-1} \ , \\
&R_K \triangleq -\mathbb{E}[(1-T)\rho''((\beta_K^*)^T u_K(\mathbf{X}))Yu_K^T(\mathbf{X})] \cdot \mathbb{E}[(1-T)\rho''((\beta_K^*)^T u_K(\mathbf{X}))u_K(\mathbf{X})u_K^T(\mathbf{X})]^{-1} \ .
\end{aligned}$$

Since $\lim_{K\to\infty} \mathbb{E}[g_K(T, \mathbf{X}, Y; \theta_K^*)]^T] = 0$ (the zero vector), by (39) in the supplementary material we could get:

$$\begin{aligned}
&\lim_{K\to\infty} Var(\sqrt{N}(\hat{\tau}_K - \tau_K^*)) \qquad (16) \\
&= \lim_{K\to\infty} (L_K, R_K, -1)\mathbb{E}[g_K(T, \mathbf{X}, Y; \theta_K^*)g_K^T(T, \mathbf{X}, Y; \theta_K^*)] \begin{pmatrix} L_K^T \\ R_K^T \\ -1 \end{pmatrix} \\
&= \lim_{K\to\infty} (L_K, R_K, -1)P_K \begin{pmatrix} L_K^T \\ R_K^T \\ -1 \end{pmatrix} \ ,
\end{aligned}$$

where $P_K \triangleq \mathbb{E}[g_K(T,\mathbf{X},Y;\theta_K^*)g_K^T(T,\mathbf{X},Y;\theta_K^*)]$.

This motivates us to define our estimator for the asymptotic variance by:

$$\hat{V}_K \triangleq (\hat{L}_K, \hat{R}_K, -1)\hat{P}_K \begin{pmatrix} \hat{L}_K^T \\ \hat{R}_K^T \\ -1 \end{pmatrix} ,$$

where

$$\hat{L}_K \triangleq \left[\frac{1}{N}\sum_{i=1}^N T_i\rho''(\hat{\lambda}_K^T u_K(\mathbf{X}_i))u_K^T(\mathbf{X}_i)Y_i\right] \times \left[\frac{1}{N}\sum_{i=1}^N T_i\rho''(\hat{\lambda}_K^T u_K(\mathbf{X}_i))u_K(\mathbf{X}_i)u_K^T(\mathbf{X}_i)\right]^{-1} ,$$

$$\hat{R}_K \triangleq -\left[\frac{1}{N}\sum_{i=1}^N (1-T_i)\rho''(\hat{\beta}_K^T u_K(\mathbf{X}_i))u_K^T(\mathbf{X}_i)Y_i\right] \times \left[\frac{1}{N}\sum_{i=1}^N (1-T_i)\rho''(\hat{\beta}_K^T u_K(\mathbf{X}_i))u_K^T(\mathbf{X}_i)u_K(\mathbf{X}_i)\right]^{-1} ,$$

$$\hat{P}_K \triangleq \frac{1}{N}\sum_{i=1}^N g_K(T_i,\mathbf{X}_i,Y_i;\hat{\theta}_K)g_K^T(T_i,\mathbf{X}_i,Y_i;\hat{\theta}_K) .$$

Although the construction of the proposed variance estimator did not start with a direct approximation of the influence function, the variance estimator can be written as $\hat{V}_K = N^{-1} \times \sum_{i=1}^N \hat{\varphi}_{CAL}^2(T_i,\mathbf{X}_i,Y_i;\hat{\theta})$ where

$$\begin{aligned} \hat{\varphi}_{CAL}(T,\mathbf{X},Y) &= -(L_K, -R_K, -1)g_K(T,\mathbf{X},Y;\hat{\theta}) \\ &= g_{K3}(T,\mathbf{X},Y;\hat{\theta}) - L_K g_{K1}(T,\mathbf{X};\hat{\lambda}) + R_K g_{K2}(T,\mathbf{X};\hat{\beta}) , \end{aligned}$$

which is an estimator of the efficient influence function:

$$\varphi_{eff}(T,\mathbf{X},Y) = \frac{TY}{\pi(\mathbf{X})} - \frac{(1-T)Y}{1-\pi(\mathbf{X})} - \tau + (T-\pi(\mathbf{X}))\beta(\mathbf{X}) , \tag{17}$$

$$\beta(\mathbf{X}) = -\left[\frac{m_1(\mathbf{X})}{\pi(\mathbf{X})} + \frac{m_0(\mathbf{X})}{1-\pi(\mathbf{X})}\right] .$$

Comparing $\varphi_{eff}$ and $\hat{\varphi}_{CAL}$, the proposed variance estimator would have a good performance if $-L_K g_{K1}(T,\mathbf{X};\hat{\lambda}) + R_K g_{K2}(T,\mathbf{X};\hat{\beta})$ is a good but indirect approximation of $(T-\pi(\mathbf{X}))\beta(\mathbf{X})$. This is particularly true because of approximation results that are established in the proof of Theorem 1 for the terms (29) and (30) given in Appendix A. In summary, we have the following theorem:

THEOREM 2. *Under Assumptions 1-8 with Assumption 2 being strengthened to* $\mathbb{E}(Y^4(1)) < \infty$ *and* $\mathbb{E}(Y^4(0)) < \infty$*,* $\hat{V}_K$ *is a consistent estimator for the asymptotic variance* $V_{semi}$*.*

The proof of Theorem 2 is given in the supplementary material. The strengthened condition in Theorem 2 is mainly used in (71) in the supplementary material so as to ensure the consistency of the asymptotic variance; and this condition is mild and naturally holds for practical samples. The results illuminated the advantages of using the proposed variance estimator for statistical inference. We note that the proposed variance estimator can pair with any globally efficient estimators for valid inference, since all of them are asymptotically equivalent. However, the computation of the proposed estimator only requires intermediate input from the calibration estimation and does not require direct estimation of propensity score or outcome regression function, therefore it pairs naturally with the calibration estimators.

## 4. Related estimation problems

In this section we illustrate that calibration weighting can be easily extended to several related problems. All proofs of the main theorems in this section are very similar to that of Theorem 1, and they are omitted. The estimators for asymptotic variances in this section, namely $V_{semi}^g, V_{semi}^{\pi}, V_{semi}^{ATT}$, and $V_{jl}$ for $i, l \in \mathcal{J} = \{0, \dots, J-1\}, J \geq 2$, and their corresponding consistent estimation results can be derived similarly by using the approach founded in Theorem 2.

### 4.1. Weighted average treatment effects

Our estimator can be easily extended to estimate a weighted average treatment effects

$$\tau_{WATE}^g = \frac{\int \mathbb{E}[Y(1) - Y(0) \mid \mathbf{X} = x] g(x) dF(x)}{\int g(x) dF(x)}$$

where $g$ is a known function of the covariates. To estimate $\tau_{WATE}^g$, we can define $\hat{p}_K = N^{-1}\rho'\left(\hat{\lambda}_K^T u_K(\mathbf{X})\right)$ and $\hat{q}_K = N^{-1}\rho'\left(\hat{\beta}_K^T u_K(\mathbf{X})\right)$, with $\hat{\lambda}_K$ and $\hat{\beta}_K$ being replaced by maximizers of slightly different objective functions:

$$\hat{G}_K(\lambda) \triangleq \frac{1}{N}\sum_{i=1}^N \left[T_i \rho(\lambda^T u_K(\mathbf{X}_i)) - \lambda^T \bar{u}_K^g\right]$$

and

$$\hat{H}_K(\beta) \triangleq \frac{1}{N}\sum_{i=1}^N \left[(1-T_i)\rho(\beta^T u_K(\mathbf{X}_i)) - \beta^T \bar{u}_K^g\right] ,$$

where $\bar{u}_K^g = \sum_{i=1}^N g(\mathbf{X}_i) u_K(\mathbf{X}_i) / \sum_{i=1}^n g(\mathbf{X}_i)$. Therefore, $\hat{p}_K$ and $\hat{q}_K$ satisfies

$$\sum_{i=1}^N T_i \hat{p}_K(\mathbf{X}_i) u_K(\mathbf{X}_i) = \bar{u}_K^g \tag{18}$$

and

$$\sum_{i=1}^N (1-T_i)\hat{q}_K(\mathbf{X}_i) u_K(\mathbf{X}_i) = \bar{u}_K^g \ .$$

Define $\hat{\tau}_{WATE}^g = \sum_{i=1}^N T_i \hat{p}_K(\mathbf{X}_i) Y_i - \sum_{i=1}^N (1-T_i)\hat{q}_K(\mathbf{X}_i) Y_i$. The following theorem states that $\hat{\tau}_{WATE}^g$ is efficient.

THEOREM 3. *Under Assumptions 1-8, $|g|$ is bounded from above and that $\mathbb{E}[g(\mathbf{X})] > 0$. Then we have*

*(a)* $\hat{\tau}_{WATE}^g \xrightarrow{p} \tau_{WATE}$;

*(b)* $\sqrt{N}\,(\hat{\tau}_{WATE}^g - \tau_{WATE}) \xrightarrow{d} \mathcal{N}(0, V_{semi}^g)$, *where*

$$V_{semi}^g \triangleq \frac{1}{\mathbb{E}[g(\mathbf{X})]^2}\mathbb{E}\left[g(\mathbf{X})^2 (m_1(\mathbf{X}) - m_0(\mathbf{X}) - \tau_{wate})^2 + \frac{\sigma_1^2(\mathbf{X}) g(\mathbf{X})^2}{\pi(\mathbf{X})} + \frac{\sigma_0^2(\mathbf{X}) g(\mathbf{X})^2}{1-\pi(\mathbf{X})}\right]$$ *attaining the semi-parametric efficiency bound as shown in Theorem 4 of Hirano et al. (2003).*

### 4.2. Treatment effects on the treated

To estimate the average treatment effects among the treated subpopulations, we estimate

$$\tau_{ATT} = \mathbb{E}(Y(1) - Y(0) | T = 1) = \frac{\int \mathbb{E}[Y(1) - Y(0) | \mathbf{X} = x]\pi(x) dF(x)}{\int \pi(x) dF(x)} ,$$

where the last equality follows from Assumption 1. Therefore, when the propensity score is known, $\tau_{ATT}$ is a special case of $\tau_{WATE}$ with $g(x) = \pi(x)$, and one can estimate $\tau_{ATT}$ by $\hat{\tau}^{\pi}_{WATE}$. Following Theorem 3, we have the following results for $\hat{\tau}^{\pi}_{WATE}$.

