# Covariate Distribution Balance via Propensity Scores*

Pedro H. C. Sant'Anna† Xiaojun Song‡ Qi Xu§

Vanderbilt University Peking University Vanderbilt University

August 28, 2018

**Abstract**

The propensity score plays an important role in causal inference with observational data. Once the propensity score is available, one can use it to estimate a variety of causal effects in a unified setting. Despite this appeal, a main practical difficulty arises because the propensity score is usually unknown, has to be estimated, and extreme propensity score estimates can lead to distorted inference procedures. To address these limitations, this article proposes to estimate the propensity score by fully exploiting its covariate balancing property. We call the resulting estimator the integrated propensity score (IPS) as it is based on integrated moment conditions. In sharp contrast with other methods that balance only some specific moments of covariates, the IPS aims to balance *all* functions of covariates. Further, the IPS estimator is data-driven, does not rely on tuning parameters such as bandwidths, admits an asymptotic linear representation, and is $\sqrt{n}$-consistent and asymptotically normal. We derive the asymptotic properties of inverse probability weighted estimators for the average, distributional and quantile treatment effects based on the IPS, and illustrate their relative performance via Monte Carlo simulations and three empirical applications. An implementation of the proposed methods is provided in the new package `IPS` for `R`.

*We thank Harold Chiang, Yuya Sasaki, and Ping Yu for valuable comments.

†Department of Economics, Vanderbilt University. E-mail: pedro.h.santanna@vanderbilt.edu. Part of this article was written when I was vising the Cowles Foundation at Yale University, whose hospitality is gratefully acknowledged.

‡Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University. E-mail: sxj@gsm.pku.edu.cn.

§Department of Economics, Vanderbilt University. E-mail: qi.xu.1@vanderbilt.edu.

# 1 Introduction

Identifying and estimating the effect of a policy, treatment or intervention on an outcome of interest is one of the main goals in applied research. Although a randomized control trial is the gold standard to identify causal effects, many times its implementation is infeasible and researchers have to rely on observational data. In such settings, the propensity score (PS), which is defined as the probability of being treated given observed covariates, plays a prominent role. Statistical methods using the PS include matching, inverse probability weighting (IPW), regression, as well as combinations thereof; for review, see e.g. Imbens and Rubin (2015).

In order to use these methods in practice, one has to acknowledge that the PS is usually unknown and has to be estimated from the observed data. Given the moderate or high dimensionality of covariates, researchers are usually coerced to adopt a parametric model for the PS. A popular approach is to assume a linear logistic model, estimate the unknown parameters by maximum likelihood (ML), check if the resulting PS estimates balance specific moments of covariates, and in case they do not, refit the PS model including higher-order and interaction terms and repeat the procedure until covariate balancing is achieved, see e.g. Rosenbaum and Rubin (1984) and Dehejia and Wahba (2002). On top of involving *ad hoc* choices of model refinements, such model selection procedures may result in distorted inference about the parameters of interest, see e.g. Leeb and Pötscher (2005). An additional challenge faced by PS estimators based on ML is that the likelihood loss function does not take into account the covariate balancing property of the PS (Rosenbaum and Rubin, 1983), implying that PS estimates can be relatively close to zero or one, and the resulting causal estimator may perform poorly. Such a drawback is particularly important under model misspecification, see e.g. Kang and Schafer (2007).

In light of these practical issues, alternative procedures that closer resemble randomization have been proposed. For instance, Graham et al. (2012), Hainmueller (2012), Imai and Ratkovic (2014), and Zubizarreta (2015) propose alternative estimation procedures that attempt to directly balance covariates among treated, control and the combined sample. Although such methods usually lead to improved finite sample properties, they only aim to balance prespecified functions of covariates. However, the covariate balancing property of the PS is considerably more powerful as it implies balance not only for some moments but for *all* measurable, integrable functions of the covariates.

In this paper we propose an alternative method to estimate the PS that aims on fully exploiting its covariate balancing property. We call the resulting PS estimator the integrated propensity score (IPS). At a conceptual level, the IPS builds on the fact that the covariate balancing property of the PS is based on conditional moment restrictions, and as so, it can be equivalently characterized by an infinite, but tractable, number of unconditional moment restrictions, see e.g. Theorem 16.10 in Billingsley (1995). Upon such an observation, we consider Cramér-von Mises type distances between these infinite integrated balancing conditions and zero, and show that their minima are uniquely achieved at the true PS parameters. These results in turn suggest that we can naturally estimate the IPS unknown parameters within the minimum distance framework, see e.g. Dominguez and Lobato (2004) and Escanciano (2006*a*, 2018).

The IPS enjoys several appealing properties. First, the IPS estimation procedure guarantees that the unknown PS parameters are globally identified. This is in contrast to the generalized method of moments approach based on finitely many balancing conditions. Second, the IPS estimator is data-driven, does not rely on tuning parameters such as bandwidths, admits an asymptotic linear representation, and is $\sqrt{n}$-consistent and asymptotically normal under relatively weak regularity conditions. Third, the IPS does not rely on outcome data, and therefore can be used in a unified manner to estimate a variety of causal parameters under different identifying assumptions. We illustrate this flexibility by deriving the asymptotic properties of inverse probability weighted (IPW) estimators for average, distributional and quantile treatment effects based on the IPS. Finally, we emphasize that our proposed method is computationally simple and easy to use as currently implemented in the new package `IPS` for `R`.

The rest of this article is organized as follows. Section 2 introduces general framework of balancing weights and explains the estimation problem of the IPS. Section 3 presents the large sample properties of the IPS estimator. This section also discusses how one can use the IPS to estimate and make inference about average, distributional and quantile treatment effects under the unconfoundedness assumption. Section 4 illustrates the comparative performance of the proposed method through simulations, whereas Section 5 presents three empirical applications. Section 6 concludes. Proofs as well as additional computational results are reported in the supplemental Appendix.

# 2 Covariate Balancing via Propensity Score

## 2.1 Background

Let $D$ be a binary random variable that indicates participation in the program, i.e. $D = 1$ if the individual participates in the treatment and $D = 0$ otherwise. Define $Y(1)$ and $Y(0)$ as the potential outcomes under treatment and control, respectively. The realized outcome of interest is $Y = DY(1) + (1 - D)Y(0)$, and $\mathbf{X}$ is an observable $k \times 1$ vector of pre-treatment covariates. Denote the support of $\mathbf{X}$ by $\mathcal{X} \subset \mathbb{R}^k$ and the propensity score $p(\mathbf{x}) = \mathbb{P}(D = 1|\mathbf{X} = \mathbf{x})$. For $d \in \{0, 1\}$, denote the distribution and quantile of the potential outcome $Y(d)$ by $F_{Y(d)}(y) = \mathbb{P}(Y(d) \leq y)$, and $q_{Y(d)}(\tau) = \inf\left\{y : F_{Y(d)}(y) \geq \tau\right\}$, respectively, where $y \in \mathbb{R}$ and $\tau \in (0, 1)$. Henceforth, assume we have a random sample $\left\{(Y_i, D_i, \mathbf{X}_i')'\right\}_{i=1}^n$ from $(Y, D, \mathbf{X}')'$, where $n$ is the sample size, and all random variables are defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For a generic random variable $Z$, let $\mathbb{E}_n[Z] = n^{-1} \sum_{i=1}^n Z_i$.

The main goal in causal inference is to assess the effect of a treatment $D$ on the outcome of interest $Y$. Perhaps the most popular parameter of interest is the overall average treatment effect, $ATE = \mathbb{E}[Y(1) - Y(0)]$. Despite its popularity, the $ATE$ can mask important treatment effect heterogeneity across different subpopulations, see e.g. Abadie (2002) and Bitler et al. (2006). Thus, in order to uncover potential treatment effect heterogeneity, one usually focuses on different treatment effect parameters beyond the mean. Leading examples include the overall distributional treatment effect,

$DTE(y) = F_{Y(1)}(y) - F_{Y(0)}(y)$, and the overall quantile treatment effect, $QTE(\tau) = q_{Y(1)}(\tau) - q_{Y(0)}(\tau)$. Given that these causal parameters depend on potential outcomes which are not jointly observed for the same individual, one cannot directly rely on the analogy principle to identify and estimate such functionals.

A commonly used identification strategies in policy evaluation to bypass this difficulty is to assume that selection into treatment is solely based on observable characteristics, and that all individuals have a positive probability of being in either the treatment or the control group - the so-called strongly ignorable setup, see e.g. Rosenbaum and Rubin (1983). Formally, strong ignorability requires the following assumption:

**Assumption 1** *(a) Given* $\mathbf{X}$*,* $(Y(1), Y(0))$ *is jointly independent from* $D$*; and (b) for all* $\mathbf{x} \in \mathcal{X}$*,* $p(\mathbf{x})$ *is uniformly bounded away from zero and one.*

Rosenbaum (1987) shows that, under Assumption 1, the $ATE$ is identified by

$$ATE = \mathbb{E}\left[\left(\frac{D}{p(\mathbf{X})} - \frac{(1-D)}{1-p(\mathbf{X})}\right)Y\right].$$

Analogously, for $d \in \{0,1\}$, $F_{Y(d)}(y)$ is identified by

$$F_{Y(d)}(y) = \mathbb{E}\left[\frac{1\{D=d\}}{dp(\mathbf{X}) + (1-d)(1-p(\mathbf{X}))} 1\{Y \leq y\}\right],$$

implying that both $DTE(y)$ and $QTE(\tau)$ can also be written as functionals of the observed data; see e.g. Firpo (2007).

These identification results suggest that, if the PS is known, one can get consistent estimators by using the sample analogue of such estimands. For instance, one can estimate the $ATE$ using the Hájek (1971) type estimator

$$\widetilde{ATE}_n = \mathbb{E}_n\left[\left(w_{1,n}^{ps}(D,\mathbf{X}) - w_{0,n}^{ps}(D,\mathbf{X})\right)Y\right],$$

where

$$w_{1,n}^{ps}(D,\mathbf{X}) = \frac{D}{p(\mathbf{X})} \Big/ \mathbb{E}_n\left[\frac{D}{p(\mathbf{X})}\right],$$

$$w_{0,n}^{ps}(D,\mathbf{X}) = \frac{1-D}{1-p(\mathbf{X})} \Big/ \mathbb{E}_n\left[\frac{1-D}{1-p(\mathbf{X})}\right].$$

Estimators for $F_{Y(d)}(y)$, $d \in \{0,1\}$, and $DTE(y)$ are formed using an analogous strategy. For the $QTE(\tau)$, one can simply invert the estimator of $F_{Y(d)}(y)$ to estimate $q_{Y(d)}(\tau)$; see e.g. Firpo (2007). Of course, estimators for other treatment effect measures can also be formed using a similar strategy, see e.g. Firpo and Pinto (2016) and Li et al. (2018). Thus, the prominent role played by the PS in causal inference is evident.

In observational studies, however, the propensity score $p(\mathbf{X})$ is usually unknown, and has to be estimated. Given that $\mathbf{X}$ is usually of moderate or high dimensions, researchers routinely adopt a parametric approach. A popular choice among practitioners is to use the logistic model, where

$$p(\mathbf{X}) = p(\mathbf{X}; \boldsymbol{\beta}_0) = \frac{\exp(\mathbf{X}'\boldsymbol{\beta}_0)}{1 + \exp(\mathbf{X}'\boldsymbol{\beta}_0)},$$

with $\boldsymbol{\beta}_0 \in \Theta \subset \mathbb{R}^k$. Next, one usually proceeds to estimate $\boldsymbol{\beta}_0$ within the maximum likelihood paradigm, i.e.,

$$\hat{\boldsymbol{\beta}}_n^{mle} = \arg\max_{\boldsymbol{\beta}\in\Theta} \mathbb{E}_n \left[ D \ln(p(\mathbf{X}; \boldsymbol{\beta})) + (1-D)\ln(1 - p(\mathbf{X}; \boldsymbol{\beta}))\right],$$

and uses the resulting PS fitted values to construct different treatment effect estimators. Despite the popularity of this procedure, it has been shown that it can lead to significant instabilities under mild PS misspecifications, particularly when some PS estimates are relatively close to zero or one, see e.g. Kang and Schafer (2007).

In light of these challenges, alternative methods to estimate the PS have emerged. A particularly fruitful direction is to exploit the covariate balancing property of the PS, that is, to exploit the fact that, for every measurable, integrable function $f$ of the covariates $\mathbf{X}$,

$$\mathbb{E}\left[\frac{D}{p(\mathbf{X}; \boldsymbol{\beta}_0)} f(\mathbf{X})\right] = \mathbb{E}\left[\frac{1-D}{1 - p(\mathbf{X}; \boldsymbol{\beta}_0)} f(\mathbf{X})\right] = \mathbb{E}\left[f(\mathbf{X})\right] \tag{1}$$

for a unique value $\boldsymbol{\beta}_0 \in \Theta$. For example, Imai and Ratkovic (2014) propose estimating the PS parameters $\boldsymbol{\beta}_0$ within the generalized method of moments framework where, for a finite vector of user-chosen functions $f(\mathbf{X})$ (e.g. $f(\mathbf{X}) = \mathbf{X}$ ),

$$\mathbb{E}\left[\left(\frac{D}{p(\mathbf{X}; \boldsymbol{\beta}_0)} - \frac{1-D}{1 - p(\mathbf{X}; \boldsymbol{\beta}_0)}\right) f(\mathbf{X})\right] = 0. \tag{2}$$

Graham et al. (2012), on the other hand, propose estimating $\boldsymbol{\beta}_0$ as the solution to a globally concave programming problem such that

$$\mathbb{E}\left[\left(\frac{D}{p\left(\mathbf{X};\boldsymbol{\beta}_0\right)}-1\right)\mathbf{X}\right]=0.$$

Note that both procedures rely on choosing a finite number of functions $f\left(\mathbf{X}\right)$, though there is little to no theoretical guidance on how to choose such functions. This potential limitation also applies to the procedures in Hainmueller (2012) and Zubizarreta (2015), which instead of using PS reweighting, they use calibrated weights to balance specific moments of $\mathbf{X}$.

While estimators that balance low-order moments of covariates usually enjoy more attractive finite sample properties than those based on the ML paradigm, it is important to emphasize that the aforementioned proposals do not fully exploit the covariate balancing property (1). Furthermore, as emphasized by Dominguez and Lobato (2004), the global identification condition for $\boldsymbol{\beta}_0$ can fail when one adopts the generalized method of moment approach, and only attempts to balance finitely many covariate moments.

In this paper we aim to estimate the PS parameters $\boldsymbol{\beta}_0$ by taking advantage of all the information contained in (1). Our proposed estimators do not rely on tuning parameters such as bandwidth, do not consult the outcome data, and can be implemented in a data-driven manner. Our estimation procedure also guarantees that the unknown PS parameters are globally identified.

## 2.2 The Integrated Propensity Score

In this section we discuss how we operationalize our proposal. The crucial step is to reexpress the infinite number of covariate balancing conditions (1) into a more tractable set of moment restrictions, and then characterize $\boldsymbol{\beta}_0$ as the unique minimizer of a (population) minimum distance function. We then leverage on this characterization, and make use of the analogy principle to suggest a natural estimator for $\boldsymbol{\beta}_0$. In what follows, we present a step-by-step description of how we achieve this.

First, note that by using the definition of conditional expectation, (1) can be rewritten as

$$\mathbb{E}\left[\mathbf{h}\left(D,\mathbf{X};\boldsymbol{\beta}_0\right)\middle|\,\mathbf{X}\right]=0\ a.s., \tag{3}$$

where $\mathbf{h}\left(D,\mathbf{X};\boldsymbol{\beta}\right)=\left(h_1\left(D,\mathbf{X};\boldsymbol{\beta}\right),h_0\left(D,\mathbf{X};\boldsymbol{\beta}\right)\right)'$, $h_d\left(D,\mathbf{X};\boldsymbol{\beta}\right)=w_d^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)-1$, $d\in\{0,1\}$, and

$$\begin{aligned}
w_1^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right) &= \frac{D}{p\left(\mathbf{X};\boldsymbol{\beta}\right)}\Big/\,\mathbb{E}\left[\frac{D}{p\left(\mathbf{X};\boldsymbol{\beta}\right)}\right],\\
w_0^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right) &= \frac{1-D}{1-p\left(\mathbf{X};\boldsymbol{\beta}\right)}\Big/\,\mathbb{E}\left[\frac{1-D}{1-p\left(\mathbf{X};\boldsymbol{\beta}\right)}\right].
\end{aligned}$$

That is, one can express the covariate balancing conditions (1) in terms of stabilized conditional moment restrictions.

Next, by exploiting the "integrated conditional moment approach" commonly adopted in the specification testing literature (González-Manteiga and Crujeiras, 2013 contains a comprehensive review), one can reduce (3) to an infinite number of unconditional covariate balancing restrictions. That is, by appropriately choosing a space $\Pi$ and a parametric family of functions $\mathcal{W}=\{w(\mathbf{X};\mathbf{u}):\mathbf{u}\in\Pi\}$, one can equivalently characterize (1) as

$$\mathbb{E}\left[\mathbf{h}\left(D,\mathbf{X};\boldsymbol{\beta}_0\right)w(\mathbf{X};\mathbf{u})\right]=0\ a.e\ in\ \Pi, \tag{4}$$

see e.g. Lemma 1 of Escanciano (2006*b*) for primitive conditions on the family $\mathcal{W}$ such that the equivalence between (3) and (4) holds. Choices of $w$ satisfying this equivalence include $(a)$ $w(\mathbf{X};\mathbf{u})=1\left\{\mathbf{X}\leq\mathbf{u}\right\}$, where $\mathbf{u}\in[-\infty,\infty]^k$, $1\left\{A\right\}$ denotes the indicator of the event $A$ and $\mathbf{X}\leq\mathbf{u}$ is understood coordinate-wise (see e.g. Stute, 1997 and Dominguez and Lobato, 2004; Domínguez and Lobato, 2015), $(b)$ $w(\mathbf{X};\mathbf{u})=\exp(i\mathbf{u}'\Phi\left(\mathbf{X}\right))$, where $\mathbf{u}\in\mathbb{R}^k$, $\Phi\left(\cdot\right)$ is a vector of bounded one-to-one maps from $\mathbb{R}^k$ to $\mathbb{R}^k$ and $i=\sqrt{-1}$ is the imaginary unit (see e.g. Bierens, 1982, 1990 and Escanciano, 2018), and $(c)$ $w(\mathbf{X};\mathbf{u})=1\left\{\boldsymbol{\gamma}'\mathbf{X}\leq u\right\}$, with $\mathbf{u}=(\boldsymbol{\gamma},u)\in\mathbb{S}_k\times[-\infty,\infty]$, $\mathbb{S}_k=\left\{\boldsymbol{\gamma}\in\mathbb{R}^k:\|\boldsymbol{\gamma}\|=1\right\}$ (see e.g. Escanciano, 2006*a*). We call (4) the "integrated covariate balancing condition" because it uses the integrated (cumulative) measure of covariate balancing.

Finally, let

$$Q_w(\boldsymbol{\beta})=\int\left|\left|\mathbf{H}_w(\boldsymbol{\beta},\mathbf{u})\right|\right|^2\,\Psi(d\mathbf{u}),\quad \boldsymbol{\beta}\in\Theta\subset\mathbb{R}^k, \tag{5}$$

where $\mathbf{H}_w(\boldsymbol{\beta}, \mathbf{u}) = \mathbb{E}\left[\mathbf{h}\left(D, \mathbf{X}; \boldsymbol{\beta}\right) w(\mathbf{X}; \mathbf{u})\right]$, $\|A\|^2 = A^c A$, $A^c$ denotes the conjugate transpose of the column vector $A$, and $\Psi(\mathbf{u})$ is an integrating probability measure that is absolutely continuous with respect to a dominating measure on $\Pi$.

With these results in hand, in the following lemma we show that

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}\in\Theta} Q_w(\boldsymbol{\beta}) \tag{6}$$

and $\boldsymbol{\beta}_0$ is the unique value such that the covariate balancing condition (1) is satisfied.

**Lemma 1** *Let $\Theta \subset \mathbb{R}^k$ be the parameter space. Then $Q_w(\boldsymbol{\beta}) \geq 0$, $\forall \boldsymbol{\beta} \in \Theta$, and $Q_w(\boldsymbol{\beta}_0) = 0$ if and only if the the covariate balancing condition (1) holds.*

Lemma 1 is a global identification result that characterizes $\boldsymbol{\beta}_0$ as the unique minimizer of a population minimum distance function, $Q_w(\boldsymbol{\beta})$. Restated, from Lemma 1 we have that $\boldsymbol{\beta}_0$ is the unique PS parameter that minimizes the imbalances of all measurable, integrable functions $f(\cdot)$ between the treated, control and the combined group. This is in contrast with the generalized method of moments approach that only employs finitely many balancing conditions; see Dominguez and Lobato (2004).

Another important implication of Lemma 1 is that it suggests a natural estimator for $\boldsymbol{\beta}_0$ based on the sample analogue of (6), that is,

$$\widehat{\boldsymbol{\beta}}_{n,w}^{ips} = \arg\min_{\boldsymbol{\beta}\in\Theta} Q_{n,w}(\boldsymbol{\beta}), \tag{7}$$

where $Q_{n,w}(\boldsymbol{\beta}) = \int \|\mathbf{H}_{n,w}(\boldsymbol{\beta}, \mathbf{u})\|^2 \ \Psi_n(d\mathbf{u})$, $\Psi_n$ is a uniformly consistent estimator of $\Psi$, $\mathbf{H}_{n,w}(\boldsymbol{\beta}, \mathbf{u}) = \mathbb{E}_n\left[\mathbf{h}_n\left(D, \mathbf{X}; \boldsymbol{\beta}\right) w(\mathbf{X}; \mathbf{u})\right]$, with $\mathbf{h}_n\left(D, \mathbf{X}; \boldsymbol{\beta}\right) = \left(h_{n,1}\left(D, \mathbf{X}; \boldsymbol{\beta}\right), h_{n,0}\left(D, \mathbf{X}; \boldsymbol{\beta}\right)\right)'$, $h_{n,d}\left(D, \mathbf{X}; \boldsymbol{\beta}\right) = w_{n,d}^{ps}\left(D, \mathbf{X}; \boldsymbol{\beta}\right) - 1$, $d \in \{0, 1\}$, and

$$w_{n,1}^{ps}\left(D, \mathbf{X}; \boldsymbol{\beta}\right) = \frac{D}{p\left(\mathbf{X}; \boldsymbol{\beta}\right)} \Bigg/ \mathbb{E}_n\left[\frac{D}{p\left(\mathbf{X}; \boldsymbol{\beta}\right)}\right], \tag{8}$$

$$w_{n,0}^{ps}\left(D, \mathbf{X}; \boldsymbol{\beta}\right) = \frac{1-D}{1-p\left(\mathbf{X}; \boldsymbol{\beta}\right)} \Bigg/ \mathbb{E}_n\left[\frac{1-D}{1-p\left(\mathbf{X}; \boldsymbol{\beta}\right)}\right]. \tag{9}$$

We call $\widehat{\boldsymbol{\beta}}_{n,w}^{ips}$ the integrated propensity score estimator of $\boldsymbol{\beta}_0$ since it is based on the integrated covariate balancing conditions (4).