COROLLARY 4. *Under Assumptions 1-8, we have*

*(a)* $\hat{\tau}^{\pi}_{WATE} \xrightarrow{p} \tau_{ATT}$*;*
*(b)* $\sqrt{N}(\hat{\tau}^{\pi}_{WATE} - \tau_{ATT}) \xrightarrow{d} \mathcal{N}(0, V^{\pi}_{semi})$*, where*

$$V^{\pi}_{semi} \triangleq \frac{1}{\mathbb{E}\left[\pi(\mathbf{X})\right]^2}\mathbb{E}\left[\pi(\mathbf{X})^2\left(m_1(\mathbf{X}) - m_0(\mathbf{X}) - \tau_{ATT}\right)^2 + \sigma_1^2(\mathbf{X})\pi(\mathbf{X}) + \frac{\sigma_0^2(\mathbf{X})\pi(\mathbf{X})^2}{1-\pi(\mathbf{X})}\right]$$ *attaining the semi-parametric efficiency bound as shown in Theorem 2 of Hahn (1998).*

Note that $\mathbb{E}(Y(1)|T = 1)$ is estimated by $\sum_{i=1}^{N} T_i \hat{p}_K(\mathbf{X}_i) Y_i$ where $\hat{p}_K$ satisfies (18) with $g(x) = \pi(x)$, and this estimate is more efficient than the estimator $\sum_{i=1}^{N} T_i Y_i / \sum_{i=1}^{N} T_i$ when $\pi(x)$ is known, because one can calibrate the treated subpopulation to $\bar{u}^{\pi}_K$ and use the full data to improve estimation efficiency.

When $\pi(x)$ is unknown, however, the weighted average treatment effects estimator cannot be used. Since we want to estimate the subpopulation of the treated, it is natural to calibrate the control group to the treated by redefining the objective function:

$$\hat{H}_K(\beta) \triangleq \frac{1}{N}\sum_{i=1}^{N}\left[(1-T_i)\rho(\beta^T u_K(\mathbf{X}_i)) - \beta^T \bar{u}_{1K}\right] ,$$

where $\bar{u}_{1K} = \sum_{i=1}^{N} T_i u(\mathbf{X}_i) / \sum_{i=1}^{N} T_i$. The calibration estimator for estimating the treatment effects on the treated is $\hat{\tau}_{ATT} = \sum_{i=1}^{N} T_i Y_i / \sum_{i=1}^{N} T_i - \sum_{i=1}^{N}(1-T_i)\hat{q}(\mathbf{X}_i)Y_i$. The next theorem states that $\hat{\tau}_{ATT}$ is globally efficient when the propensity score is unknown.

THEOREM 5. *Under Assumptions 1 to 8, we have*

*(a)* $\hat{\tau}_{ATT} \xrightarrow{p} \tau_{ATT}$*;*
*(b)* $\sqrt{N}(\hat{\tau}_{ATT} - \tau_{ATT}) \xrightarrow{d} \mathcal{N}(0, V^{ATT}_{semi})$*, where*

$$V^{ATT}_{semi} \triangleq \frac{1}{\mathbb{E}\left[\pi(\mathbf{X})\right]^2}\mathbb{E}\left[\pi(\mathbf{X})\left(m_1(\mathbf{X}) - m_0(\mathbf{X}) - \tau_{ATT}\right)^2 + \sigma_1^2(\mathbf{X})\pi(\mathbf{X}) + \frac{\sigma_0^2(\mathbf{X})\pi(\mathbf{X})^2}{1-\pi(\mathbf{X})}\right]$$ *attaining the semi-parametric efficiency bound as shown in Theorem 1 of Hahn (1998).*

### 4.3. Multiple treatment groups

The calibration methods can also be easily generalized to situations with multiple treatment groups. Let $T_i \in \mathcal{J} = \{0, \ldots, J-1\}$ where $J \geq 2$ is an integer. Define $\mu_j = \mathbb{E}[Y(j)]$, $m_j(x) = \mathbb{E}[Y(j)|\mathbf{X} = x]$, $\pi_j(x) = \mathbb{P}\left(T = j|\mathbf{X} = x\right)$, $\sigma_j^2(x) = Var(Y(j)|\mathbf{X} = x)$, $j \in \mathcal{J}$, and $\tau_{jl} = \mu_j - \mu_l$ which is the average treatment effects between treatments $j$ and $l$. Calibration weights can be defined for any $j \in \mathcal{J}$ by

$$\hat{p}^j_K(\mathbf{X}_i) = \frac{1}{N}\rho'(\hat{\lambda}^j_K u_K(\mathbf{X}_i))$$

where $\hat{\lambda}^j_K \in \mathbb{R}^K$ maximizes the objective function:

$$\hat{G}^j_K(\lambda^j) = \frac{1}{N}\sum_{i=1}^{N}[I(T_i = j)\rho((\lambda^j)^T u_K(\mathbf{X}_i)) - \lambda^T \bar{u}_K], \quad j \in \mathcal{J} .$$

That is, we calibrate moments of $u_K(\mathbf{X})$ for each group to the full data. Estimators for $\mu_j, j \in \mathcal{J}$ are defined as

$$\hat{\mu}_j \triangleq \sum_{i=1}^{N} I(T_i = j)\hat{p}_K^j(\mathbf{X}_i)Y_i,\ j \in \mathcal{J}\ ,$$

and the estimator for the average treatment effects between treatment $j$ and $l$ is $\hat{\tau}_{jl} \triangleq \hat{\mu}_j - \hat{\mu}_l$.

THEOREM 6. *Under Assumptions 1 to 8 with $\pi(x)$ and $(m_1(x), m_0(x))$ replaced by $\pi_j(x)$ and $m_j(x)$ for $j \in \mathcal{J}$, respectively,*

*(a)* $\sum_{i=1}^{N}\Big\{I(T_i = j)\hat{p}_K^j(\mathbf{X}_i)Y_i - I(T_i = l)\hat{p}_K^l(\mathbf{X}_i)Y_i\Big\} \xrightarrow{p} \tau_{jl}$;

*(b)* $\sqrt{N}\left(\sum_{i=1}^{N}\{I(T_i = j)\hat{p}_K^j(\mathbf{X}_i)Y_i - I(T_i = l)\hat{p}_K^l(\mathbf{X}_i)Y_i\} - \tau_{jl}\right) \xrightarrow{d} \mathcal{N}(0, V_{jl})$, *where*

$$V_{jl} \triangleq \mathbb{E}\left[(m_j(\mathbf{X}) - m_l(\mathbf{X}) - \tau_{jl})^2 + \frac{\sigma_j^2(\mathbf{X})}{\pi_j(\mathbf{X})} + \frac{\sigma_l^2(\mathbf{X})}{\pi_l(\mathbf{X})}\right].$$

Estimators for the weighted average treatment effects and the treatment effects of the treated can also be easily extended to multiple treatment groups.

## 5. Simulations studies

In this section we present results from simulation studies to investigate the finite sample performance of various weighting estimators and standard error estimators.

The first set of simulation scenarios was similar to those in Kang and Schafer (2007) for the estimation of a population mean. Sample size for each simulated data set was 200 or 1000, and 10000 Monte Carlo datasets were generated for each scenario. For each observation, a random vector $Z = (Z_1, Z_2, Z_3, Z_4)$ was generated from the standard multivariate normal distribution. The potential outcome $Y(1)$ was generated from a normal distribution with mean $210+b(Z)$ and unit variance, where $b(Z) = 27.4Z_1+13.7Z_2+13.7Z_3+13.7Z_4$; $Y(0)$ was generated from a normal distribution with mean $200 - 0.5b(Z)$ and unit variance. An individual was assigned to treatment $T = 1$ with probability $\exp(\eta_0(Z))/(1+\exp(\eta_0(Z)))$, where $\eta_0(Z) = -Z_1+0.5Z_2-0.25Z_3 - 0.1Z_4$. The observed outcome was $Y = TY(1) + (1 - T)Y(0)$. Instead of observing covariates $Z$, we were only able to observe non-linear transformations $X_1 = \exp(Z_1/2), X_2 = Z_2/(1 + \exp(Z_1)), X_3 = (Z_1Z_3/25 + 0.6)^3$ and $X_4 = (Z_2 + Z_4 + 20)^2$. Denote $\mathbf{X} \triangleq (X_1, X_2, X_3, X_4)$. We compared the Horvitz-Thompson estimator (HT)

$$\hat{\tau}_{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_iY_i}{\hat{\pi}(\mathbf{X}_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-T_i)Y_i}{1-\hat{\pi}(\mathbf{X}_i)}\ ,$$

the ratio-type inverse probability weighting estimator (IPW)

$$\hat{\tau}_{IPW} = \frac{\sum_{i=1}^{N}T_i(\hat{\pi}(\mathbf{X}_i))^{-1}Y_i}{\sum_{i=1}^{N}T_i(\hat{\pi}(\mathbf{X}_i))^{-1}} - \frac{\sum_{i=1}^{N}(1-T_i)(1-\hat{\pi}(\mathbf{X}_i))^{-1}Y_i}{\sum_{i=1}^{N}(1-T_i)(1-\hat{\pi}(\mathbf{X}_i))^{-1}}\ ,$$

and the calibration estimators (CAL) with $\rho(v) = -e^{-v}$ (exponential tilting; ET), $\rho(v) = \log(1+v)$ (empirical likelihood; EL), $\rho(v) = -(1-v)^2/2$ (quadratic; Q) and $\rho(v) = v - e^{-v}$ (inverse logistic; IL). For HT and IPW estimators, we used a working logistic regression for propensity score model, and we considered six different ways to estimate the propensity score parameters: i. The maximum likelihood estimator ($\hat{\gamma}_{MLE}$); ii. The moment estimator solving the empirical counterpart of (11) that balances the covariates of the treated and the full data ($\hat{\gamma}_{1F}$); iii. The moment estimator solving the empirical counterpart of (12) that balances the covariates of the controls and the full data ($\hat{\gamma}_{0F}$); iv. The moment estimator solving the empirical counterpart of (13) that balances the covariates of the treated and the controls ($\hat{\gamma}_{10}$); v. The generalized method of moment estimator for the overidentified system (11) and (12) that balances the covariates of the treated, the controls, and the full data ($\hat{\gamma}_{10F}$); vi. The covariate balancing propensity score estimator of Imai and Ratkovic (2014) for an overidentified system defined by (13) and the likelihood score equation ($\hat{\gamma}_{CBPS}$). We presented the average bias and root mean squared error of the estimators over 10000 simulations, and the following standardized imbalance measures (Rosenbaum and Rubin; 1985; Imai and Ratkovic; 2014):

$$Imb_{10} = \left\{ \left( \frac{1}{N} \sum_{i=1}^{N} [(T_i w_{1i} - (1-T_i) w_{0i}) \mathbf{X}_i]^T \right) \left( \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^T \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^{N} (T_i w_{1i} - (1-T_i) w_{0i}) \mathbf{X}_i \right) \right\}^{1/2},$$

$$Imb_{1F} = \left\{ \left( \frac{1}{N} \sum_{i=1}^{N} [(T_i w_{1i} - 1) \mathbf{X}_i]^T \right) \left( \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^T \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^{N} (T_i w_{1i} - 1) \mathbf{X}_i \right) \right\}^{1/2},$$

and

$$Imb_{0F} = \left\{ \left( \frac{1}{N} \sum_{i=1}^{N} [1 - (1-T_i) w_{0i}) \mathbf{X}_i]^T \right) \left( \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^T \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^{N} [1 - (1-T_i) w_{0i}] \mathbf{X}_i \right) \right\}^{1/2},$$

where $w_{1i}$ are the weights for the treated, which is $\hat{\pi}_i^{-1}$ for HT; $(\hat{\pi}_i \times \sum_{j=1}^{N} T_j \hat{\pi}_j^{-1})^{-1}$ for IPW; and $\hat{p}_i$ for calibration. And $w_{0i}$ are the weights for the controls, which is $(1-\hat{\pi}_i)^{-1}$ for HT; $((1-\hat{\pi}_i) \times \sum_{j=1}^{N}(1-T_j)(1-\hat{\pi}_j)^{-1})^{-1}$ for IPW and $\hat{q}_i$ for calibration. Three imbalance measures were considered which measure three-way imbalance between the treated, the controls, and the full data. We examined all three measures because there is no guarantee that minimizing one imbalance measure could lead to a reduction in the others, as discussed in Section 2.4. A total imbalance measure is defined by $(Imb_{10}^2 + Imb_{1F}^2 + Imb_{0F}^2)^{1/2}$. Table 1 and 2 show the simulation results for $N = 200$ and $N = 1000$ respectively, using a linear specification of covariates $u_5 = (1, X_1, X_2, X_3, X_4)$ for both propensity score modeling and for calibration estimation.