From (7) one can conclude that different PS estimators that fully exploit the covariate balancing property (1) can be constructed by choosing different $w$ and $\Psi_n$. In this article we focus on three different combinations that are intuitive, computationally simple, and that perform well in practice:

(*i*) $w(\mathbf{X};\mathbf{u}) = 1\{\mathbf{X} \leq \mathbf{u}\}$ and $\Psi_n(\mathbf{u}) = F_{n,\mathbf{X}}(\mathbf{u}) \equiv n^{-1}\sum_{i=1}^{n} 1(\mathbf{X}_i \leq \mathbf{u})$, leading to the IPS estimator

$$\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips} = \arg\min_{\boldsymbol{\beta}\in\Theta} \int_{[-\infty,\infty]^k} \left\| \mathbb{E}_n \left[ \mathbf{h}_n\left(D, \mathbf{X}; \boldsymbol{\beta}\right) 1(\mathbf{X} \leq \mathbf{u}) \right] \right\|^2 F_{n,X}(d\mathbf{u}); \qquad (10)$$

(*ii*) $w(\mathbf{X};\mathbf{u}) = 1\{\boldsymbol{\gamma}'\mathbf{X} \leq u\}$ with $\Psi_n(\mathbf{u})$ the product measure of $F_{n,\boldsymbol{\gamma}}(u) \equiv n^{-1}\sum_{i=1}^{n} 1(\boldsymbol{\gamma}'\mathbf{X}_i \leq u)$ and the uniform distribution on $\mathbb{S}_k$, leading to the IPS estimator

$$\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips} = \arg\min_{\boldsymbol{\beta}\in\Theta} \int_{[-\infty,\infty]\times\mathbb{S}_k} \left\| \mathbb{E}_n \left[ \mathbf{h}_n\left(D, \mathbf{X}; \boldsymbol{\beta}\right) 1(\boldsymbol{\gamma}'\mathbf{X} \leq u) \right] \right\|^2 F_{n,\boldsymbol{\gamma}}(du)d\boldsymbol{\gamma}\ ; \qquad (11)$$

(*iii*) $w(\mathbf{X};\mathbf{u}) = \exp(i\mathbf{u}'\Phi\left(\mathbf{X}\right))$ with $\Psi_n(\mathbf{u}) \equiv \Psi(\mathbf{u})$, the CDF of $k$-variate standard normal distribution, $\Phi\left(\mathbf{X}\right) = \left(\Phi\left(\widetilde{X}_1\right), \ldots, \Phi\left(\widetilde{X}_k\right)\right)'$, $\widetilde{X}_p$ the studentized $X_p$, and $\Phi$ the univariate CDF of the standard normal distribution, leading to the IPS estimator

$$\widehat{\boldsymbol{\beta}}_{n,\exp}^{ips} = \arg\min_{\boldsymbol{\beta}\in\Theta} \int_{\mathbb{R}^k} \left\| \mathbb{E}_n \left[ \mathbf{h}_n\left(D, \mathbf{X}; \boldsymbol{\beta}\right) \exp(i\mathbf{u}'\Phi\left(\mathbf{X}\right)) \right] \right\|^2 \frac{\exp\left(-\frac{1}{2}\mathbf{u}'\mathbf{u}\right)}{\left(2\pi\right)^{k/2}} d\mathbf{u}. \qquad (12)$$

The estimators (10)-(12) build on Dominguez and Lobato (2004) and Escanciano (2006*a*, 2018), respectively. Despite the apparent differences, they all aim to minimize covariate distribution imbalances: (10) aims to directly minimize imbalances of the joint distribution of covariates; (11) exploits the Cramér-Wold theorem and focuses on minimizing imbalances of the distribution of all one-dimensional projections of covariates; and (12) focuses on the imbalances of the (transformed) covariates' joint characteristic function. From the Cramér-Wold theorem and the fact that the characteristic function completely defines the distribution function (and vice-versa), (10)-(12) are indeed intrinsically related. Furthermore, we emphasize that our estimators are data-driven, and neither $w$ nor $\Psi$ plays the role of a bandwidth since they do not affect the convergence rate of the IPS estimator.

From the computational perspective, (10)-(12) are easy to estimate since they do not involve matrix inversion nor nonparametric estimation. In the supplemental Appendix A we show that the objective functions in (10)-(12) can be written in closed form, which in turn implies a more straightforward implementation. In practice the IPS is easy to use as it is already implemented in the new package `IPS` for `R`. At the present time, `IPS` is available upon request, and soon it will be publicly available through CRAN and GitHub.

**Remark 1** It is important to stress that the covariate balancing property (1) follows directly from the definition of the PS and does not depend on the unconfoundedness assumption 1. Thus, one can use our proposed IPS estimators even in contexts where Assumption 1 does not hold, though in such cases the resulting (second step) estimators may be only descriptive, see e.g. DiNardo et al. (1996), Barsky et al. (2002), and Kline (2011).

**Remark 2** It is interesting to compare (2) with (4) beyond the fact that (4) is based on infinitely many balancing conditions whereas (2) is not. First, note that (4) is based on normalized (or stabilized) weights, whereas (2) is not. We prefer to use stabilized weights since treatment effect estimators based on them usually have improved finite sample properties (see e.g. Busso et al., 2014). Second, note that (4) implies three-way balance (treated, control and combined groups), whereas (2) only imposes a two-way balance (treated and control). We note that (2) can lead to relatively smaller/larger PS estimates since a "small" PS estimate in the treated group can be offset by a "large" PS estimate in the control group. By using (4) such a potential drawback is avoided.

**Remark 3** Other authors have also aimed to fully explore (1) using different procedures. For instance, Chan et al. (2016) propose a doubly robust approach to estimate the $ATE$ such that the number of balancing moments increases with sample size at a certain rate. Wong and Chan (2018) propose a related balancing procedure via reproducing kernel Hilbert space modelling. It is worth mentioning that these approaches involve choosing tuning parameters, though their sensitivity to these choices is still unknown. The IPS does not rely on tuning parameters, but imposes a parametric specification on the PS.

Another difference between our proposal and Chan et al. (2016) and Wong and Chan (2018) is that their estimators rely on assumptions about the outcome model whereas the IPS does not. Thus, by separating the design stage from the analysis stage, our proposal is in line with Rubin (2007, 2008) recommendations to achieve an objective design of observational studies for causal effects. Another advantage of not relying on outcome data is that we can consider a variety of causal parameters in a unified setup, whereas Chan et al. (2016) and Wong and Chan (2018) focus only the $ATE$ and the $ATT$.

# 3 Large Sample Properties

In this section, we first derive the asymptotic properties of the IPS estimators, namely the consistency, asymptotic linear representation, and asymptotic normality of $\widehat{\boldsymbol{\beta}}_{n,w}^{ips}$. We then discuss how one can build on these results to conduct asymptotically valid inference for overall average, distributional and quantile treatment effects, using inverse probability weighted estimators. Finally, we also present results for IPW estimators of the average, distributional and quantile treatment effects on the treated subpopulation. The difference between overall treatment effects and treatment effects on the treated is that the former is the treatment effect for the whole population under consideration and the latter is the treatment effect only for those individuals subject to treatment. Although our proposal can also be used to estimate other treatment effects of interest such as those discussed in Firpo and Pinto (2016) and Li et al. (2018), we omit such a discussion for the sake of brevity.

## 3.1 Asymptotic Theory for IPS estimator

Here we derive the asymptotic properties of the IPS estimator. Let the score of $\mathbf{H}_w(\boldsymbol{\beta}, \mathbf{u})$ be defined as $\dot{\mathbf{H}}_w(\boldsymbol{\beta}, \mathbf{u}) = \left(\dot{\mathbf{H}}_{1,w}'(\boldsymbol{\beta}, \mathbf{u}), \dot{\mathbf{H}}_{0,w}'(\boldsymbol{\beta}, \mathbf{u})\right)'$ where, for $d \in \{0, 1\}$, $\dot{\mathbf{H}}_{d,w}(\boldsymbol{\beta}, \mathbf{u}) =$

$\mathbb{E}\left[\dot{\mathbf{h}}_d\left(D,\mathbf{X};\boldsymbol{\beta}\right)w(\mathbf{X};\mathbf{u})\right]$, with

$$\dot{\mathbf{h}}_1\left(D,\mathbf{X};\boldsymbol{\beta}\right)=-\frac{w_1^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)}{p\left(\mathbf{X};\boldsymbol{\beta}\right)}\dot{p}\left(\mathbf{X};\boldsymbol{\beta}\right)'+w_1^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)\cdot\mathbb{E}\left[\frac{w_1^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)}{p\left(\mathbf{X};\boldsymbol{\beta}\right)}\dot{p}\left(\mathbf{X};\boldsymbol{\beta}\right)'\right],$$

$$\dot{\mathbf{h}}_0\left(D,\mathbf{X};\boldsymbol{\beta}\right)=\frac{w_0^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)}{1-p\left(\mathbf{X};\boldsymbol{\beta}\right)}\dot{p}\left(\mathbf{X};\boldsymbol{\beta}\right)'-w_0^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)\cdot\mathbb{E}\left[\frac{w_0^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)}{1-p\left(\mathbf{X};\boldsymbol{\beta}\right)}\dot{p}\left(\mathbf{X};\boldsymbol{\beta}\right)'\right],$$

and $\dot{p}\left(\cdot;\boldsymbol{\beta}\right)=\partial p\left(\cdot;\boldsymbol{b}\right)/\left.\partial\boldsymbol{b}\right|_{\boldsymbol{b}=\boldsymbol{\beta}}$. We make the following set of assumptions.

**Assumption 2** *$(i)$ $p\left(\mathbf{x}\right)=p\left(\mathbf{x};\boldsymbol{\beta}_0\right)$, where $\boldsymbol{\beta}_0$ is an interior point of a compact set $\Theta\subset\mathbb{R}^k$; $(ii)$ for some $0<\delta<0.5$, $\delta\leq p\left(\mathbf{x};\boldsymbol{\beta}\right)\leq1-\delta$ for all $\mathbf{x}\in\mathcal{X}$, $\boldsymbol{\beta}\in int\left(\Theta\right)$; $(iii)$ with probability one, $p\left(\mathbf{x};\boldsymbol{\beta}\right)$ is continuous at each $\boldsymbol{\beta}\in\Theta$; $(iv)$ with probability one, $p\left(\mathbf{x};\boldsymbol{\beta}\right)$ is once continuously differentiable in a neighborhood $\Theta_0\subset\Theta$ of $\boldsymbol{\beta}$; $(v)$ for $d\in\{0,1\}$*

$$\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in\Theta_0}\left\|\left(\frac{w_d^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)}{d\cdot p\left(\mathbf{X};\boldsymbol{\beta}\right)+(1-d)\cdot\left(1-p\left(\mathbf{X};\boldsymbol{\beta}\right)\right)}\right)\cdot\dot{p}\left(\mathbf{X};\boldsymbol{\beta}\right)\right\|\right]<\infty.$$

**Assumption 3** *The family of weighting functions and integrating probability measures satisfy one of the following:*

*$(i)$ $\mathcal{W}_{ind}\equiv\left\{\mathbf{x}\in\mathcal{X}\mapsto1(\mathbf{x}\leq\mathbf{u}):\mathbf{u}\in[-\infty,\infty]^k\right\}$, $\Psi_n(\mathbf{u})=F_{n,\mathbf{X}}(\mathbf{u})$, and $\Psi(\mathbf{u})=F_{\mathbf{X}}(\mathbf{u})$, where $F_{n,\mathbf{X}}(\mathbf{u})\equiv n^{-1}\sum_{i=1}^{n}1(\mathbf{X}_i\leq\mathbf{u})$, and $F_{\mathbf{X}}(\mathbf{u})\equiv\mathbb{E}\left[1(\mathbf{X}\leq\mathbf{u})\right]$;*

*$(ii)$ $\mathcal{W}_{proj}\equiv\left\{\mathbf{x}\in\mathcal{X}\mapsto1\left\{\boldsymbol{\gamma}'\mathbf{x}\leq u\right\}:(\boldsymbol{\gamma},u)\in\mathbb{S}_k\times[-\infty,\infty]\right\}$, $\Psi_n\left(\mathbf{u}\right)=F_{n,\boldsymbol{\gamma}}\left(u\right)\times\Upsilon$, and $\Psi\left(\mathbf{u}\right)=F_{\boldsymbol{\gamma}}\left(u\right)\times\Upsilon$, where $\mathbb{S}_k\equiv\left\{\boldsymbol{\gamma}\in\mathbb{R}^k:\|\boldsymbol{\gamma}\|=1\right\}$, $F_{n,\boldsymbol{\gamma}}\left(u\right)\equiv n^{-1}\sum_{i=1}^{n}1(\boldsymbol{\gamma}'\mathbf{X}_i\leq u)$, $F_{\boldsymbol{\gamma}}\left(u\right)\equiv\mathbb{E}\left[1(\boldsymbol{\gamma}'\mathbf{X}\leq u)\right]$ and $\Upsilon$ is the uniform distribution on $\mathbb{S}_k$;*

*$(iii)$ $\mathcal{W}_{exp}\equiv\left\{\mathbf{x}\in\mathcal{X}\mapsto\exp(i\mathbf{u}'\Phi\left(\mathbf{x}\right)):\mathbf{u}\in\Pi\right\}$, and $\Psi_n(\mathbf{u})=\Psi(\mathbf{u})$, where $\Pi$ is a compact, convex subset $\mathbb{R}^k$ with a non-empty interior, and $\Psi(\mathbf{u})$ is the CDF of k-variate standard normal distribution.*

Assumption 2 is standard in the literature, see e.g. Theorems 2.6 and 3.4 of Newey and McFadden (1994), Example 5.40 of van der Vaart (1998), and Graham et al. (2012). Assumption 2$(i)$ states that the true PS is known up to finite dimensional parameters $\boldsymbol{\beta}_0$, that is, that we are in a parametric setup. Assumption 2$(ii)$ imposes that the PS is bounded from above and from below. This assumption can be relaxed by assuming that

$(D/p(\mathbf{X};\boldsymbol{\beta}),(1-D)/(1-p(\mathbf{X};\boldsymbol{\beta})))' \leq \mathbf{b}(\mathbf{X})$ such that $\mathbb{E}\left[\|\mathbf{b}(\mathbf{X})\|^2\right] < \infty$. Assumptions 2($iii$)-($iv$) impose additional smoothness conditions on the PS, whereas Assumptions 2($v$) implies that, in a small neighborhood of $\boldsymbol{\beta}_0$ and for all $u \in \Pi$, the score $\dot{\mathbf{H}}_w(\boldsymbol{\beta},\mathbf{u})$ is bounded by an integrable function.

Assumption 3 restricts our attention to the IPS estimators (10)-(12). As mentioned before, we focus on such estimators due to their computational simplicity and transparency. Nonetheless, other IPS estimators can also be formed, provided that the weighting function $w$ and integrating measure $\Psi$ satisfy some high-level regularity conditions.

The next theorem characterizes the asymptotic properties of the IPS estimators $\widehat{\boldsymbol{\beta}}^{ips}_{n,w}$. Define

$$C_{w,\Psi} = \int_{\mathbb{R}^k} \left( \dot{\mathbf{H}}^c_w(\boldsymbol{\beta}_0,\mathbf{u})\dot{\mathbf{H}}_w(\boldsymbol{\beta}_0,\mathbf{u}) + \dot{\mathbf{H}}'_w(\boldsymbol{\beta}_0,\mathbf{u}) \left(\dot{\mathbf{H}}'_w(\boldsymbol{\beta}_0,\mathbf{u})\right)^c \right) \Psi(d\mathbf{u}),$$

and

$$l_{w,\Psi}(D,\mathbf{X};\boldsymbol{\beta}_0) = -C^{-1}_{w,\Psi} \cdot \int \left( \dot{\mathbf{H}}^c_w(\boldsymbol{\beta}_0,\mathbf{u}) w(\mathbf{X};\mathbf{u}) + \dot{\mathbf{H}}'_w(\boldsymbol{\beta}_0,\mathbf{u}) w^c(\mathbf{X},\mathbf{u}) \right) \Psi(d\mathbf{u}) \cdot \mathbf{h}(D,\mathbf{X};\boldsymbol{\beta}_0).$$

**Theorem 1** *Under Assumptions 2 - 3,*

$$\widehat{\boldsymbol{\beta}}^{ips}_{n,w} - \boldsymbol{\beta}_0 = o_p(1).$$

*Furthermore, provided that $C_{w,\Psi}$ is positive definite, as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{ips}_{n,w} - \boldsymbol{\beta}_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_{w,\Psi}(D_i,\mathbf{X}_i;\boldsymbol{\beta}_0) + o_p(1), \tag{13}$$

*and*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{ips}_{n,w} - \boldsymbol{\beta}_0\right) \xrightarrow{d} N\left(0, \Omega^{ips}_{w,\Psi}\right),$$

*where $\Omega^{ips}_{w,\Psi} \equiv \mathbb{E}\left[l_{w,\Psi}(D,\mathbf{X};\boldsymbol{\beta}_0)\, l_{w,\Psi}(D,\mathbf{X};\boldsymbol{\beta}_0)'\right]$.*

From Theorem 1, we conclude that the proposed IPS estimator is consistent, admits an asymptotically linear representation, and converges to a normal distribution. The asymptotic linear representation (13) plays a major role in establishing the asymptotic properties of causal parameters such as average, distributional, and quantile treatment effects; see Section 3.2. It is worth mentioning that it is not clear whether such an attractive feature is shared by the calibration type estimators such as those in Hainmueller (2012), Zubizarreta (2015), and Wong and Chan (2018).

## 3.2 Estimating Overall Treatment Effects

In this section we illustrate how one can estimate and make asymptotically valid inference about overall average, distributional, and quantile treatment effects under the strong ignorability Assumption 1, using IPW estimators based on the IPS.

Based on the discussion in Section 2.1, the IPW estimators for $ATE$, $DTE(y)$ and $QTE(\tau)$ are:

$$\widehat{ATE}_n^{ips} = \mathbb{E}_n\left[\left(w_{1,n}^{ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right) - w_{0,n}^{ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right)\right)Y\right], \tag{14}$$

$$\widehat{DTE}_n^{ips}(y) = \mathbb{E}_n\left[\left(w_{1,n}^{ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right) - w_{0,n}^{ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right)\right)1\{Y\leq y\}\right], \tag{15}$$

$$\widehat{QTE}_n^{ips}(\tau) = \widehat{q}_{n,Y(1)}^{ips}(\tau) - \widehat{q}_{n,Y(0)}^{ips}(\tau), \tag{16}$$

where, for $d\in\{0,1\}$,

$$\widehat{q}_{n,Y(d)}^{ips} = \arg\min_{q\in\mathbb{R}}\mathbb{E}_n\left[w_{d,n}^{ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right)\cdot\rho_\tau(Y-q)\right],$$

with $\rho_\tau(a) = a\cdot(\tau - 1\{a\leq 0\})$ the check function as in Koenker and Bassett (1978), and the weights $w_{1,n}^{ps}$ and $w_{0,n}^{ps}$ are as in (8)-(9).

In order to derive the asymptotic properties of (14)-(16), we need to make an additional assumption about the underlying distribution of the potential outcomes $Y(1)$ and $Y(0)$.

**Assumption 4** *For $d\in\{0,1\}$, (i) $\mathbb{E}\left[Y(d)^2\right] < M < \infty$ for some $M>0$, (ii)*

$$\mathbb{E}\left[\sup_{\beta\in\Theta_0}\left\|\frac{w_d^{ps}(D,\mathbf{X};\boldsymbol{\beta})(Y(d)-\mathbb{E}[Y(d)])}{d\cdot p(\mathbf{X};\boldsymbol{\beta}) + (1-d)(1-p(\mathbf{X};\boldsymbol{\beta}))}\cdot\dot{p}(\mathbf{X};\boldsymbol{\beta})\right\|\right] < \infty,$$

*and (iii) for some $\varepsilon>0$, $0<a_1<a_2<1$, $F_{Y(d)}$ is continuously differentiable on $\left[q_{Y(d)}(a_1)-\varepsilon, q_{Y(d)}(a_2)+\varepsilon\right]$ with strictly positive derivative $f_{Y(d)}$.*

Assumption 4$(i)$ requires potential outcomes to be square-integrable, whereas Assumption 4$(ii)$ is a mild regularity condition which guarantees that, in a small neighborhood of $\boldsymbol{\beta}_0$, the score of the IPW estimator for the $ATE$ is bounded by an integrable function. Assumption 4$(iii)$ requires potential outcomes to be continuously distributed and guarantees the uniqueness of quantiles. Assumption 4$(iii)$ only plays a role for quantile treatment effects, and can be relaxed at the cost of more complex arguments, see Chernozhukov et al. (2017).

Before stating the results as a theorem, let us define some important quantities. Let

$$\psi_{w,\Psi}^{ate}(Y,D,\mathbf{X}) = g^{ate}(Y,D,\mathbf{X}) - l_{w,\Psi}(D,\mathbf{X};\boldsymbol{\beta}_0)' \cdot \mathbf{G}_{\boldsymbol{\beta}}^{ate}, \tag{17}$$

$$\psi_{w,\Psi}^{dte}(Y,D,\mathbf{X};y) = g^{dte}(Y,D,\mathbf{X};y) - l_{w,\Psi}(D,\mathbf{X};\boldsymbol{\beta}_0)' \cdot \mathbf{G}_{\boldsymbol{\beta}}^{dte}(y), \tag{18}$$

$$\psi_{w,\Psi}^{qte}(Y,D,\mathbf{X};\tau) = -\left(g^{qte}(Y,D,\mathbf{X};\tau) - l_{w,\Psi}(D,\mathbf{X};\boldsymbol{\beta}_0)' \cdot \mathbf{G}_{\boldsymbol{\beta}}^{qte}(\tau)\right) \tag{19}$$

where, for $j \in \{ate, dte, qte\}$, $g^j(Y,D,\mathbf{X}) = g_1^j(Y,D,\mathbf{X}) - g_0^j(Y,D,\mathbf{X})$, with

$$g_d^{ate}(Y,D,\mathbf{X}) = w_d^{ps}(D,\mathbf{X};\boldsymbol{\beta}_0) \cdot (Y - \mathbb{E}[Y(d)]),$$

$$g_d^{dte}(Y,D,\mathbf{X};y) = w_d^{ps}(D,\mathbf{X};\boldsymbol{\beta}_0) \cdot \left(1\{Y \le y\} - F_{Y(d)}(y)\right),$$

$$g_d^{qte}(Y,D,\mathbf{X};\tau) = \frac{w_d^{ps}(D,\mathbf{X};\boldsymbol{\beta}_0) \cdot \left(1\left\{Y \le q_{Y(d)}(\tau)\right\} - \tau\right)}{f_{Y(d)}\left(q_{Y(d)}(\tau)\right)},$$

and

$$\mathbf{G}_{\boldsymbol{\beta}}^{ate} = \mathbb{E}\left[\left(\frac{g_1^{ate}(Y,D,\mathbf{X})}{p(\mathbf{X};\boldsymbol{\beta}_0)} + \frac{g_0^{ate}(Y,D,\mathbf{X})}{1-p(\mathbf{X};\boldsymbol{\beta}_0)}\right) \cdot \dot{p}(\mathbf{X};\boldsymbol{\beta}_0)\right],$$

$$\mathbf{G}_{\boldsymbol{\beta}}^{dte}(y) = \mathbb{E}\left[\left(\frac{g_1^{dte}(Y,D,\mathbf{X};y)}{p(\mathbf{X};\boldsymbol{\beta}_0)} + \frac{g_0^{dte}(Y,D,\mathbf{X};y)}{1-p(\mathbf{X};\boldsymbol{\beta}_0)}\right) \cdot \dot{p}(\mathbf{X};\boldsymbol{\beta}_0)\right],$$

$$\mathbf{G}_{\boldsymbol{\beta}}^{qte}(\tau) = \mathbb{E}\left[\left(\frac{g_1^{qte}(Y,D,\mathbf{X};\tau)}{p(\mathbf{X};\boldsymbol{\beta}_0)} + \frac{g_0^{qte}(Y,D,\mathbf{X};\tau)}{1-p(\mathbf{X};\boldsymbol{\beta}_0)}\right) \cdot \dot{p}(\mathbf{X};\boldsymbol{\beta}_0)\right].$$

The functions $g^{ate}$, $g^{dte}$ and $g^{qte}$ would be the influence functions of the $ATE$, $DTE(y)$ and $QTE(\tau)$ estimators, respectively, if the PS parameters $\boldsymbol{\beta}_0$ were known. With some abuse of notation, denote $\Omega_{w,\Psi}^{ate} = \mathbb{E}\left[\psi_{w,\Psi}^{ate}(Y,D,\mathbf{X})^2\right]$, $\Omega_{w,\Psi,y}^{dte} = \mathbb{E}\left[\psi_{w,\Psi}^{dte}(Y,D,\mathbf{X};y)^2\right]$, and $\Omega_{w,\Psi,\tau}^{qte} = \mathbb{E}\left[\left(\psi_{w,\Psi}^{qte}Y,D,\mathbf{X};\tau\right)^2\right]$.