The results showed that the Horvitz-Thompson estimators could worsen the problem of imbalance when the propensity score was estimated by maximum likelihood, or by method of moment that only balances a particular group to the full data. Balancing the treated to the full data only led to a noticeable improvement in estimating $\mathbb{E}(Y(1))$ but the estimator of $\mathbb{E}(Y(0))$ performed very poorly, as the imbalance between the controls and the full data actually increased. Therefore, the performance of the average treatment effects estimator was also very poor. Directly balancing the treated and the controls performed much better, and reduced the imbalance between the particular groups and the full data since the covariate distribution of the full data is a convex combination of the covariate distributions of the treated and controls. Generalized method of moment estimators ($\hat{\gamma}_{10F}, \hat{\gamma}_{CBPS}$) had more imbalance than balancing solely for the treated and controls, because the generalized method of moment estimates did not exactly solve the corresponding overidentified systems of equations. The calibration estimators achieved the exact three-way balance by design. In general, the mean squared errors of the estimates of average treatment effects were positively correlated with the overall imbalance. Finally, the exponential tilting estimator performed the best among the calibration estimators.

Next, we focused on the performance of weighting estimators under three sets of covariate specifications:

(a) $u_5(\mathbf{X}) = (1, X_1, X_2, X_3, X_4)$,
(b) $u_9(\mathbf{X}) = (1, X_1, X_2, X_3, X_4, X_1^2, X_2^2, X_3^2, X_4^2)$,
(c) $u_{15}(\mathbf{X}) = (1, X_1, X_2, X_3, X_4, X_1^2, X_2^2, X_3^2, X_4^2, X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, X_3X_4)$.

**Table 1.** Comparison of weighting estimators for the Kang and Schafer scenario, $n = 200$.

| | E(Y(1)) | | E(Y(0)) | | E(Y(1)-Y(0)) | | Imbalance*100 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Estimator | Bias | RMSE | Bias | RMSE | Bias | RMSE | (1,0) | (1,F) | (0,F) | Total |
| Unweighted | -10.09 | 10.68 | -5.06 | 5.34 | -5.03 | 6.38 | 85 | 43 | 42 | 104 |
| HT($\hat{\gamma}_{MLE}$) | 25.52 | 264.87 | -4.93 | 8.36 | 30.45 | 266.19 | 165 | 145 | 31 | 228 |
| HT($\hat{\gamma}_{1F}$) | -2.2 | 3.87 | 78.98 | $> 999$ | -81.17 | $> 999$ | 454 | 0 | 454 | 642 |
| HT($\hat{\gamma}_{0F}$) | $> 999$ | $> 999$ | -1.25 | 1.92 | $> 999$ | $> 999$ | $> 999$ | $> 999$ | 1 | $> 999$ |
| HT($\hat{\gamma}_{10}$) | 1.3 | 7.38 | 0.65 | 7.79 | 0.65 | 4.9 | 0 | 21 | 21 | 30 |
| HT($\hat{\gamma}_{10F}$) | -3.38 | 7.36 | -6.68 | 8.31 | 3.3 | 10.97 | 41 | 25 | 29 | 59 |
| HT($\hat{\gamma}_{CBPS}$) | -5.33 | 10.57 | -4.89 | 7.81 | -0.44 | 13.59 | 54 | 35 | 31 | 74 |
| IPW($\hat{\gamma}_{MLE}$) | 1.7 | 9.65 | -1.87 | 2.5 | 3.57 | 10.72 | 32 | 33 | 12 | 49 |
| IPW($\hat{\gamma}_{1F}$) | -2.17 | 3.86 | -1.38 | 3.65 | -0.79 | 4.96 | 33 | 0 | 34 | 47 |
| IPW($\hat{\gamma}_{0F}$) | 10.53 | 21.49 | -1.3 | 1.96 | 11.83 | 22.46 | 89 | 90 | 1 | 126 |
| IPW($\hat{\gamma}_{10}$) | -1.33 | 3.41 | -1.87 | 2.65 | 0.54 | 4.44 | 0 | 11 | 11 | 16 |
| IPW($\hat{\gamma}_{10F}$) | -2.41 | 4.06 | -2.14 | 2.73 | -0.27 | 4.43 | 11 | 12 | 11 | 20 |
| IPW($\hat{\gamma}_{CBPS}$) | -2.79 | 4.7 | -2.24 | 2.85 | -0.56 | 4.67 | 18 | 13 | 13 | 26 |
| CAL(ET) | -1.45 | 3.56 | -1.95 | 2.46 | 0.5 | 4.29 | 0 | 0 | 0 | 0 |
| CAL(EL) | -1.72 | 3.76 | -1.54 | 2.17 | -0.19 | 4.36 | 0 | 0 | 0 | 0 |
| CAL(Q) | -0.52 | 3.38 | -2.49 | 2.9 | 1.97 | 4.77 | 0 | 0 | 0 | 0 |
| CAL(IL) | -2.08 | 3.83 | -1.18 | 1.87 | -0.9 | 4.34 | 0 | 0 | 0 | 0 |

RMSE: root mean squared error; HT: Horvitz-Thompson estimators; IPW: ratio-type inverse probability weighting estimators; CAL: calibration estimators. For HT and IPW estimators, propensity score parameters were estimated in six ways: i. The maximum likelihood estimator ($\hat{\gamma}_{MLE}$); ii. The moment estimator that balances the covariates of the treated and the full data ($\hat{\gamma}_{1F}$); iii. The moment estimator that balances the covariates of the controls and the full data ($\hat{\gamma}_{0F}$); iv. The moment estimator that balances the covariates of the treated and the controls ($\hat{\gamma}_{10}$); v. The generalized method of moment estimator for an overidentified system that balances the covariates of the treated, the controls, and the full data ($\hat{\gamma}_{10F}$); vi. The covariate balancing propensity score estimator of Imai and Ratkovic (2014) ($\hat{\gamma}_{CBPS}$). For calibration estimators, ET: exponential tilting; EL: empirical likelihood; Q: quadratic; IL: inverse logistic.

**Table 2.** Comparison of weighting estimators for the Kang and Schafer scenario, $n = 1000$.

| | E(Y(1)) | | E(Y(0)) | | E(Y(1)-Y(0)) | | Imbalance*100 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Estimator | Bias | RMSE | Bias | RMSE | Bias | RMSE | (1,0) | (1,F) | (0,F) | Total |
| Unweighted | -10.04 | 10.16 | -5.01 | 5.07 | -5.03 | 5.32 | 82 | 41 | 41 | 101 |
| HT($\hat{\gamma}_{MLE}$) | 47.35 | 414.45 | -4.76 | 5.38 | 52.1 | 415.16 | 241 | 227 | 19 | 334 |
| HT($\hat{\gamma}_{1F}$) | -2.79 | 3.19 | -4.69 | 9.71 | 1.9 | 8.37 | 36 | 0 | 36 | 51 |
| HT($\hat{\gamma}_{0F}$) | 285.46 | $> 999$ | -1.11 | 1.28 | 286.57 | $> 999$ | $> 999$ | $> 999$ | 0 | $> 999$ |
| HT($\hat{\gamma}_{10}$) | -2.05 | 3.01 | -2.81 | 3.37 | 0.77 | 2.09 | 0 | 10 | 10 | 15 |
| HT($\hat{\gamma}_{10F}$) | 2.03 | 5.05 | -6.62 | 7.37 | 8.65 | 11.37 | 44 | 24 | 25 | 57 |
| HT($\hat{\gamma}_{CBPS}$) | 1.92 | 6.67 | -4.74 | 5.48 | 6.66 | 10.37 | 39 | 27 | 19 | 52 |
| IPW($\hat{\gamma}_{MLE}$) | 5.08 | 12.22 | -1.84 | 1.97 | 6.91 | 13.21 | 44 | 47 | 7 | 66 |
| IPW($\hat{\gamma}_{1F}$) | -2.79 | 3.19 | -2.51 | 2.69 | -0.27 | 1.9 | 13 | 0 | 13 | 19 |
| IPW($\hat{\gamma}_{0F}$) | 13.25 | 21.36 | -1.12 | 1.29 | 14.38 | 22.15 | 98 | 98 | 0 | 138 |
| IPW($\hat{\gamma}_{10}$) | -1.44 | 2.05 | -2.24 | 2.4 | 0.8 | 2.11 | 0 | 8 | 8 | 12 |
| IPW($\hat{\gamma}_{10F}$) | -1.25 | 2.14 | -2.32 | 2.45 | 1.07 | 2.5 | 7 | 11 | 9 | 16 |
| IPW($\hat{\gamma}_{CBPS}$) | -0.92 | 2.33 | -2.25 | 2.41 | 1.33 | 2.76 | 10 | 12 | 9 | 18 |
| CAL(ET) | -1.97 | 2.49 | -1.87 | 1.99 | -0.09 | 1.96 | 0 | 0 | 0 | 0 |
| CAL(EL) | -2.11 | 3 | -1.39 | 1.56 | -0.73 | 2.59 | 0 | 0 | 0 | 0 |
| CAL(Q) | -0.81 | 1.7 | -2.47 | 2.56 | 1.66 | 2.56 | 0 | 0 | 0 | 0 |
| CAL(IL) | -2.78 | 3.18 | -1.1 | 1.27 | -1.68 | 2.59 | 0 | 0 | 0 | 0 |

RMSE: root mean squared error; HT: Horvitz-Thompson estimators; IPW: ratio-type inverse probability weighting estimators; CAL: calibration estimators. For HT and IPW estimators, propensity score parameters were estimated in six ways: i. The maximum likelihood estimator ($\hat{\gamma}_{MLE}$); ii. The moment estimator that balances the covariates of the treated and the full data ($\hat{\gamma}_{1F}$); iii. The moment estimator that balances the covariates of the controls and the full data ($\hat{\gamma}_{0F}$); iv. The moment estimator that balances the covariates of the treated and the controls ($\hat{\gamma}_{10}$); v. The generalized method of moment estimator for an overidentified system that balances the covariates of the treated, the controls, and the full data ($\hat{\gamma}_{10F}$); vi. The covariate balancing propensity score estimator of Imai and Ratkovic (2014) ($\hat{\gamma}_{CBPS}$). For calibration estimators, ET: exponential tilting; EL: empirical likelihood; Q: quadratic; IL: inverse logistic.