**Theorem 2** *Under Assumptions 1 - 4, for each $y \in \mathbb{R}$, $\tau \in [\varepsilon, 1-\varepsilon]$, we have that, as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{ATE}_n^{ips} - ATE\right) \xrightarrow{d} N\left(0, \Omega_{w,\Psi}^{ate}\right),$$

$$\sqrt{n}\left(\widehat{DTE}_n^{ips} - DTE\right)(y) \xrightarrow{d} N\left(0, \Omega_{w,\Psi,y}^{dte}\right),$$

$$\sqrt{n}\left(\widehat{QTE}_n^{ips} - QTE\right)(\tau) \xrightarrow{d} N\left(0, \Omega_{w,\Psi,\tau}^{qte}\right).$$

Theorem 2 indicates that one can use our proposed IPS estimator to estimate a variety of causal parameters that are able to highlight treatment effect heterogeneity. Furthermore, Theorem 2 also suggests that in order to conduct asymptotically valid inference for these causal parameters, one simply needs to estimate the asymptotic variance $\Omega_{w,\Psi}^{ate}$, $\Omega_{w,\Psi,y}^{dte}$, and $\Omega_{w,\Psi,\tau}^{qte}$. Under additional smoothness conditions (for instance, the PS being twice continuously differentiable with bounded second derivatives), one can show that their sample analogues are consistent using standard empirical process arguments. We omit the details for the sake of brevity.

### 3.3 Estimating Treatment Effects on the Treated

In this section we focus on treatment effect parameters for the treated subpopulation instead of the overall population. Heckman et al. (1997) argue that analyzing treatment effects on the treated instead of overall treatment effects is more interesting when the policy intervention is directed at individuals with certain characteristics. For instance, if an employment program (or a clinical treatment) is directed at individuals who face barriers to employment (or who has some specific symptoms), perhaps there is little interest in analyzing the effect of this intervention on individuals with strong labor market attachment (or on individual who does not have these symptoms). Another potential advantage of focusing on the treated subpopulation is that one can weaken the overlap condition by allowing the PS to be close or even exactly equal to zero. This is particularly important in one of our applications in Section 5.

Analogous to the discussion in the previous section, here the goal is to make inference about the average, distributional and quantile treatment effect on the treated, defined as $ATT = \mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=1]$, $DTT(y) = F_{Y(1)|D=1}(y) - F_{Y(0)|D=1}(y)$, and $QTT(\tau) = q_{Y(1)|D=1}(\tau) - q_{Y(0)|D=1}(\tau)$, respectively, where, for $d \in \{0,1\}$, $F_{Y(d)|D=1}(y) = \mathbb{E}[1\{Y(d) \leq y\}|D=1]$, $y \in \mathbb{R}$, and $q_{Y(d)|D=1}(\tau) = \inf\left\{y : F_{Y(d)|D=1}(y) \geq \tau\right\}$, $\tau \in (0,1)$.

Let $w_1^{tt,ps}(D,\mathbf{X}) = D/\mathbb{E}[D]$ and

$$w_0^{tt,ps}(D,\mathbf{X}) = \frac{(1-D)\,p(\mathbf{X})}{1-p(\mathbf{X})} \Big/ \mathbb{E}\left[\frac{(1-D)\,p(\mathbf{X})}{1-p(\mathbf{X})}\right].$$

Note that functionals of the distribution of $Y(1)$ for the treated subpopulation can be directly estimated from the data using the analogy principle. Thus, when analyzing treatment effects on the treated, the main challenge faced is to identify and make inference about functionals of the distribution of $Y(0)$ for the treated subpopulation. Towards this end, we make the following assumptions.

**Assumption 5** *(a) Given $\mathbf{X}$, $Y(0)$ is independent from $D$; and (b) for all $\mathbf{x} \in \mathcal{X}$, $p(\mathbf{x})$ is uniformly bounded away from one.*

**Assumption 6** *For $d \in \{0,1\}$, (i) $\mathbb{E}\left[Y(d)^2|D=1\right] < M < \infty$ for some $M > 0$, (ii)*

$$\mathbb{E}\left[\sup_{\beta \in \Theta_0}\left\|\frac{w_0^{tt,ps}(D,\mathbf{X};\boldsymbol{\beta})\,(Y - \mathbb{E}[Y(0)|D=1])}{p(\mathbf{X};\boldsymbol{\beta})\,(1-p(\mathbf{X};\boldsymbol{\beta}))} \cdot \dot{p}(\mathbf{X};\boldsymbol{\beta})\right\|\right] < \infty,$$

*and (iii) for some $\varepsilon > 0$, $0 < a_1 < a_2 < 1$, $F_{Y(d)|D=1}$ is continuously differentiable on $\left[q_{Y(d)|D=1}(a_1) - \varepsilon, q_{Y(d)|D=1}(a_2) + \varepsilon\right]$ with strictly positive derivative $f_{Y(d)|D=1}$.*

Assumption 5 is a weaker version of Assumption 1, where we do not impose any lower bound on the PS, nor make any assumption about the relationship of $Y(1)$, $D$, and $\mathbf{X}$. Assumption 6 is the analogue of Assumption 4.

As shown by Heckman et al. (1997), under Assumptions 5 - 6, the $ATT$ is identified by

$$ATT = \mathbb{E}\left[\left(w_1^{tt,ps}(D,\mathbf{X}) - w_0^{tt,ps}(D,\mathbf{X})\right)Y\right].$$

Analogously, $F_{Y(0)|D=1}(y)$ is identified by

$$F_{Y(0)|D=1}(y) = \mathbb{E}\left[w_0^{tt,ps}(D,\mathbf{X})\,1\{Y \leq y\}\right],$$

implying that both $DTT(y)$ and $QTT(\tau)$ can also be written as functionals of the observed data; see e.g. Firpo (2007). Such identification results suggest that we can estimate the $ATT$, $DTT(y)$ and $QTT(\tau)$ by

$$\widehat{ATT}_n^{ips} = \mathbb{E}_n\left[\left(w_{n,1}^{tt,ps}(D,\mathbf{X}) - w_{n,0}^{tt,ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right)\right)Y\right],$$

$$\widehat{DTT}_n^{ips}(y) = \mathbb{E}_n\left[\left(w_{n,1}^{tt,ps}(D,\mathbf{X}) - w_{n,0}^{tt,ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right)\right)1\{Y \leq y\}\right],$$

$$\widehat{QTT}_n^{ips}(\tau) = \widehat{q}_{n,Y(1)|D=1}(\tau) - \widehat{q}_{n,Y(0)|D=1}^{ips}(\tau),$$

where

$$\widehat{q}_{n,Y(1)|D=1} = \arg\min_{q\in\mathbb{R}} \mathbb{E}_n\left[w_{n,1}^{tt,ps}(D,\mathbf{X}) \cdot \rho_\tau(Y-q)\right],$$

$$\widehat{q}_{n,Y(0)|D=1}^{ips} = \arg\min_{q\in\mathbb{R}} \mathbb{E}_n\left[w_{n,0}^{tt,ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right) \cdot \rho_\tau(Y-q)\right],$$

$w_{n,1}^{tt,ps}(D,\mathbf{X}) = D/\,\mathbb{E}_n[D]$, and

$$w_{n,0}^{tt,ps}(D,\mathbf{X};\boldsymbol{\beta}) = \frac{(1-D)\,p(\mathbf{X};\boldsymbol{\beta})}{1-p(\mathbf{X};\boldsymbol{\beta})} \Big/ \mathbb{E}_n\left[\frac{(1-D)\,p(\mathbf{X};\boldsymbol{\beta})}{1-p(\mathbf{X};\boldsymbol{\beta})}\right].$$

The next theorem summarizes the asymptotic properties of the IPW estimators for the treatment effect on the treated based on the IPS. For $j \in \{att, dtt, qtt\}$, let $g^j(Y,D,\mathbf{X}) = g_1^j(Y,D,\mathbf{X}) - g_0^j(Y,D,\mathbf{X})$, with, for $d \in \{0,1\}$,

$$g_d^{att}(Y,D,\mathbf{X}) = w_d^{tt,ps}(D,\mathbf{X};\boldsymbol{\beta}_0)\cdot(Y - \mathbb{E}[Y(d)|D=1]),$$

$$g_d^{dtt}(Y,D,\mathbf{X};y) = w_d^{tt,ps}(D,\mathbf{X};\boldsymbol{\beta}_0)\cdot\left(1\{Y\leq y\} - F_{Y(d)|D=1}(y)\right),$$

$$g_d^{qtt}(Y,D,\mathbf{X};\tau) = \frac{w_d^{tt,ps}(D,\mathbf{X};\boldsymbol{\beta}_0)\cdot\left(1\left\{Y \leq q_{Y(d)|D=1}(\tau)\right\} - \tau\right)}{f_{Y(d)|D=1}\left(q_{Y(d)|D=1}(\tau)\right)}.$$

Finally, let $\Omega_{w,\Psi}^{att} = \mathbb{E}\left[\psi_{w,\Psi}^{att}(Y,D,\mathbf{X})^2\right]$, $\Omega_{w,\Psi,y}^{dtt} = \mathbb{E}\left[\psi_{w,\Psi}^{dtt}(Y,D,\mathbf{X};y)^2\right]$, and $\Omega_{w,\Psi,\tau}^{qtt} = \mathbb{E}\left[\psi_{w,\Psi}^{qtt}(Y,D,\mathbf{X};\tau)^2\right]$, where $\psi_{w,\Psi}^{att}$, $\psi_{w,\Psi}^{dtt}$, and $\psi_{w,\Psi}^{qtt}$ are defined analogously to (17)-(19), but with $g^{att}$, $g^{dtt}$, $g^{qtt}$ playing the role of $g^{ate}$, $g^{dte}$, $g^{qte}$, respectively, and

$$\mathbf{G}_{\boldsymbol{\beta}}^{att} = \mathbb{E}\left[\frac{g_0^{att}(Y,D,\mathbf{X})}{p(\mathbf{X};\boldsymbol{\beta}_0)(1-p(\mathbf{X};\boldsymbol{\beta}_0))}\cdot\dot{p}(\mathbf{X};\boldsymbol{\beta}_0)\right],$$

$$\mathbf{G}_{\boldsymbol{\beta}}^{dtt}(y) = \mathbb{E}\left[\frac{g_0^{dtt}(Y, D, \mathbf{X}; y)}{p(\mathbf{X}; \boldsymbol{\beta}_0)(1 - p(\mathbf{X}; \boldsymbol{\beta}_0))} \cdot \dot{p}(\mathbf{X}; \boldsymbol{\beta}_0)\right],$$

$$\mathbf{G}_{\boldsymbol{\beta}}^{qtt}(\tau) = \mathbb{E}\left[\frac{g_0^{qtt}(Y, D, \mathbf{X}; \tau)}{p(\mathbf{X}; \boldsymbol{\beta}_0)(1 - p(\mathbf{X}; \boldsymbol{\beta}_0))} \cdot \dot{p}(\mathbf{X}; \boldsymbol{\beta}_0)\right],$$

playing the role of $\mathbf{G}_{\boldsymbol{\beta}}^{ate}$, $\mathbf{G}_{\boldsymbol{\beta}}^{dte}$, and $\mathbf{G}_{\boldsymbol{\beta}}^{qte}$, respectively.

**Theorem 3** *Under Assumptions 2, 3, 5, and 6, for each $y \in \mathbb{R}$, $\tau \in [\varepsilon, 1 - \varepsilon]$, we have that, as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{ATT}_n^{ips} - ATT\right) \xrightarrow{d} N\left(0, \Omega_{w,\Psi}^{att}\right),$$

$$\sqrt{n}\left(\widehat{DTT}_n^{ips} - DTT\right)(y) \xrightarrow{d} N\left(0, \Omega_{w,\Psi,y}^{dtt}\right),$$

$$\sqrt{n}\left(\widehat{QTT}_n^{ips} - QTT\right)(\tau) \xrightarrow{d} N\left(0, \Omega_{w,\Psi,\tau}^{qtt}\right).$$

**Remark 4** When average, distributional and quantile treatment effect on the treated are the main parameters of interest, instead of using (7), one may wish to estimate $\boldsymbol{\beta}_0$ such that, for every measurable, integrable function $f$ of the covariates,

$$\mathbb{E}\left[\left(\left(\frac{(1 - D)\, p(\mathbf{X}; \boldsymbol{\beta}_0)}{1 - p(\mathbf{X}; \boldsymbol{\beta}_0)} \Big/ \mathbb{E}\left[\frac{(1 - D)\, p(\mathbf{X}; \boldsymbol{\beta}_0)}{1 - p(\mathbf{X}; \boldsymbol{\beta}_0)}\right]\right) - \frac{D}{\mathbb{E}[D]}\right) f(\mathbf{X})\right] = 0. \tag{20}$$

From the discussion in Section 2, and the fact that

$$\frac{(1 - D)\, p(\mathbf{X}; \boldsymbol{\beta}_0)}{1 - p(\mathbf{X}; \boldsymbol{\beta}_0)} - D = \frac{(1 - D)}{1 - p(\mathbf{X}; \boldsymbol{\beta}_0)} - 1,$$

and $\mathbb{E}[(1 - D)\, p(\mathbf{X}; \boldsymbol{\beta}_0)/(1 - p(\mathbf{X}; \boldsymbol{\beta}_0))] = \mathbb{E}[D]$, we can conclude that one can use

$$H_{0,w}(\boldsymbol{\beta}, \mathbf{u}) = \mathbb{E}\left[\left(\left(\frac{(1 - D)}{1 - p(\mathbf{X}; \boldsymbol{\beta})} \Big/ \mathbb{E}\left[\frac{(1 - D)}{1 - p(\mathbf{X}; \boldsymbol{\beta})}\right]\right) - 1\right) w(\mathbf{X}; \mathbf{u})\right]$$

to construct a minimum distance estimator for $\boldsymbol{\beta}_0$ analogous to (5). In order to avoid additional cumbersome notation, the results stated in Theorem 3 do not use this alternative IPS estimator, though such a modification is straightforward.

# 4 Simulations

In this section, we conduct a series of Monte Carlo experiments in order to study the finite sample properties of our proposed treatment effect estimators based on the IPS. In particular, we compare the performance of different IPW estimators for the $ATE$ and the $QTE(\tau)$, $\tau \in \{0.25, 0.5, 0.75\}$ when one estimates the PS using our proposed IPS estimators (10)-(12), Imai and Ratkovic (2014) covariate balancing propensity score (CBPS) as in (2) with $f(\mathbf{X}) = \mathbf{X}$, and the classical maximum likelihood (ML) approach. In all cases, we consider a logistic PS model where all available covariates enter linearly. All treatment effect estimators use stabilized weights (8) and (9).

We consider sample size $n$ equal to 500. For each design, we conduct $10,000$ Monte Carlo simulations, and calculate the average bias, root mean square error (RMSE), empirical 95% coverage probability, and the average length of a 95% confidence interval. The confidence intervals are based on the normal approximation in Theorem 2, with the variances being estimated by their sample analogue. For the $QTE(\tau)$ variance, we estimate the potential outcome densities using the Gaussian kernel coupled with Silverman's rule-of-thumb bandwidth. These are the default choices of the `density` function in the `stats` package in `R`.

Our simulation design is largely based on Kang and Schafer (2007). Let $\mathbf{X} = (X_1, X_2, X_3, X_4)'$ be distributed as $N(0, I_4)$, and $I_4$ be the $4 \times 4$ identity matrix. The true PS is given by

$$p(\mathbf{X}) = \frac{\exp(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)}{1 + \exp(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)},$$

and the treatment status $D$ is generated as $D = 1\{p(\mathbf{X}) > U\}$, where $U$ follows a uniform $(0, 1)$ distribution. The potential outcomes $Y(1)$ and $Y(0)$ are given by

$$Y(1) = 210 + m(\mathbf{X}) + \varepsilon(1),$$

$$Y(0) = 200 - m(\mathbf{X}) + \varepsilon(0),$$

where $m(\mathbf{X}) = 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4$, $\varepsilon(1)$ and $\varepsilon(1)$ are independent

$N(0,1)$ random variables. The $ATE$ and the $QTE(\tau)$ are equal to 10, for all $\tau \in (0,1)$.

We consider two different scenarios to assess the sensibility of the proposed estimators under misspecified models that are "nearly correct". In the first experiment the observed data is $\{(Y_i, D_i, \mathbf{X}_i)\}_{i=1}^n$, and therefore all IPW estimators are correctly specified. In the second experiment the observed data is $\{(Y_i, D_i, \mathbf{Z}_i)\}_{i=1}^n$, where $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)'$ with $Z_1 = \exp(X_1/2)$, $Z_2 = X_2/(1+\exp(X_1))$, $Z_3 = (X_1X_3/25+0.6)^3$, and $Z_4 = (X_2+X_4+20)^2$. In this second scenario, the IPW estimators for $ATE$ and $QTE(\tau)$ are misspecified.

Table 1 displays the simulation results for both scenarios. When the PS model is correctly specified, all estimators perform well. Among the considered estimators, the ones based on the IPS with exponential or projection weighting tend to dominate the others in all considered criteria. For example, the RMSE of estimators based on the IPS with projection weight function tend to be at least 10% smaller than those based on the CBPS, with the exception of the $QTE(0.25)$. On the other hand, IPW estimators based on the IPS with indicator function tend to give slightly larger confidence intervals than when using other IPS estimators, perhaps because there are multiple covariates, implying that many $1\{\mathbf{X}_i \leq \mathbf{u}\}$ are equal to zero when $\mathbf{u}$ is evaluated at the sample observations. Nonetheless, the length of such confidence intervals is of similar magnitude as those based on ML estimates, and no significant size distortions are present.

When the PS model is misspecified, our Monte Carlo results suggest that the potential gains of using the IPS can be more pronounced. In this scenario, we note that estimators based on ML tend to be substantially biased, have relatively high RMSE, and inference tends to be misleading. These findings are in line with the results in Kang and Schafer (2007). Overall, estimators based on CBPS improve upon ML, though undercoverage is still an issue even when one focuses only on the $ATE$. Finally, we note that our proposed IPS estimators further improve upon CBPS in terms of bias, RMSE, and coverage probability. In particular, estimators based on the IPS with the projection weight function have the lowest bias and RMSE, and their confidence intervals are close to the nominal coverage.

**Table 1:** Monte Carlo study of the performance of IPW estimators for $ATE$ and $QTE$ based on different propensity score estimation methods. Sample size: $n = 500$.

| | Corectly Specified Model | | | | Misspecified Model | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias | RMSE | COV | ACIL | Bias | RMSE | COV | ACIL |
| (a) $ATE$ | | | | | | | | |
| $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$ | -0.066 | 3.730 | 0.940 | 14.081 | 1.869 | 4.196 | 0.917 | 14.266 |
| $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$ | 0.812 | 3.656 | 0.966 | 15.553 | 2.679 | 4.811 | 0.943 | 17.840 |
| $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$ | -0.058 | 3.599 | 0.942 | 13.602 | 0.302 | 3.371 | 0.952 | 13.524 |
| $\widehat{\boldsymbol{\beta}}_{n}^{cbps}$ | -0.103 | 4.090 | 0.931 | 14.997 | 2.583 | 4.571 | 0.871 | 13.959 |
| $\widehat{\boldsymbol{\beta}}_{n}^{mle}$ | -0.124 | 4.465 | 0.941 | 16.419 | 5.930 | 10.819 | 0.858 | 20.587 |
| (b) $QTE(0.25)$ | | | | | | | | |
| $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$ | 0.016 | 4.374 | 0.972 | 19.656 | -2.170 | 4.845 | 0.953 | 20.221 |
| $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$ | 0.565 | 4.613 | 0.987 | 22.696 | -1.052 | 4.550 | 0.981 | 22.483 |
| $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$ | 0.022 | 4.372 | 0.971 | 19.605 | -1.127 | 4.348 | 0.968 | 20.036 |
| $\widehat{\boldsymbol{\beta}}_{n}^{cbps}$ | -0.018 | 4.342 | 0.971 | 19.469 | -1.402 | 4.540 | 0.963 | 20.135 |
| $\widehat{\boldsymbol{\beta}}_{n}^{mle}$ | -0.035 | 4.440 | 0.973 | 19.899 | 0.856 | 8.075 | 0.978 | 23.667 |
| (c) $QTE(0.50)$ | | | | | | | | |
| $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$ | -0.010 | 4.461 | 0.958 | 18.091 | 0.964 | 4.465 | 0.958 | 19.381 |
| $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$ | 0.813 | 4.510 | 0.973 | 19.856 | 1.922 | 4.880 | 0.954 | 21.409 |
| $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$ | -0.011 | 4.373 | 0.956 | 17.743 | -0.001 | 4.129 | 0.969 | 17.803 |
| $\widehat{\boldsymbol{\beta}}_{n}^{cbps}$ | -0.049 | 4.706 | 0.955 | 18.946 | 1.812 | 4.805 | 0.928 | 19.534 |
| $\widehat{\boldsymbol{\beta}}_{n}^{mle}$ | -0.058 | 4.906 | 0.958 | 19.859 | 5.235 | 12.220 | 0.911 | 26.179 |
| (d) $QTE(0.75)$ | | | | | | | | |
| $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$ | -0.133 | 5.696 | 0.928 | 20.587 | 5.262 | 7.693 | 0.864 | 22.214 |
| $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$ | 1.042 | 5.440 | 0.952 | 21.360 | 5.809 | 8.126 | 0.897 | 25.217 |
| $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$ | -0.129 | 5.462 | 0.928 | 19.715 | 1.480 | 4.973 | 0.951 | 19.760 |
| $\widehat{\boldsymbol{\beta}}_{n}^{cbps}$ | -0.178 | 6.217 | 0.922 | 22.094 | 6.052 | 8.279 | 0.804 | 21.513 |
| $\widehat{\boldsymbol{\beta}}_{n}^{mle}$ | -0.194 | 6.674 | 0.933 | 23.835 | 11.481 | 18.546 | 0.793 | 32.267 |

Note: Simulations based on 10,000 Monte Carlo experiments. Bias, Monte Carlo Bias; RMSE, root mean square error; COV, Monte Carlo coverage of 95% normal confidence interval; ACIL, Monte Carlo average of 95% normal confidence interval length; ATE, average treatment effect; QTE($\tau$), quantile treatment effect at $\tau$ quantile. All propensity scores follow a logistic link function. $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$, IPW estimator based on IPS estimator (10); $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$, IPW estimator based on IPS estimator (11); $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$, IPW estimator based on IPS estimator (12); $\widehat{\boldsymbol{\beta}}_{n}^{cbps}$, IPW estimator based on CBPS estimator with moment equation (2), with $f(\mathbf{X}) = \mathbf{X}$; $\widehat{\boldsymbol{\beta}}_{n}^{mle}$, IPW estimator based on MLE.

In summary, our simulation results highlight that, by fully exploiting the covariate balancing property of the PS, one can indeed obtain treatment effect estimators with improved finite sample properties. Interestingly, such improvements can be more pronounced when the PS model is misspecified but "nearly correct". Within the IPS class of estimators, our simulation results suggest that using the projection weight function provides a good compromise in terms of bias, RMSE, and coverage probability. Nonetheless, it is also worth mentioning that, in general, the relative performance of IPW estimators based on different PS estimation methods can be application dependent, and one procedure may not always dominate the others, especially under propensity score misspecification. Thus, one should view these different PS estimators as complements, and not necessarily substitutes.

# 5 Empirical Applications

In this section, we apply our proposed tools to three different datasets. First, we revisit Ichino et al. (2008) and use Italian data from the early 2000s to study if temporary work agency (TWA) assignment affects the probability of finding a stable job later on. Second, we revisit Connors et al. (1996) and analyze the effectiveness of right heart catheterization (RHC) in the intensive care unit (ICU) of critically ill patients on survival at 30 days after admission. Finally, using data from Blattman and Annan (2010), we study the effect of child abduction by a militant group on future wages.

## 5.1 Effect of Temporary Work Assignment on Future Stable Employment

In temporary agency work, a company that needs employees signs a contract with a TWA, which in turn is in charge of hiring and subsequently leasing these workers to the company. In contrast to "traditional" jobs, the TWA is in charge of paying the workers salary and fringe benefits, whereas the company's responsibility is to train and guide the workers. One of the main arguments of introducing temporary agency work is that it helps workers facing barriers to employment find a stable job later on.