**Table 3.** Comparisons of weighting estimators for various covariate configurations

| | | $u_5$ | | $u_9$ | | $u_{15}$ | | $(u_5, u_9)$ | $(u_5, u_{15})$ | $(u_9, u_{15})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Bias | RMSE | Bias | RMSE | Bias | RMSE | Corr | Corr | Corr |
| HT | N=200 | 30.45 | 266.19 | -1.41 | 19.29 | 3.41 | 20.47 | 0.04 | <0.01 | 0.37 |
| | N=1000 | 52.1 | 415.16 | -0.8 | 7.35 | 4.63 | 10.17 | 0.01 | <0.01 | 0.35 |
| IPW | N=200 | 3.57 | 10.72 | -0.10 | 4.73 | -0.16 | 5.04 | 0.45 | 0.32 | 0.79 |
| | N=1000 | 6.91 | 13.21 | 0.36 | 2.10 | 0.51 | 2.47 | 0.20 | 0.11 | 0.75 |
| CAL | N=200 | 0.5 | 4.29 | -1.39 | 4.46 | -0.66 | 4.50 | 0.82 | 0.70 | 0.79 |
| | N=1000 | -0.09 | 1.96 | -0.76 | 1.97 | -0.24 | 1.88 | 0.84 | 0.76 | 0.87 |

HT: Horvitz-Thompson estimator with propensity score estimated from maximum likelihood; IPW: ratio-type inverse probability weighting estimator with propensity score estimated from maximum likelihood; CAL: calibration estimator with exponential tilting.

In theory, global efficiency is attained when the number of moment conditions $K$ increases with the sample size $N$, but intuition and theory both suggest that $K$ cannot be too large. Hirano et al. (2003) and Imbens et al. (2006) both suggest using $K = N^{\nu}$ where $\nu < 1/9$ for the inverse probability weighting estimators, therefore the theory is in favor of $u_5$ for $N = 200$ and 1000. Since vigorous theories for nonparametric inverse probability weighting estimators have only been developed for maximum likelihood estimation, we limit our discussion here to estimators where the propensity score parameters are estimated by maximum likelihood. The results are shown in Table 3. While the existing theory suggested that $K$ should be small for the sample sizes considered, the performance of both HT and IPW estimators for $u_5$ was quite poor, while calibration estimations worked well, even for $u_5$. Propensity score weighting estimators performed better for $u_9$ and $u_{15}$, but the calibration estimators had uniformly smaller mean squared errors compared to that through the propensity score methods. We examined the stability of the estimation procedure by estimating the sample correlation of the same estimators under different covariate specifications. The correlations of HT estimators between $u_5$ and each of $u_9$ and $u_{15}$ were negligible, indicating that the inclusion of additional covariates could change the estimates arbitrarily. The correlations of the IPW estimators were also quite low, except for the correlation between $u_9$ and $u_{15}$. In contrast, the calibration estimators had high correlations under different covariate specifications. Therefore, adding higher order terms of covariates did not dramatically change the calibration estimate and its performance was minimally affected by the choice of $K$.

Next, we studied the performance of the proposed estimator for the efficient asymptotic variance $V_{semi}$, compared with a few other existing estimators. To describe the other estimators, we note that $V_{semi} = \mathbb{E}(\varphi_{eff}^2)$ where $\varphi_{eff}$ is the efficient influence function given in (17). The variance estimator of Hirano et al. (2003) is based on plugging in a propensity score estimate $\hat{\pi}(\mathbf{X})$ and a polynomial series estimate $\hat{\beta}(\mathbf{X})$. To estimate $\beta(\mathbf{X})$, they note that $\beta(\mathbf{X}) = \mathbb{E}(Y^*|\mathbf{X})$ where

$$Y^* = -\frac{TY}{\pi^2(\mathbf{X})} - \frac{(1-T)Y}{(1-\pi(\mathbf{X}))^2} \ .$$

Therefore, they calculate $Y^*$ by substituting $\pi(\mathbf{X})$ with a nonparametric estimate $\hat{\pi}(\mathbf{X})$, and $\beta(\mathbf{X})$ can be estimated by a linear regression of $Y^*$ on $u(\mathbf{X})$, which is the same design matrix as in the logistic regression model for $\pi(\mathbf{X})$. The estimator of Hirano et al. (2003) is

$$\hat{V}_{HIR} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{T_iY_i}{\hat{\pi}(\mathbf{X}_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(\mathbf{X}_i)} - \hat{\tau}^2 + (T_i - \pi(\mathbf{X}_i))\hat{\beta}(\mathbf{X}_i)\right]^2 \ .$$

An alternative plug-in estimator is to directly plug in $\hat{\pi}(\mathbf{X})$, $\hat{m}_1(\mathbf{X})$ and $\hat{m}_0(\mathbf{X})$ into the influence function, instead of estimating $\beta(\mathbf{X})$ which is a function of $(\pi(\mathbf{X}), m_1(\mathbf{X}), m_0(\mathbf{X}))$. The conditional expectations $m_1(\mathbf{X})$ and $m_0(\mathbf{X})$ can be estimated by polynomial series linear regression models of $Y$ on $u(\mathbf{X})$ among the treated and controls respectively. We compared the plug-in estimators and the proposed estimator in Table 4. The Hirano et al. estimator performed poorly for all covariate specifications and sample sizes, the direct plug-in estimator performed poorly in covariate specification $u_5$, and the proposed estimator performed well in all covariate specifications and sample sizes. A hypothesis for explaining the failure of the Hirano et al. estimator is that $Y^*$ depends on the squared inverse of propensity score, which is estimated very poorly and creates highly influential outlying values when the fitted propensity score is close to zero. To understand whether the problem is solely caused by poorly estimated propensity scores, we studied whether the performance of plug-in variance estimators can be improved when the propensity score was known. A known propensity score did not solve the entire problem for the Hirano et al. estimator because $\beta(\mathbf{X})$ is still highly nonlinear in $\mathbf{X}$ and the sieve estimator for $\beta(\mathbf{X})$ did not approximate $\beta(\mathbf{X})$ well enough in the given situations. To further understand the performance of estimators, we studied the correlations between the true and the estimated influence functions. For estimators that showed a good performance, the correlations were above 0.8 in general. The correlations were very low for the Hirano et al. estimator because the sieve estimator for $\beta(\mathbf{X})$ performed poorly. The direct plug-in estimator did not perform well for covariate specification $u_5$, because the propensity score was poorly estimated in that case.

**Table 4.** Comparisons of variance estimators, where the true asymptotic variance was 54.

| | | Covariate specifications | | | | | |
|---|---|---|---|---|---|---|---|
| | | $u_5$ | | $u_9$ | | $u_{15}$ | |
| Estimators | N | Estimate | Correlation | Estimate | Correlation | Estimate | Correlation |
| HIR(MLE) | 200 | >999 | 0.06 | 822 | 0.11 | 792 | 0.07 |
| | 1000 | >999 | -0.02 | >999 | 0.19 | >999 | 0.17 |
| HIR(True) | 200 | 511 | 0.12 | 888 | -0.16 | >999 | -0.16 |
| | 1000 | 328 | 0.19 | 532 | -0.16 | 732 | -0.15 |
| Plug-in(MLE) | 200 | 133 | 0.72 | 59 | 0.87 | 61 | 0.91 |
| | 1000 | 494 | 0.47 | 59 | 0.88 | 59 | 0.92 |
| Plug-in(True) | 200 | 74 | 0.85 | 64 | 0.87 | 68 | 0.91 |
| | 1000 | 71 | 0.80 | 65 | 0.85 | 61 | 0.92 |
| Proposed | 200 | 58 | 0.94 | 55 | 0.92 | 61 | 0.91 |
| | 1000 | 59 | 0.92 | 56 | 0.95 | 57 | 0.96 |

HIR: the estimator of Hirano, Imbens and Ridder (2003), MLE: Maximum likelihood estimator for propensity score; True: substituting in true propensity score and average treatment effects.

The results showed that $\hat{\varphi}_{CAL}$ and the true (but practically unknown) efficient influence function were highly correlated with correlation coefficients being greater than 0.9 for all simulation scenarios, and the estimated standard deviations are consistently close to the true asymptotic standard deviation. Further simulations (not included in the manuscript) showed that averages of the proposed variance estimates were very close to the sampling variances, and the coverage of confidence intervals based on normal approximations were close to the nominal levels.

We further considered an additional simulation scenario as in Hainmueller (2012). Six covariates $X_j, j = 1, \ldots, 6$ were generated, where $(X_1, X_2, X_3)$ were multivariate normal with means zero, variances of $(2, 1, 1)$, and the covariances of $X_1$ and $X_2$, $X_1$ and $X_3$, $X_2$ and $X_3$, were 1, -1 and -0.5 respectively; $X_4$ was uniformly distributed on (-3,3), $X_5$ was $\chi^2(1)$-distributed and $X_6$ was Bernoulli random variable with mean 0.5. The treatment indicator followed $T = I(X_1 + 2X_2 - 2X_3 - X_4 - 0.5X_5 + X_6 + \epsilon > 0)$ where $\epsilon \sim N(0, 30)$. This corresponds to the case with the largest imbalance between the treated and control groups in Hainmueller's simulation setting. The outcome distribution was $Y = (X_1 + X_2 + X_5)^2 + \eta$ where $\eta$ was the standard normal random variable. The outcome did not differ between the treated and controls, and the average treatment effects was zero. We compared the same set of estimators as in Table 1 and 2. We report the results for $n = 300$ with a linear covariate specification $(X_1, \ldots, X_6)$ in Table 5. Similar to Tables 1 and 2 for the Kang and Schafer scenario, the calibration estimators performed the best in terms of mean squared error and created an exact three-way covariate balance.

## 6. Data analysis

### *6.1. A childhood nutrition study*

We studied the performance of various weighting estimators using the 2007-2008 National Health and Nutrition Examination Survey (NHANES), which is a study designed to assess the health and nutrition statuses of children and adults in the United States. We studied whether participation of the National School Lunch or the School Breakfast programs (hereinafter, “school meal programs”) would lead to an increase in body mass index (BMI) for children and youths aged at 4 to 17. The school meal programs are intended to address the problem of insufficient food access of children in low-income families. However, a potential unintended consequence of the program is excessive food consumption which may cause childhood obesity. We analyzed 2330 children and youth at ages between 4 and 17, with a median age of 10, with 1284 (55%) participated in school meal programs.

Covariates in the data include: child age, child gender, race (black, Hispanic versus others), families above 200% of the federal poverty level, participation in Special Supplemental Nutrition Program for Women,

**Table 5.** Comparison of weighting estimators for the Hainmueller scenario, $n = 300$.

| | E(Y(1)-Y(0)) | | | Imbalance*100 | | |
|---|---|---|---|---|---|---|
| Estimator | Bias | RMSE | (1,0) | (1,F) | (0,F) | Total |
| Unweighted | 2.95 | 3.39 | 111 | 55 | 55 | 136 |
| HT($\gamma_{MLE}$) | 0.15 | 3.07 | 25 | 19 | 18 | 38 |
| HT($\gamma_{1F}$) | <-999 | > 999 | > 999 | 0 | > 999 | > 999 |
| HT($\gamma_{0F}$) | > 999 | > 999 | > 999 | > 999 | 0 | > 999 |
| HT($\gamma_{10}$) | 0.1 | 1.9 | 0 | 14 | 14 | 20 |
| HT($\gamma_{10F}$) | 0.34 | 1.81 | 17 | 14 | 13 | 26 |
| HT($\gamma_{CBPS}$) | 0.38 | 1.84 | 21 | 16 | 16 | 31 |
| IPW($\gamma_{MLE}$) | 0.21 | 2.34 | 23 | 18 | 17 | 35 |
| IPW($\gamma_{1F}$) | -1.96 | 8 | 59 | 0 | 59 | 84 |
| IPW($\gamma_{0F}$) | 0.5 | 5.76 | 61 | 61 | 0 | 86 |
| IPW($\gamma_{10}$) | 0.1 | 1.86 | 0 | 13 | 13 | 19 |
| IPW($\gamma_{10F}$) | 0.36 | 1.77 | 17 | 14 | 13 | 26 |
| IPW($\gamma_{CBPS}$) | 0.43 | 1.75 | 20 | 16 | 15 | 31 |
| CAL(ET) | 0.08 | 1.42 | 0 | 0 | 0 | 0 |
| CAL(EL) | 0.14 | 1.67 | 0 | 0 | 0 | 0 |
| CAL(Q) | 0.04 | 1.42 | 0 | 0 | 0 | 0 |
| CAL(IL) | 0.15 | 1.54 | 0 | 0 | 0 | 0 |

RMSE: root mean squared error; HT: Horvitz-Thompson estimators; IPW: ratio-type inverse probability weighting estimators; CAL: calibration estimators. For HT and IPW estimators, propensity score parameters were estimated in six ways: i. The maximum likelihood estimator ($\hat{\gamma}_{MLE}$); ii. The moment estimator that balances the covariates of the treated and the full data ($\hat{\gamma}_{1F}$); iii. The moment estimator that balances the covariates of the controls and the full data ($\hat{\gamma}_{0F}$); iv. The moment estimator that balances the covariates of the treated and the controls ($\hat{\gamma}_{10}$); v. The generalized method of moment estimator for an overidentified system that balances the covariates of the treated, controls and the full data ($\hat{\gamma}_{10F}$); vi. The covariate balancing propensity score estimator of Imai and Ratkovic (2014) ($\hat{\gamma}_{CBPS}$). For calibration estimators, ET: exponential tilting; EL: empirical likelihood; Q: quadratic; IL: inverse logistic.

**Table 6.** The difference in average BMI between participants and non-participants of school meal programs: NHANES 2007-2008 data

| | | | | Imbalance ×100 | | | |
|---|---|---|---|---|---|---|---|
| | Estimate | SE | 95% C. I. | (1,0) | (1,F) | (0,F) | Total |
| Naive | 0.53 | 0.21 | (0.11,0.95) | 103 | 47 | 57 | 127 |
| HT(MLE) | -1.48 | 0.51 | (-2.50,-0.46) | 49 | 16 | 57 | 63 |
| HT(CBPS) | -1.38 | 0.50 | (-2.38,-0.38) | 42 | 20 | 36 | 53 |
| IPW(MLE) | -0.14 | 0.24 | (-0.62,0.34) | 7 | 7 | 26 | 16 |
| IPW(CBPS) | -0.1 | 0.23 | (-0.56,0.36) | 10 | 11 | 12 | 17 |
| CAL(ET) | -0.04 | 0.22 | (-0.48,0.40) | 0 | 0 | 0 | 0 |
| CAL(EL) | -0.06 | 0.22 | (-0.50,0.38) | 0 | 0 | 0 | 0 |
| CAL(Q) | 0 | 0.22 | (-0.44,0.44) | 0 | 0 | 0 | 0 |
| CAL(IL) | -0.08 | 0.22 | (-0.52,0.36) | 0 | 0 | 0 | 0 |

SE: standard error; CI: confidence interval; HT: Horvitz-Thompson estimator; IPW: inverse probability weighting estimator; CAL: calibration estimator; MLE: maximum likelihood estimator; CBPS: covariate balancing propensity score; ET: exponential tilting; EL: empirical likelihood; Q: quadratic; IL: inverse logistic. Standardized covariate imbalance between treated and controls (10), treated and full data (1F), control and full data (0F) and total imbalance are reported.

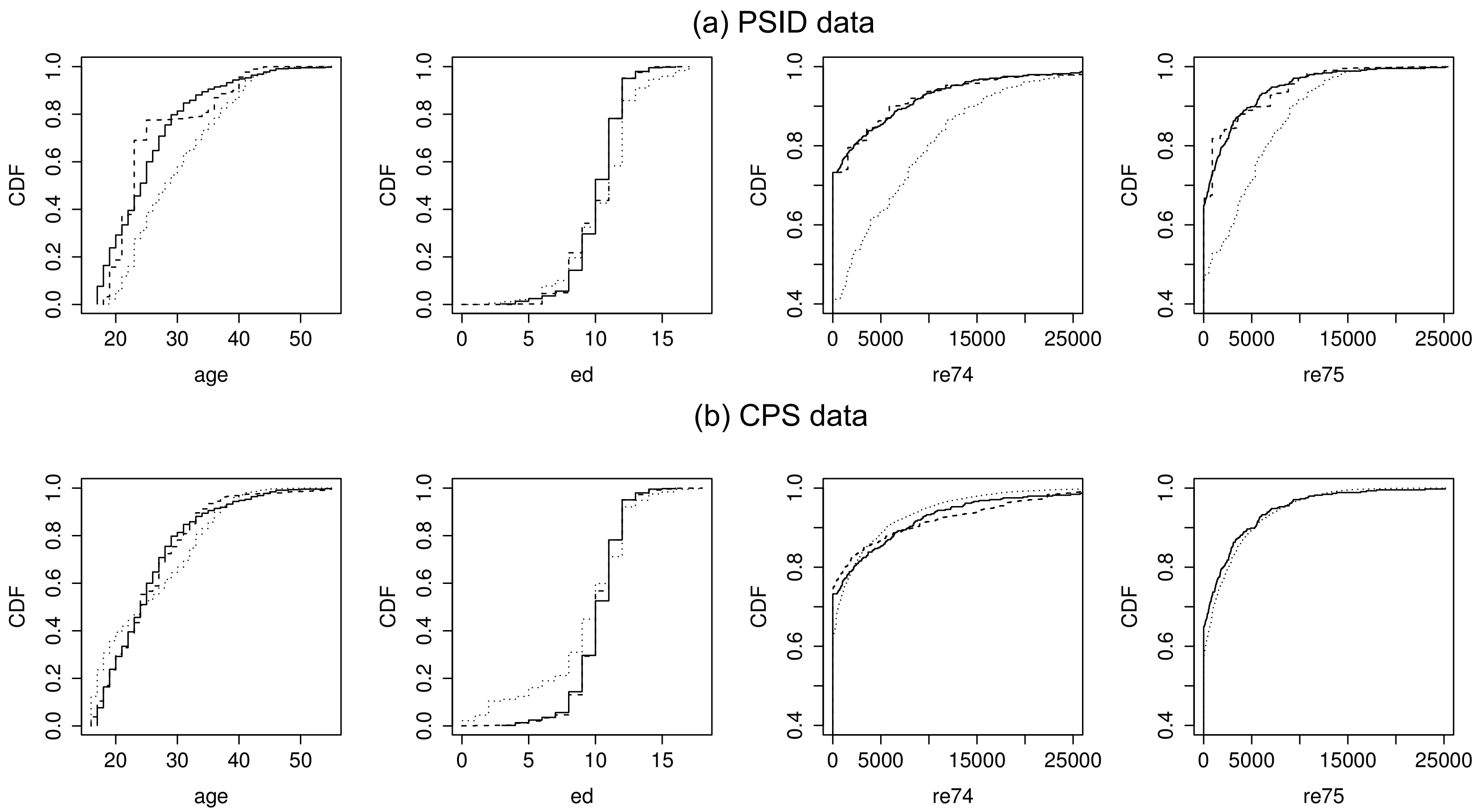
Infants and Children, participation in the Food Stamp Program, a childhood food security measurement which is an indicator of two or more affirmative responses to eight child-specific questions in the NHANES Food Security Questionnaire Module, health insurance coverage, and the age and sex of survey respondents (usually an adult in the family). The estimated average difference in BMI between participants and non-participants, together with standard error estimates and imbalance measures, are all given in Table 6. Direct comparison showed that the mean BMI of participants was significantly higher than that of non-participants, indicating that the program may lead to excessive food consumption. This particular finding has a policy implication in that there is a need to redesign the school meal programs to promote a healthier diet. Using a logistic propensity score model with a linear covariate specification, the Horvitz-Thompson estimators using maximum likelihood estimation and the covariate balancing propensity score estimates of Imai and Ratvokic (2014) both yielded a consistent, but opposite conclusion that the participation in school meal programs led to a significantly lower BMI and possible malnutrition. This particular finding has policy implications in that current school meal programs may fail in reducing the health disparities for poorer children. The inverse propensity score weighted estimates (IPW) were much closer to zero than the corresponding Horvitz-Thompson estimators, yielding a non-significant difference in BMI between the participants and the non-participants. The use of different propensity score weighting estimators leads to an inconclusive finding. The calibration estimators gave a consistent result that there is a negligible mean BMI difference between the participants and the non-participants. A policy implication is that the current school meal programs are implemented in an appropriate manner that provides assistance for the needed without an unintended consequence of increasing childhood obesity.

### 6.2. A job training study

We further demonstrate the performance of various weighting estimators by estimating the treatment impact of a labor training program data previously analyzed in Lalonde (1986) and Dehejia and Wahba (1999), among many others.

The National Supported Work (NSW) Demonstration was a randomized experiment implemented in the mid-1970s to study whether a systematic job training program would increase post-intervention income levels among workers. Both intervention and control groups were present in the original NSW study. Lalonde (1986) examined the extent to which analyses using observational data sets as controls would agree with the unbiased results of a randomized experiment. His nonexperimental estimates were based on two observational cohorts: the Panel Study of Income Dynamics (PSID) and Westat's Matched Current Population Survey –

**Fig. 1.** Weighted covariate distributions for the Lalonde (1986) data



(a) PSID data

(b) CPS data

Solid lines (———): Empirical distribution from the NSW data. Dashed lines ( - - - - - - ): Calibration weighted distributions. Dotted lines (··········): Inverse probability weighted distributions. CDF: Cumulative distribution function.