In order to evaluate whether TWA assignment has a positive impact on employment, Ichino et al. (2008) collected data for two Italian regions, Tuscany and Sicily, in the early 2000s. The dataset contains 2030 individuals, 511 of them treated and 1519 controls. Here, the treated group consists of individuals who were on a TWA assignment during the first 6 months of 2001, whereas the control group contains individuals aged 18 - 40, who belonged to the labour force but did not have a stable job on January 2001, and who did not have a TWA assignment during the first semester of 2001. Thus, both treatment groups were drawn from the same local labour market. The outcome of interest is having a permanent job at the end of 2002. A rich set of variables related to demographic characteristics, family background, educational achievements, and work experience before the treatment period were collected to adjust for potential confounding (see Table 1 in Ichino et al. (2008)). Using PS matching, Ichino et al. (2008) find evidence that TWA assignment has a positive effect on permanent employment, especially in Tuscany. The results for Sicily are sensitive to small violations of the strong ignorability assumption. Therefore, in what follows we focus on the Tuscany sub-sample[1].

We use the results in Sections 3.2 and 3.3 to estimate $ATE$ and $ATT$. We compare different IPW estimators based on the same PS estimation methods as in the simulation studies in Section 4. Table 2 shows the point estimates and standard errors (in parentheses) for the whole Tuscany sample, and also presents some heterogeneity results based on gender. The PS specification we use is the one adopted by Ichino et al. (2008) which includes all the pre-treatment variables mentioned in Table 1 of Ichino et al. (2008), distance squared, and an interaction between self-employment and one of the provinces.

The results in Table 2 suggest that both the $ATE$ and $ATT$ are positive, and statistically significant at the conventional levels, regardless of the estimation procedure adopted. The overall average effect of TWA assignment on the probability of having a permanent job ranges from 15 to 21, 14 to 23, and 15 to 19 percentage points when using the whole sample, the male subpopulation, and the female subpopulation, respectively.

1 The data is publicly available at http://qed.econ.queensu.ca/jae/2008-v23.3/ichino-mealli-nannicini/.

**Table 2:** Effect of TWA assignment on the probability to find a permanent job: IPW estimators using different propensity score estimation methods.

| | $(a)$ Results for ATE | | | | | $(b)$ Results for ATT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\widehat{\boldsymbol{\beta}}_n^{mle}$ | $\widehat{\boldsymbol{\beta}}_n^{cbps}$ | $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$ | $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$ | $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$ | $\widehat{\boldsymbol{\beta}}_n^{mle}$ | $\widehat{\boldsymbol{\beta}}_n^{cbps}$ | $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$ | $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$ | $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$ |
| Whole Sample | 17.83 | 20.67 | 15.85 | 19.23 | 18.33 | 15.89 | 16.24 | 12.28 | 15.25 | 16.93 |
| | (4.62) | (3.90) | (3.25) | (3.52) | (4.52) | (3.73) | (4.05) | (3.67) | (4.13) | (4.24) |
| Male | 14.40 | 22.79 | 14.39 | 19.17 | 17.63 | 17.37 | 17.33 | 12.89 | 16.03 | 14.49 |
| | (7.22) | (5.43) | (4.61) | (5.01) | (6.00) | (5.42) | (5.39) | (5.41) | (5.39) | (5.95) |
| Female | 16.01 | 18.58 | 15.43 | 15.28 | 14.84 | 17.72 | 20.22 | 14.91 | 11.41 | 16.93 |
| | (5.64) | (5.95) | (4.31) | (4.84) | (4.59) | (5.22) | (5.56) | (5.20) | (6.54) | (4.98) |

Note: Same data used by Ichino et al. (2008). All propensity score follows a logistic link function. Panel $(a)$ presents results for the Average Treatment Effect for the overall population, whereas panel $(b)$ present the analogous results for the treated subpopulation. Standard errors in parentheses. $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$, IPW estimator based on IPS estimator (10); $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$, IPW estimator based on IPS estimator (11); $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$, IPW estimator based on IPS estimator (12); $\widehat{\boldsymbol{\beta}}_n^{cbps}$, IPW estimator based on CBPS estimator with moment equation (2), with $f(\mathbf{X}) = \mathbf{X}$; $\widehat{\boldsymbol{\beta}}_n^{mle}$, IPW estimator based on MLE.

The $ATE$ and $ATT$ tend to achieve their smaller point estimates when coupled with the IPS estimator with indicator weight, $\widehat{\boldsymbol{\beta}}_{n,ind}^{ips}$, and its highest point estimate when one uses the CBPS estimator, $\widehat{\boldsymbol{\beta}}_n^{cbps}$. Interestingly, the IPS estimators can provide gains of efficiency when compared to both the MLE and CBPS estimators. Such gains are especially notable in the analysis of $ATE$, when one uses the IPS estimator with indicator weight function. These findings suggests that the IPS can indeed lead to improved treatment effect estimators in relevant settings.

## 5.2 Effect of Right Heart Catheterization on 30 Days Survival

In a right heart catheterization, the physician places a catheter in right-side of the heart in order to measure the pressure in the heart and lungs of critically ill patients. The catheter is usually left in place for days, so it can continuously provide information that help doctors to diagnose heart conditions and to guide therapy. RHC also involves risks since it is an invasive style of care, and can have complications such as vein thrombosis, line sepsis, and bacterial endocarditis. Furthermore, the information collected by RHC may lead to false diagnoses, which in turn, may lead to inappropriate changes in therapy. Overall, it is not clear if RHC leads to better patient outcomes or not.

In this section we revisit Connors et al. (1996), and reanalyze data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUP-

PORT) conducted in five U.S. hospitals between 1989 and 1994 to assess the impact of RHC on survival at 30 days after admission of critically ill patients. The study collected data on 5735 patients, where 2184 of them received RHC within 24 hours of admission (treated group), and 3551 did not (control group). Based on expert information, a rich set of 72 variables relating to the RHC decision was also collected; see e.g. Tables 1 and 2 in Hirano and Imbens (2001)[2].

At the time of the study by Connors et al. (1996), RHC was thought to lead to better patient outcomes, though its benefits had not been demonstrated in any randomized clinical trial. Using the rich set of covariates from SUPPORT, Connors et al. (1996) use propensity score matching and find that, for the treated patients, RHC appears to lead to lower survival than not performing RHC. In what follows, we use IPW estimators instead.

Given that RHC is usually directed at patients with certain health conditions, here we focus on analyzing the *ATT* as discussed in Section 3.3. The outcome of interest is an indicator of survival 30 days after ICU admission. As in previous studies, we estimate the PS under a logistic model with all 72 covariates in Table 2 of Hirano and Imbens (2001). We consider different PS estimators: MLE, CBPS and our three proposed IPS. Given the *ATT* is the main parameter of interest, we follow Remark 4, and aim to weight the control group such that their weighted covariate distribution is balanced with that of the treatment group; for the CBPS, we replace $f(\mathbf{X})$ in (20) with $\mathbf{X}$.

Table 3 shows the *ATT* point estimates and standard errors (in parentheses). Overall, the analysis suggests that RHC had a negative, statistically significant, effect on 30-days survival probability among the treated. Over the different PS methods, the *ATT* point estimates range from –7.68 (CBPS) to -3.37 percentage points (IPS with projection weight function). Note that the asymptotic relative efficiency (ARE) of the *ATT* estimator based on CBPS to the one based on MLE is 1.52 ($2.05^2/1.66^2$), suggesting that estimators based on the CBPS are substantially more efficient than those based on ML. Nonetheless, the ARE of *ATT* estimators based on IPS with indicator, exponential, and projection weights

2 The data is publicly available at http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/rhc.html.

**Table 3:** Effect of RHC on the 30-days survival probability: IPW estimators using different propensity score estimation methods.

| Results for the Average Treatment Effect on the Treated | | | | | |
|---|---|---|---|---|---|
| | $\widehat{\boldsymbol{\beta}}_n^{mle}$ | $\widehat{\boldsymbol{\beta}}_n^{cbps}$ | $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$ | $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$ | $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$ |
| | -5.81 | -7.68 | -6.44 | -6.79 | -3.37 |
| | (2.05) | (1.66) | (1.53) | (1.38) | (1.64) |

Note: Same data used by Connors et al. (1996). All propensity score follows a logistic link function. Standard errors in parentheses. All IPS and CBPS estimators are modified in accordance of Remark 4. $\widehat{\boldsymbol{\beta}}_{n,\text{ind}}^{ips}$, IPW estimator based on IPS estimator (10); $\widehat{\boldsymbol{\beta}}_{n,\text{proj}}^{ips}$, IPW estimator based on IPS estimator (11); $\widehat{\boldsymbol{\beta}}_{n,\text{exp}}^{ips}$, IPW estimator based on IPS estimator (12); $\widehat{\boldsymbol{\beta}}_n^{cbps}$, IPW estimator based on CBPS estimator with moment equation (20), with $f(\mathbf{X}) = \mathbf{X}$; $\widehat{\boldsymbol{\beta}}_n^{mle}$, IPW estimator based on MLE.

to the one based on MLE are, respectively, 1.79, 2.19, and 1.55, implying that, by fully exploiting the covariate balancing property of the PS, one can get important additional efficiency gains relative to both MLE and CBPS. Such findings are in line with our simulation results, and highlight the potential attractiveness of the IPS methodology.

## 5.3 Effect of Child Soldiering on Future Earnings

Understanding the impact of combat experience on human capital is crucial from the economic, psychological, and social points of view. The dominant view holds that individuals involved in combat are traumatized, violent, and isolated from society. On the other hand, there is some ethnographic evidence that resilience rather than disabling psychological trauma is the prevailing effect; see Blattman and Annan (2010) and references therein.

In order to assess the impact of combat on educational, labor market, psychological, and health outcomes, Blattman and Annan (2010) use data from Phase I of the Survey of War Affected Youth in northern Uganda (SWAY). Blattman and Annan (2010) argue that forced recruitment in Uganda is mostly due to random night raids on rural homes, and therefore, conditional on a vector of observed characteristics, treatment (abduction) is as good as random. Thus, one can use the results in Sections 3.2 and 3.3 to assess the effect of abduction on different outcomes. For conciseness, we focus solely on the impact of abduction on wages (measured in 2005 Ugandan shillings).

The Phase I SWAY data were collected from 1216 males in Uganda during 2005-2006, but wage data is available for 504 observations only[3]. We focus on this subset. Among the 504 observations, 320 had been abducted by militant groups before 2005 but had escaped by the time of the study. Covariates include dummy variables for geographical regions, whether the father is a farmer, whether parents (father, mother, or both) had died during or before 1996, age in 1996, father's years of education, mother's years of education, household size in 1996, household wealth in 1996, and household land, stock and cattle. All these covariates enter linearly in our logit PS model. As in Blattman and Annan (2010), we use survey weights to account for the stratified sampling, selective non-survival and attrition.

Table 4 presents results for average and quantile treatment effects for the overall and the treated subpopulation using the IPW estimators based on the PS estimation methods as in the simulation studies in Section 4. We show both point estimates and standard errors (in parentheses). The results reveal interesting features. For the $ATE$ and $ATT$, all estimators are negative, with similar magnitudes. The $ATE$ estimates range from -1845 to -1231 Ugandan shillings, whereas the $ATT$ range from -2195 to -1221 Ugandan shillings. However, only the estimators based on CBPS are statistically different from zero at the 90% confidence level. For the quantile treatment effects, the results also agree in terms of sign and magnitude, though most estimates are not statistically different from zero at the 90% confidence level. The exceptions are the $QTE\,(0.25)$ estimates based on MLE, CBPS, and IPS with indicator weight function, and the $QTT(0.25)$ estimates based on MLE, CBPS, and IPS with projection weight function. Interesting, we find that $QTE$ and $QTT$ point estimates are substantially lower than the $ATE$ and $ATT$, suggesting that the effect of child abduction on wages in Uganda may be heterogeneous.

The results in Table 4 reveal conflicting conclusions about the effect of abduction on future wages: although point estimates are generally in agreement, their statistical significance depend on the PS estimation method. Given that all PS specifications are the same (logistic model with linear effects), PS misspecification may have caused these differences.