Social Security Administration File (CPS). Detailed description of the two data sets was given in Lalonde (1986) and Dehejia and Wahba (1999).

Dehejia and Wahba (1999) and others had analyzed the data set using different propensity score methodologies. Since calibration estimators are weighting estimators, we again limit our comparison only to other weighting estimators based on propensity score modeling. We combined the three data sets and created a group variable $G$ having four categories: $G = 1$ for the treated in the NSW data ($N = 185$), $G = 2$ for the controls in the NSW data ($N = 260$), $G = 3$ for the PSID data ($N = 2490$) and $G = 4$ for the CPS data ($N = 15992$). Categories 2, 3 and 4 all served as control data because individuals in those groups did not participate in the job training program offered in NSW. However, Categories 2, 3, 4 had substantially different covariate distributions. We considered the four categories in the combined data to illustrate that the calibration methodology can be applied to handle multiple treatment groups. We studied whether PSID and CPS can be used as the controls in the original NSW data, and we compared the estimates for the average treatment effects in the NSW study population, which was treatment effects on the treated. We noted that the treated and the control groups in the NSW data should have the same covariate distribution because of randomization. Using this information, we calibrated the weights of the four groups to the combined covariate distribution of groups 1 and 2. We also compared our results to calibration estimators with weights calibrated to the treatment group $G = 1$ only. Since there were four nominal group categories, we used the multinomial logit model for propensity score modeling. Two covariate configurations were considered, where the calibration estimators matched the same variables as in the propensity score models. The first (linear) specification included age, an indicator for black race (black), an indicator for Hispanic race (hisp), years of education (ed), an indicator for being married (married), an indicator for not having an academic degree (nodegr), income in 1974 (re74), income in 1975 (re75), an indicator for zero income in 1974 (u74), and an indicator for zero income in 1975 (u75). The second specification included all variables in the linear specification with the following additional higher-order variables: age$^2$, age$^3$, ed$^2$, re74$^2$, re75$^2$, ed×re74 and u74×black. These variables were included in the final models for either PSID or CPS data in Dehejia and Wahba (1999).

We estimated the average difference of income in 1978 between the treatment group ($G = 1$) and the control groups ($G = 2, 3, 4$). We further examined the evaluation bias defined as the estimated mean difference of outcome between the NSW control group ($G = 2$) and the observational control groups ($G = 3, 4$). The results are shown in Table 7. Direct comparisons were known to be severely biased, and had a huge evaluation bias. All weighting estimators greatly reduced the estimated evaluation bias, but the Horvitz-Thompson and inverse probability weighted estimator can yield very different estimates under the same model. Also, the estimates can change dramatically in comparing the two specifications of propensity score models. The calibration estimators yielded very similar estimates for both model specifications. Calibration to the combined NSW group had lower estimated standard errors compared to calibration only to the treated group. However, the estimates for the two calibration procedures were slightly different, probably because they were referring to two slightly different populations. This was possibly due to the fact that the analysis file of Dehejia and Wahba (1999) removed observations with a missing 1974 income, and the missingness may be different in the treatment and control groups. Hence, the NSW treated and controls may not have the same covariate distribution. In general, the standard errors of the propensity score estimators were much larger than that of the calibration estimators.

An advantage of using weighting estimators is that statisticians can graphically assess whether covariate balance is achieved. Figure 1 shows the weighted distributions of the four continuous covariates age, education (ed), income in 1974 (re74) and income in 1975 (re75). We compared the IPW and the calibration weighted distributions for the PSID and CPS data, with the empirical distributions of the combined NSW sample. Propensity score modeling and calibration were done using the non-linear model specification as in Dehejia and Wahba (1999). The calibration weighted distributions in both PSID and CPS data were close to the empirical distributions from the NSW data. However, the IPW weighted distributions showed a substantial difference for some variables, such as age, re74 and re75 in the PSID data and ed in the CPS data. Even when we matched a small number of moment constraints, the calibration weights performed well in matching the full covariate distributions between the non-experimental groups and the NSW data.

**Table 7.** Average treatment effects and evaluation biases for the Lalonde (1986) data.

(a) Treatment effects

| | | NSW data | | PSID data | | CPS data | |
|---|---|---|---|---|---|---|---|
| Estimators | Model | Estimates | SE | Estimates | SE | Estimates | SE |
| Unweighted | | 1794 | 671 | -15205 | 657 | -8498 | 583 |
| HT | 1 | 1600 | 827 | 2421 | 917 | 551 | 862 |
| HT | 2 | 1936 | 872 | 2136 | 1018 | 1071 | 711 |
| IPW | 1 | 1702 | 734 | 310 | 1095 | 1515 | 1014 |
| IPW | 2 | 1287 | 863 | -830 | 1207 | 1534 | 745 |
| CAL | 1 | 1572 | 667 | 2557 | 716 | 1233 | 668 |
| CAL* | 1 | 1712 | 707 | 2425 | 743 | 1406 | 675 |
| CAL | 2 | 1454 | 642 | 2504 | 699 | 1178 | 754 |
| CAL* | 2 | 1874 | 705 | 2285 | 795 | 1527 | 738 |

(b) Evaluation biases

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Unweighted | | | | -17000 | 391 | -10292 | 349 |
| HT | 1 | | | 626 | 809 | -1244 | 710 |
| HT | 2 | | | 342 | 998 | -724 | 917 |
| IPW | 1 | | | -1485 | 926 | -279 | 527 |
| IPW | 2 | | | 2624 | 1085 | -261 | 577 |
| CAL | 1 | | | 986 | 556 | -338 | 487 |
| CAL* | 1 | | | 712 | 600 | -306 | 512 |
| CAL | 2 | | | 1049 | 605 | -276 | 617 |
| CAL* | 2 | | | 411 | 677 | -347 | 592 |

SE: standard error; CI: confidence interval; HT: Horvitz-Thompson estimator with propensity score estimated from maximum likelihood; IPW: ratio-type inverse probability weighting estimator with propensity score estimated from maximum likelihood estimation; CAL: calibration with exponential tilting, moments are calibrated to the combined NSW group; CAL*: calibration with exponential tilting, moments are calibrated to the NSW treatment group. Two model specifications are considered, 1: linear in covariates; 2: linear and higher-order covariates as in Dehajia and Wahba (1999). Standard errors for HT and IPW estimators were calculated by bootstrapping with 1000 replicates.

**Table 8.** Comparisons of standard error estimates for the Lalonde (1986) data.

(a) Treatment Effects

| | NSW data | PSID data | CPS data |
|---|---|---|---|
| Proposed | 667 | 716 | 668 |
| Bootstrap | 672 | 737 | 666 |
| Fixed | 643 | 300 | 111 |

(b) Evaluation Biases

| | | | |
|---|---|---|---|
| Proposed | | 556 | 487 |
| Bootstrap | | 612 | 500 |
| Fixed | | 254 | 92 |

Bootstrap estimates were based on 1000 replicates.

We further examined the performance of the proposed standard error estimators by comparing them to bootstrap estimates based on 1000 replications and standard error estimates that treat the weights as given and fixed. The corresponding results for the calibration estimator under the linear specification are given in Table 8. The proposed standard error estimates were very close to the bootstrap standard errors, but the standard error estimators which treated the weights as if they were fixed can greatly underestimate the variability of the calibration estimators. For instance, when we ignored the variability of the estimated weights, the standard error estimates for the CPS data were much smaller because of their large sample sizes. However, the true estimation variability can be more than 6 times larger because the weights were constructed by calibrating to the combined NSW data which had a much smaller sample size. Therefore, ignoring estimation variability of calibration weights can lead to a misleading inference. For a correct inference, one cannot rely on variance estimates that treat the weights as fixed, which was suggested in Section 3.4 of Hainmueller (2012).

## 7. Discussions

We studied a large class of calibration estimators for efficient inference of the average treatment effects. Calibration weights removes imbalance in pretreatment covariates among the treated, controls and the combined group. By directly modifying the misspecified uniform weights, we do not directly model or estimate the propensity score. We show that balancing covariate distribution alone can achieve global semiparametric efficiency, and we also propose a consistent asymptotic variance estimator which outperforms other estimators that involve direct approximation of the influence function.

While we considered calibration estimators that modify the uniform weights, calibration can be constructed to modify the Horvitz-Thompson weights. However, this formulation requires an additional direct modeling and estimation of propensity score. If $\hat{\pi}(x)$ is an estimated propensity score function, a class of calibration weights can be defined by

$$\hat{p}_K(\mathbf{X}_i) = \frac{1}{N\hat{\pi}(\mathbf{X}_i)}\rho'\left(\hat{\lambda}^T u_K(\mathbf{X}_i)\right), \quad \text{for } i \text{ when } T_i = 1 \ ,$$

where $\hat{\lambda}_K \in \mathbb{R}^k$ maximizes the objective function

$$\hat{G}_K(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{T_i}{\hat{\pi}(\mathbf{X}_i)}\rho\left(\lambda^T u_K(\mathbf{X}_i)\right) - \lambda^T u_K(\mathbf{X}_i)\right] \ ;$$

and

$$\hat{q}_K(\mathbf{X}_i) = \frac{1}{N(1-\hat{\pi}(\mathbf{X}_i))}\rho'\left(\hat{\beta}^T u_K(\mathbf{X}_i)\right), \quad \text{for } i \text{ when } T_i = 0 \ ,$$

where $\hat{\beta}_K \in \mathbb{R}^K$ maximizes the objective function

$$\hat{H}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1-T_i}{1-\hat{\pi}(\mathbf{X}_i)}\rho\left(\beta^T u_K(\mathbf{X}_i)\right) - \beta^T u_K(\mathbf{X}_i)\right] \ .$$

This type of calibration procedure is discussed in detail in Chan and Yam (2014). We now focus on the special case that $K = 1$ and $u \triangleq u_K \equiv 1$, together with $\lambda$ and $\beta$ being set as scalar parameters. The balancing equation for the treated becomes

$$\frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\hat{\pi}(\mathbf{X}_i)}\rho(\hat{\lambda}) = 1 \ .$$

Therefore, $\rho(\hat{\lambda}) = (N^{-1} \times \sum_{i=1}^{N} T_i\hat{\pi}^{-1}(\mathbf{X}_i))^{-1}$ and the corresponding calibration weight is $\hat{p}_K(\mathbf{X}_i) = (\hat{\pi}(\mathbf{X}_i) \times \sum_{i=1}^{N} T_i\hat{\pi}^{-1}(\mathbf{X}_i))^{-1}$. Similarly, $\hat{q}_K(\mathbf{X}_i) = ((1-\hat{\pi}(\mathbf{X}_i)) \times \sum_{i=1}^{N}(1-T_i)(1-\hat{\pi}^{-1}(\mathbf{X}_i))^{-1})^{-1}$. This yields the ratio-type IPW estimator. Therefore, the construction of IPW from HT estimators can be viewed as a calibration procedure. When $\hat{\pi}$ is a series logit estimator, Hirano et al. (2003) and Imbens et al. (2006) show that the ratio-type IPW estimator is globally asymptotically efficient. Note that $u$ is one-dimensional regardless of sample sizes. Therefore, when $\pi$ is estimated by a series logit estimator, global efficiency can be achieved when $u_K$ has a fixed dimension. However, our global efficiency results require $u_K$ to have an increasing dimension with $N$ because propensity score was not directly estimated.