3 The data is publicly available at https://chrisblattman.com/projects/sway/.

**Table 4:** Effect of abduction on monthly wage (Ugandan schillings): IPW estimators using different propensity score estimation methods.

| | (*a*) Results for the Overall Population | | | | | (*b*) Results for the Treated Subpopulation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\widehat{\beta}_n^{mle}$ | $\widehat{\beta}_n^{cbps}$ | $\widehat{\beta}_{n,\text{ind}}^{ips}$ | $\widehat{\beta}_{n,\text{exp}}^{ips}$ | $\widehat{\beta}_{n,\text{proj}}^{ips}$ | $\widehat{\beta}_n^{mle}$ | $\widehat{\beta}_n^{cbps}$ | $\widehat{\beta}_{n,\text{ind}}^{ips}$ | $\widehat{\beta}_{n,\text{exp}}^{ips}$ | $\widehat{\beta}_{n,\text{proj}}^{ips}$ |
| Average Effects | -1650 | -1845 | -1493 | -1231 | -1438 | -1841 | -1306 | -1574 | -1221 | -2195 |
| | (1109) | (926) | (1120) | (851) | (1127) | (1247) | (680) | (1471) | (853) | (1677) |
| Quantile $\tau$ Effects | | | | | | | | | | |
| $\tau = 0.25$ | -233 | -233 | -238 | -200 | -200 | -269 | -333 | -214 | -214 | -269 |
| | (126) | (124) | (127) | (131) | (146) | (139) | (134) | (144) | (141) | (163) |
| $\tau = 0.50$ | -125 | -229 | -179 | 0 | 150 | -179 | -250 | -83 | 0 | -83 |
| | (208) | (225) | (208) | (207) | (208) | (247) | (248) | (256) | (208) | (260) |
| $\tau = 0.75$ | -500 | -385 | -385 | -143 | -83 | -143 | -143 | -143 | -143 | -143 |
| | (432) | (431) | (448) | (486) | (440) | (486) | (470) | (549) | (499) | (699) |

Note: Same data used by Blattman and Annan (2010). All propensity score follows a logistic link function. Panel (*a*) presents results for the overall population (ATE, QTE(0.25), QTE(0.5), and QTE(0.75)), whereas panel (*b*) present the analogous results for the treated subpopulation. Standard errors in parentheses. $\widehat{\beta}_{n,\text{ind}}^{ips}$, IPW estimator based on IPS estimator (10); $\widehat{\beta}_{n,\text{exp}}^{proj}$, IPW estimator based on IPS estimator (11); $\widehat{\beta}_{n,\text{exp}}^{ips}$, IPW estimator based on IPS estimator (12); $\widehat{\beta}_n^{cbps}$, IPW estimator based on CBPS estimator with moment equation (2), with $f(\mathbf{X}) = \mathbf{X}$; $\widehat{\beta}_n^{mle}$, IPW estimator based on MLE.

In order to assess if this is the case, we use Sant'Anna and Song (2018) specification test but fail to reject the null hypothesis of the PS model being correctly specified. Given the lack of evidence of PS misspecification, an alternative reason for these conflicting results is that the unconfoundedness assumption does not hold in this particular application. Although such an assumption is not directly testable, the sensitivity analysis in Masten and Poirier (2017) suggests that this may be the case.

# 6 Discussion and Extensions

In this article, we proposed a framework to estimate propensity score parameters such that, instead of targeting to balance only some specific moments of covariates, it aims to balance *all* functions of covariates. The proposed estimator is of the minimum distance type, and is data-driven, $\sqrt{n}$-consistent, asymptotically normal, and admits an asymptotic linear representation that facilitates the study of inverse probability weighted estimators in a unified manner. We derived the large sample properties of average, distributional and quantile treatment effects estimator based on the proposed integrated propensity scores, and illustrated its attractive properties via a Monte Carlo study and three empirical applications.

Our results can be extended to other situations of practical interest. For instance, the results of Section 2.2 and 3 can be readily extended to multivalued treatments. Following Imbens (2000), we can define the generalized propensity score $P(D = d|\mathbf{X})$, $d = \{0, 1, \ldots, K\}$, $K \geq 2$, and as in the binary case, we can exploit that, for each $d \in \{0, 1, \ldots, K\}$, $\mathbb{E}[f(\mathbf{X}) \cdot 1\{D = d\}/P(D = d|\mathbf{X})] = \mathbb{E}[f(\mathbf{X})]$ for every measurable, integrable function $f(\cdot)$. As discussed in Section 2.2, we can characterize these moments as an infinite number of unconditional moment restrictions and combine them to form a minimum distance estimator in the same spirit of (5).

Although we focused on policy evaluation parameters identified under the strong ignorability setup, we can extend our analysis to allow for treatment assignment non-compliance, i.e., to the local treatment effect setup introduced by Imbens and Angrist (1994). In this context, a binary instrument variable $Z$ is available and the instrument propensity score $P(Z = d|\mathbf{X})$, $d \in \{0, 1\}$, can be used to identify local treatment effect measures; see e.g. Abadie (2003), Tan (2006), and Frölich and Melly (2013). Given the binary nature of $Z$, we can estimate $P(Z = d|\mathbf{X})$ using our proposed IPS procedure. With this estimator in hand, one can then proceed with plug-in estimators analogous to those discussed in Sections 3.2 and 3.3.

Another possible extension is to use our IPS estimator to propose doubly robust estimators with improved properties; for a recent overview, see e.g. Słoczyński and Wooldridge (2018). For instance, on top of using IPS, one can also use the minimum distance approach to estimate the outcome regression (OR) model. Then, by building on Robins et al. (1994) and Scharfstein et al. (1999), we can form new doubly robust estimators by appropriately combining the PS and OR minimum distance estimators. These extensions are currently being explored by us.

# References

Abadie, A. (2002), "Bootstrap tests for distributional treatment effects in instrumental variable models," *Journal of the American Statistical Association*, 97(457), 284–292.

Abadie, A. (2003), "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113, 231–263.

Barsky, R., Bound, J., Charles, K. K., and Lupton, J. P. (2002), "Accounting for the black-white wealth gap: A nonparametric approach," *Journal of the American Statistical Association*, 97(459), 663–673.

Bierens, H. J. (1982), "Consistent model specification tests," *Journal of Econometrics*, 20(1982), 105–134.

Bierens, H. J. (1990), "A consistent conditional moment test of functional form," *Econometrica*, 58(6), 1443–1458.

Billingsley, P. (1995), *Probability and Measure*, New York: Wiley and Sons.

Bitler, M. O., Gelbach, J. B., and Hoynes, H. H. (2006), "What mean impacts miss: Distributional effects of welfare reform experiments," *The American Economic Review*, 96(4), 988–1012.

Blattman, C., and Annan, J. (2010), "The consequences of child soldiering," *Review of Economics and Statistics*, 92(4), 882–898.

Busso, M., Dinardo, J., and McCrary, J. (2014), "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *The Review of Economics and Statistics*, 96(5), 885–895.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016), "Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3), 673–700.

Chernozhukov, V., Fernández-Val, I., Melly, B., and Wüthrich, K. (2017), "Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes," *Arxiv preprint arXiv:1608.05142*, pp. 1–30.

Connors, A. F., Speroff, T., Dawson, N., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, W. J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. A. (1996), "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically III Patients," *Journal of the American Medical Association*, 276(11), 889–897.

Dehejia, R., and Wahba, S. (2002), "Propensity score-matching methods for nonexperimental causal studies," *The Review of Economics and Statistics*, 84(1), 151–161.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996), "Labor Market Institutions and the Distribution of Wages , 1973-1992 : A Semiparametric Approach," *Econometrica*, 64(5), 1001–1044.

Dominguez, M. A., and Lobato, I. N. (2004), "Consistent Estimation of Models Defined by Conditional Moment Restrictions," *Econometrica*, 72(5), 1601–1615.

Domínguez, M. A., and Lobato, I. N. (2015), "A Simple Omnibus Overidentification Specification Test for Time Series Econometric Models," *Econometric Theory*, 31(04), 891–910.

Escanciano, J. C. (2006*a*), "A consistent diagnostic test for regression models using projections," *Econometric Theory*, 22, 1030–1051.

Escanciano, J. C. (2006*b*), "Goodness-of-Fit Tests for Linear and Nonlinear Time Series Models," *Journal of the American Statistical Association*, 101(474), 531–541.

Escanciano, J. C. (2018), "A simple and robust estimator for linear regression models with strictly exogenous instruments," *Econometrics Journal*, 21(1), 36–54.

Firpo, S. (2007), "Efficient semiparametric estimation of quantile treatment effects," *Econometrica*, 75(1), 259–276.

Firpo, S., and Pinto, C. (2016), "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures," *Journal of Applied Econometrics*, 31(3), 457–486.

Frölich, M., and Melly, B. (2013), "Unconditional Quantile Treatment Effects Under Endogeneity," *Journal of Business & Economic Statistics*, 31(3), 346–357.

González-Manteiga, W., and Crujeiras, R. M. (2013), "An updated review of Goodness-of-Fit tests for regression models," *Test*, 22(3), 361–411.

Graham, B. S., De Xavier Pinto, C. C., and Egel, D. (2012), "Inverse Probability Tilting for Moment Condition Models with Missing Data," *The Review of Economic Studies*, 79(3), 1053–1079.

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.

Hájek, J. (1971), "Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu," in *Foundations of Statistical Inference*, eds. V. P. Godambe, and D. A. Sprott, Toronto: Holt, Rinehart, and Winston.

Heckman, J. J., Ichimura, H., and Todd, P. (1997), "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," *The Review of Economic Studies*, 64(4605-654).

Hirano, K., and Imbens, G. W. (2001), "Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization," *Health Services & Outcomes Research Methodology*, 2(3), 259–278.

Ichino, A., Mealli, F., and Nannicini, T. (2008), "From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity?," *Journal of Applied Econometrics*, 23(3), 305–327.

Imai, K., and Ratkovic, M. (2014), "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 243–263.

Imbens, G. (2000), "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87(3), 706–710.

Imbens, G. W., and Angrist, J. D. (1994), "Identification and estimation of local average treatment effects," *Econometrica*, 62(2), 467–475.

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social and Biometical Sciences*, Cambridge, MA: Cambridge University Press.

Kang, J. D. Y., and Schafer, J. L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.," *Statistical Science*, 22(4), 569–573.

Kline, P. (2011), "Oaxaca-Blinder as a reweighting estimator," *American Economic Review*, 101(3), 532–537.

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46(1), 33–50.

Leeb, H., and Pötscher, B. M. (2005), "Model selection and inference: Facts and fiction," *Econometric Theory*, 21(1), 21–59.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018), "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association*, 113(521), 390–400.

Masten, M., and Poirier, A. (2017), "Inference on Breakdown Frontiers," *Arxiv preprint arXiv:1705.04765*, .

Newey, W. K., and McFadden, D. (1994), "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, Vol. 4, Amsterdam: North-Holland: Elsevier, chapter 36, pp. 2111–2245.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89(427), 846–866.

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82(398), 387–394.

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," *Journal of the American Statistical Association*, 79(387), 516–524.

Rubin, D. B. (2007), "The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials," *Statistics in Medicine*, 26(1), 20–36.

Rubin, D. B. (2008), "For objective causal inference, design trumps analysis," *Annals of Applied Statistics*, 2(3), 808–840.

Sant'Anna, P. H. C., and Song, X. (2018), "Specification Tests for the Propensity Score," *Arxiv preprint arXiv:1611.06217*, .

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94(448), 1096–1120.

Słoczyński, T., and Wooldridge, J. M. (2018), "A General Double Robustness Result for Estimating Average Treatment Effects," *Econometric Theory*, 34(1), 112–133.

Stute, W. (1997), "Nonparametric model checks for regression," *The Annals of Statistics*, 25(2), 613–641.

Tan, Z. (2006), "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 101(476), 1607–1618.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

Wong, R. K. W., and Chan, K. C. G. (2018), "Kernel-based covariate functional balancing for observational studies," *Biometrika*, 105(1), 199–213.

Zubizarreta, J. R. (2015), "Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data," *Journal of the American Statistical Association*, 110(511), 910–922.