Our results show that a class of calibration weights yield globally efficient estimators which are first-order equivalent. An interesting extension is to study higher-order bias and optimality for the general class of estimators. Although the class of calibration estimators is related to generalized empirical likelihood, we cannot apply the higher-order theory of Newey and Smith (2004) for two reasons. First, moment conditions are assumed to be correctly specified in Newey and Smith (2004) while the moment conditions for calibration are misspecified for any finite $K$. While Schennach (2007) studied model misspecification, the number of moment conditions is fixed in both Newey and Smith (2004) and Schennach (2007), but is increasing for our calibration estimator. The latter problem is particularly thorny because there can be multiple first-order negligible terms that affect the higher-order bias in different ways. This challenging problem will be studied in a separate investigation.

In practice, we suggest to choose $u_K(\mathbf{X})$ to be the first and possibly higher moments of candidate covariates. When the covariate distributions of the treated and the controls differ only by a mean shift, matching the first moment of $\mathbf{X}$ would be sufficient for removing imbalance. When the variances differ, one can also match the second moment. Matching moments of covariate distributions are intuitive to non-statisticians. We can also graphically check whether the choice of $u_K(\mathbf{X})$ is sufficient as in Figure 1. A noticeable difference in the weighted distributions comparing the treated and the controls would suggest that additional moment conditions are needed. Furthermore, we can choose $K$ by a graphical method or by cross-validation. Since the parameter $K$ controls the number of moment conditions for matching to eliminate the bias from confounding, the choice of $K$ is therefore analogous to the selection of number of confounders in regression modeling, for which a graphical method was discussed in Crainiceanu et al. (2008). Inspired by that paper, we propose the following graphical method for choosing $K$. From the unweighted sample, we calculate the total standardized imbalance measure for each candidate $u_K(\mathbf{X})$, and rearrange them in a descending order of the imbalance measure. Then for $k = 1, 2, \ldots$, we plot the point estimates and the 95% confidence intervals by matching the first $k$ moment conditions. The bias would vanish when enough moment conditions are balanced, therefore the difference between consecutive point estimates will stabilize. Identify a region such that the the difference between consecutive point estimates is small, and within this region we can choose $K$ such that the corresponding confidence interval is the shortest. Alternatively, $K$ can be chosen by cross-validation. The moment conditions are matched in a training subset of the full data, while the weights are created in a complementary testing set and a total imbalance measure is calculated for the test data. This process is repeated over different partitions of data, and we choose $K$ such that the average imbalance measure is minimized. While $K$ can be chosen by the above methods, we would like to remark that Theorems 1 and 2 hold for a broad ranges of $K$ and our simulation results show that the performance of the calibration estimators are insensitive to the choice of $K$ when all relevant covariates are included. This is because our estimator involves the summation of linear functions of $\hat{p}_K(\mathbf{X}_i)$ and $\hat{q}_K(\mathbf{X}_i)$, and the summation of functions of nonparametric estimators are typically insensitive to the selection of the tuning parameter (Maity et al.; 2007), known as the double-smoothing phenomenon.

## References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects, *Econometrica* **74**(1): 235–267.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models, *Biometrics* **61**(4): 962–973.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates, *Journal of Human resources* **8**(4): 436–455.

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L. and Kulich, M. (2009). Improved Horvitz–Thompson Estimation of Model Parameters from Two-phase Stratified Samples: Applications in Epidemiology, *Statistics in Biosciences* **1**(1): 32–49.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data, *Biometrika* **96**(3): 723–734.

Cassel, C., Särndal, C. and Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations, *Biometrika* **63**(3): 615–620.

Chan, K. C. G. (2012). Uniform improvement of empirical likelihood for missing response problem, *Electronic Journal of Statistics* **6**: 289–302.

Chan, K. C. G. (2013). A simple multiply robust estimator for missing response problem, *Stat* **2**(1): 143–149.

Chan, K. C. G. and Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem, *Statist. Sci.* **29**(3): 380–396.

Chen, J. and Sitter, R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys, *Statistica Sinica* **9**: 385–406.

Chen, J., Sitter, R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, *Biometrika* **89**(1): 230–237.

Chen, S. X., Qin, J. and Tang, C. Y. (2013). Mann–whitney test with adjustments to pretreatment variables for missing values and observational study, *J. R. Statist. Soc. B* **75**(1): 81–102.

Chen, X., Hong, H. and Tarozzi, A. (2008). Semiparametric efficiency in gmm models with auxiliary data, *Ann. Statist.* **36**(2): 808–843.

Crainiceanu, C. M., Dominici, F. and Parmigiani, G. (2008). Adjustment uncertainty in effect estimation, *Biometrika* **95**(3): 635–651.

Dehejia, R. and Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs, *J. Am. Statist. Ass.* **94**(448): 1053–1062.

Deming, W. and Stephan, F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Ann. Math. Statist.* **11**(4): 427–444.

Deville, J. and Särndal, C. (1992). Calibration estimators in survey sampling, *J. Am. Statist. Ass.* **87**(418): 376–382.

Deville, J., Särndal, C. and Sautory, O. (1993). Generalized raking procedures in survey sampling, *J. Am. Statist. Ass.* **88**(423): 1013–1020.

Geman, S. and Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves, *Ann. Statist.* **10**(2): 401–414.

Graham, B., Pinto, C. and Egel, D. (2012). Inverse probability tilting for moment condition models with missing data, *The Review of Economic Studies* **79**(3): 1053–1079.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects, *Econometrica* **66**(2): 315–331.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies, *Political Analysis* **20**(1): 25–46.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data., *J. Am. Statist. Ass.* **109**(4): 1159–1173.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness, *Biometrika* **100**(2): 417–430.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators, *Econometrica* **50**: 1029–1054.

Hansen, L., Heaton, J. and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators, *Journal of Business & Economic Statistics* **14**(3): 262–280.

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

Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica* **71**(4): 1161–1189.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *J. Am. Statist. Ass.* **47**(260): 663–685.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score, *J. R. Statist. Soc. B* **76**(1): 243–263.

Imbens, G., Newey, W. and Ridder, G. (2006). Mean-squared-error calculations for average treatment effects, *Unpublished manuscript, University of California Berkeley* .

Imbens, G., Spady, R. and Johnson, P. (1998). Information theoretic approaches to inference in moment condition models, *Econometrica* **66**(2): 333–357.

Kang, J. and Schafer, J. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data, *Statist. Sci.* **22**(4): 523–539.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling, *International Statistical Review* **78**(1): 21–39.

Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation, *Econometrica* **65**(4): 861–874.

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs, *American Economic Review* **76**: 604–620.

Lumley, T., Shaw, P. A. and Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data, *International Statistical Review* **79**(2): 200–220.

Maity, A., Ma, Y. and Carroll, R. J. (2007). Efficient estimation of population-level summaries in general semiparametric regression models, *J. Am. Statist. Ass.* **102**(477): 123–139.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators, *Econometrica* **62**(6): 1349–1382.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics* **79**: 147–168.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators, *Econometrica* **72**(1): 219–255.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets, *International economic review* **14**(3): 693–709.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika* **75**(2): 237–249.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations, *Ann. Statist.* **22**: 300–325.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies, *J. R. Statist. Soc. B* **69**(1): 101–122.

Ridgeway, G. and McCaffrey, D. F. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, *Statist. Sci.* **22**(4): 540–543.

Robins, J., Rotnitzky, A. and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed, *J. Am. Statist. Ass.* **89**(427): 846–866.

Rosenbaum, P. R. (1987). Model-based direct adjustment, *J. Am. Statist. Ass.* **82**(398): 387–394.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies, *J. R. Statist. Soc. B (Methodological)* **53**(3): 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**(1): 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**(387): 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, *The American Statistician* **39**(1): 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**(1): 20–36.

Saegusa, T. and Wellner, J. A. (2013). Weighted likelihood estimation under two-phase sampling, *Ann. Statist.* **41**(1): 269–295.

Schennach, S. (2007). Point estimation with exponentially tilted empirical likelihood, *Ann. Statist.* **35**(2): 634–672.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting, *Biometrika* **97**(3): 661–682.

Tseng, P. and Bertsekas, D. P. (1987). Relaxation methods for problems with strictly convex separable costs and linear constraints, *Mathematical Programming* **38**(3): 303–321.

Vansteelandt, S., Bekaert, M. and Claeskens, G. (2012). On model selection and model misspecification in causal inference, *Statistical methods in medical research* **21**(1): 7–30.

Wu, C. and Sitter, R. (2001). A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data., *J. Am. Statist. Ass.* **96**(453): 185–193.

## A. Asymptotic expansion of the calibration estimators

The technical proofs for the lemmas and theorems are given in the supplementary materials. Here we present an asymptotic expansion of the empirical balancing estimator, which will be the key to these proofs.

Define

$$
\begin{aligned}
G_K^*(\lambda) &\triangleq \mathbb{E}[\pi(\mathbf{X})\rho(\lambda^T u_K(\mathbf{X})) - \lambda^T u_K(\mathbf{X})] = \mathbb{E}[\hat{G}_K(\lambda)] \ , \\
\lambda_K^* &\triangleq \arg\max_{\lambda\in\mathbb{R}^K} G_K^*(\lambda) \quad \text{and} \quad p_K^*(x) \triangleq \frac{1}{N}\rho'((\lambda_K^*)^T u_K(x)) \ , \\
H_K^*(\beta) &\triangleq \mathbb{E}[(1-\pi(\mathbf{X}))\rho(\beta^T u_K(\mathbf{X})) - \beta^T u_K(\mathbf{X})] = \mathbb{E}[\hat{H}_K(\beta)] \ , \\
\beta_K^* &\triangleq \arg\max_{\beta\in\mathbb{R}^K} H_K^*(\beta) \quad \text{and} \quad q_K^*(x) \triangleq \frac{1}{N}\rho'((\beta_K^*)^T u_K(x)) \ , \\
\Sigma_K &\triangleq (G_K^*)''(\lambda_K^*) = \mathbb{E}[\pi(\mathbf{X})\rho''((\lambda_K^*)^T u_K(\mathbf{X}))u_K(\mathbf{X})u_K(\mathbf{X})^T] \ , \\
\Psi_K &\triangleq -\mathbb{E}[m_1(\mathbf{X})\pi(\mathbf{X})\rho''((\lambda_K^*)^T u_K(\mathbf{X}))u_K(\mathbf{X})] \ , \\
Q_K(x) &\triangleq \Psi_K^T \Sigma_K^{-1} u_K(x) \ , \\
\tilde{\Pi}_K &\triangleq \hat{H}_K''(\tilde{\beta}_K) = \frac{1}{N}\sum_{i=1}^N (1-T_i)\rho''(\tilde{\beta}_K^T u_K(\mathbf{X}_i))u_K(\mathbf{X}_i)u_K(\mathbf{X}_i)^T \ , \\
\tilde{\Omega}_K &\triangleq -\int_{\mathcal{X}} m_0(x)(1-\pi(x))\rho''(\tilde{\beta}_K^T u_K(x))u_K(x)dF_0(x) \ , \\
\tilde{D}_K(x) &\triangleq \tilde{\Omega}_K^T \tilde{\Pi}_K^{-1} u_K(x) \ , \\
\Pi_K &\triangleq (H_K^*)''(\beta_K^*) = \mathbb{E}[(1-\pi(\mathbf{X}))\rho''((\beta_K^*)^T u_K(\mathbf{X}))u_K(\mathbf{X})u_K(\mathbf{X})^T] \ , \\
\Omega_K &\triangleq -\mathbb{E}[m_0(\mathbf{X})(1-\pi(\mathbf{X}))\rho''((\beta_K^*)^T u_K(\mathbf{X}))u_K(\mathbf{X})] \ , \\
D_K(x) &\triangleq \Omega_K^T \Pi_K^{-1} u_K(x).
\end{aligned}
$$

Now we have the following decomposition of our empirical balancing estimator in Theorem 1(b),

$$
\begin{aligned}
&\sqrt{N}\left(\sum_{i=1}^N \{T_i\hat{p}_K(\mathbf{X}_i)Y_i - (1-T_i)\hat{q}_K(\mathbf{X}_i)Y_i\} - \tau\right) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^N (\{NT_i\hat{p}_K(\mathbf{X}_i)Y_i - N(1-T_i)\hat{q}_K(\mathbf{X}_i)Y_i\} - \mu_1 + \mu_0) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^N \left\{ (N\hat{p}_K(\mathbf{X}_i) - Np_K^*(\mathbf{X}_i))\, T_iY_i - \int_{\mathcal{X}} m_1(x)\pi(x)(N\hat{p}_K(x) - Np_K^*(x))dF_0(x) \right\} \qquad (19) \\
&- \frac{1}{\sqrt{N}}\sum_{i=1}^N \left\{ (N\hat{q}_K(\mathbf{X}_i) - Nq_K^*(\mathbf{X}_i))\,(1-T_i)Y_i - \int_{\mathcal{X}} m_0(x)(1-\pi(x))(N\hat{q}_K(x) - Nq_K^*(x))dF_0(x) \right\} \qquad (20) \\
&+ \frac{1}{\sqrt{N}}\sum_{i=1}^N \left\{ \left(Np_K^*(\mathbf{X}_i) - \frac{1}{\pi(\mathbf{X}_i)}\right) T_iY_i - \mathbb{E}\left[m_1(\mathbf{X})\pi(\mathbf{X})\left(Np_K^*(\mathbf{X}) - \frac{1}{\pi(\mathbf{X})}\right)\right] \right\} \qquad (21) \\
&- \frac{1}{\sqrt{N}}\sum_{i=1}^N \left\{ \left(Nq_K^*(\mathbf{X}_i) - \frac{1}{1-\pi_0(\mathbf{X}_i)}\right)(1-T_i)Y_i - \mathbb{E}\left[m_0(\mathbf{X})(1-\pi(\mathbf{X}))\left(Nq_K^*(\mathbf{X}) - \frac{1}{1-\pi_0(\mathbf{X})}\right)\right] \right\} \\
&\qquad (22)
\end{aligned}
$$

$$+\sqrt{N}\mathbb{E}\left[m_1(\mathbf{X})\pi(\mathbf{X})\left(Np_K^*(\mathbf{X})-\frac{1}{\pi(\mathbf{X})}\right)\right] \tag{23}$$

$$-\sqrt{N}\mathbb{E}\left[m_0(\mathbf{X})(1-\pi(\mathbf{X}))\left(Nq_K^*(\mathbf{X})-\frac{1}{1-\pi_0(\mathbf{X})}\right)\right] \tag{24}$$

$$+\sqrt{N}\int_{\mathcal{X}}m_1(x)\pi(x)(N\hat{p}_K(x)-Np_K^*(x))dF_0(x)-\frac{1}{\sqrt{N}}\sum_{i=1}^N[T_i\rho'((\lambda_K^*)^Tu_K(\mathbf{X}_i))-1]\tilde{Q}_K(\mathbf{X}_i) \tag{25}$$

$$-\sqrt{N}\int_{\mathcal{X}}m_0(x)(1-\pi(x))(N\hat{q}_K(x)-Nq_K^*(x))dF_0(x)+\frac{1}{\sqrt{N}}\sum_{i=1}^N[(1-T_i)\rho'((\beta_K^*)^Tu_K(\mathbf{X}_i))-1]\tilde{D}_K(\mathbf{X}_i) \tag{26}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N[T_i\rho'((\lambda_K^*)^Tu_K(\mathbf{X}_i))-1](\tilde{Q}_K(\mathbf{X}_i)-Q_K(\mathbf{X}_i)) \tag{27}$$

$$-\frac{1}{\sqrt{N}}\sum_{i=1}^N[(1-T_i)\rho'((\beta_K^*)^Tu_K(\mathbf{X}_i))-1](\tilde{D}_K(\mathbf{X}_i)-D_K(\mathbf{X}_i)) \tag{28}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N\left\{[T_i\rho'((\lambda_K^*)^Tu_K(\mathbf{X}_i))-1]Q_K(\mathbf{X}_i)+\frac{m_1(\mathbf{X}_i)}{\pi(\mathbf{X}_i)}(T_i-\pi(\mathbf{X}_i))\right\} \tag{29}$$

$$-\frac{1}{\sqrt{N}}\sum_{i=1}^N\left\{[(1-T_i)\rho'((\beta_K^*)^Tu_K(\mathbf{X}_i))-1]D_K(\mathbf{X}_i)+\frac{m_0(\mathbf{X}_i)}{1-\pi(\mathbf{X}_i)}(\pi(\mathbf{X}_i)-T_i)\right\} \tag{30}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N\left(\frac{T_iY_i}{\pi(\mathbf{X}_i)}-\mu_1-\frac{m_1(\mathbf{X}_i)}{\pi(\mathbf{X}_i)}(T_i-\pi(\mathbf{X}_i))\right) \tag{31}$$

$$-\frac{1}{\sqrt{N}}\sum_{i=1}^N\left(\frac{(1-T_i)Y_i}{1-\pi(\mathbf{X}_i)}-\mu_0-\frac{m_0(\mathbf{X}_i)}{1-\pi(\mathbf{X}_i)}(\pi(\mathbf{X}_i)-T_i)\right) \tag{32}$$

Since $\sum_{i=1}^N T_i\hat{p}_K(\mathbf{X}_i)Y_i-\mu_1$ and $\sum_{i=1}^N(1-T_i)\hat{q}_K(\mathbf{X}_i)Y_i-\mu_0$ have a symmetric structure, we only need to consider the terms (19), (21), (23), (25), (27) and (29), and then apply the similar arguments to the terms (20), (22), (24), (26), (28) and (30). We shall show that the sum (31) + (32) behaves like an asymptotic normal random variable, and the remaining terms are all of order $o_p(1)$. A key challenge of the proof is to show the asymptotic order of (29) and (30), because they link all the unknown functions $(\pi(x),m_1(x),m_0(x))$ with the calibration weights and balancing moment conditions. This is overcome by using a novel weighted projection argument.

## B. Dual formulation of calibration estimators

We derive the dual of the constrained optimization problem (5) by using the methodology introduced in Tseng and Bertsekas (1987); the dual of (6) follows by a similar argument. Define $E_{K\times N}\triangleq(u_K(\mathbf{X}_1),\dots,u_K(\mathbf{X}_N))$, $s_i\triangleq 1-T_iNp_i, i=1,\dots,N$, $\mathbf{s}\triangleq(s_1,\dots,s_N)^T$ and $f(v)\triangleq D(1-v)$, then we can rewrite the problem (5) as

$$\min_{\mathbf{s}}\sum_{i=1}^N T_if(s_i)\quad\text{subject to}\quad E_{K\times N}\cdot\mathbf{s}=0\ .$$

For every $j\in\{1,\dots,N\}$, we define the conjugate convex function (Tseng and Bertsekas; 1987) of $T_jf(\cdot)$

to be

$$\begin{aligned} g_j(z_j) =& \sup_{s_j} \{z_j s_j - T_j f(s_j)\} = \sup_{p_j} \{-T_j N p_j z_j + z_j - T_j f(1 - T_j N p_j)\} \\ =& \sup_{p_j} \{-T_j N p_j z_j + z_j - T_j f(1 - N p_j)\} \\ =& - T_j N p_j^* z_j + z_j - T_j f(1 - N p_j^*) \ , \end{aligned}$$

where the third equality follows by noting that $Tf(1-TNp_j) = Tf(1-Np_j)$, and $p_j^*$ satisfies the first order condition:

$$-T_j z_j = -T_j f'(1 - N p_j^*) \Rightarrow p_j^* = \frac{1}{N}\left\{1 - (f')^{-1}(z_j)\right\} \ .$$

By defining $\rho(z) \triangleq f\left((f')^{-1}(z)\right) + z - z \cdot (f')^{-1}(z)$, then

$$g_j(z_j) = -T_j \rho(z_j) + z_j \ .$$

By Tseng and Bertsekas (1987), the dual problem of (5) is

$$\begin{aligned} \min_{\lambda} \sum_{j=1}^{N} g_j(\lambda^T E_j) =& - \max_{\lambda} \sum_{j=1}^{N} \left\{T_j \rho\left(\lambda^T u_K(\mathbf{X}_j)\right) - \lambda^T u_K(\mathbf{X}_j)\right\} \\ =& - \max_{\lambda} \hat{G}_K(\lambda) \ , \end{aligned}$$

where $E_j$ is the $j$-th column of $E_{K\times N}$, i,e., $E_j = u_K(\mathbf{X}_j)$.

Since $D$ is strictly convex, $f''(v) = D''(1-v)$, and hence $f$ is also strictly convex and $f'$ is strictly increasing. Note that

$$\rho(v) = f((f')^{-1}(v)) + v - v(f')^{-1}(v) \Leftrightarrow \rho\left(f'(v)\right) = f(v) + f'(v) - vf'(v) \ ;$$

differentiating with respect to $v$ both sides of the latter equation yields:

$$\rho'\left(f'(v)\right) f''(v) = f'(v) + f''(v) - f'(v) - vf''(v) = (1-v)f''(v) \ ,$$

which also implies

$$\rho'\left(f'(v)\right) = 1 - v \ ,$$

since $f'' > 0$. Further differentiating with respect to $v$ of the above equation, we get $\rho''\left(f'(v)\right) f''(v) = -1$, which implies

$$\rho''(v) = -\frac{1}{f''\left((f')^{-1}(v)\right)} < 0 \ .$$

By also working backward, the convexity of $D$ is equivalent to the concavity of $\rho$.