Covariate Distribution Balance via Propensity Scores*

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

November 9, 2021

Abstract

This paper proposes new estimators for the propensity score that aim to maximize the covariate distribution balance among different treatment groups. Heuristically, our proposed procedure attempts to estimate a propensity score model by making the underlying covariate distribution of different treatment groups as close to each other as possible. Our estimators are data-driven and can be used to estimate different treatment effect parameters under different identifying assumptions, including unconfoundedness and local treatment effects. We derive the asymptotic properties of inverse probability weighted estimators for the average, distributional, and quantile treatment effects based on the proposed propensity score estimator and illustrate their finite sample performance via Monte Carlo simulations and an empirical application.

*We thank the Editor, three anonymous referees, Harold Chiang, Cristine Pinto, Yuya Sasaki, Ping Yu, and the seminar and conference participants at many institution for their valuable comments and suggestions.

†Microsoft and 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 (RCT) is commonly deemed as 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).

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 available 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), and, as a result, treatment effect estimators based of ML PS estimates can be very sensitive to model misspecifications, see, e.g., Kang and Schafer (2007).

In light of these practical issues, alternative estimation procedures that are able to resemble randomization in a closer fashion have been proposed. For instance, Graham, Pinto and Egel (2012), Hainmueller (2012), Imai and Ratkovic (2014), Zubizarreta (2015), and Zhao (2019) propose alternative estimation procedures that attempt to directly balance covariates among the treated, untreated and, combined sample. Although such methods usually lead to treatment effect estimators with improved finite sample properties, they only aim to balance *some specific* functions of covariates. However, the covariate balancing property of the PS is considerably more powerful as it implies balance not only for some particular moments but for *all* measurable, integrable functions of the covariates. Indeed, the balancing property of the propensity score resembles randomization: when the data come from a randomized control trial (RCT) with perfect compliance, the entire covariate distributions among different treatment groups are balanced and, therefore, all measurable, integrable functions of the covariates are indeed balanced.

In this paper, we propose an alternative framework for estimating the PS that fully exploits the covariate balancing property of the PS. We call the resulting PS estimator the integrated propensity score (IPS). At a conceptual level, the IPS builds on the observation that the covariate balancing property of the PS can be equivalently characterized by balancing covariate distributions, namely, by an infinite, but tractable, number of unconditional moment restrictions. Upon such an observation, we consider Cramér-von Mises-type distances between these infinite 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 estimate the unknown PS

parameters within the minimum distance framework, as in, for example, Dominguez and Lobato (2004) and Escanciano (2006*a*, 2018). We emphasize that the IPS can be used under different "research designs", including not only the unconfounded treatment assignment setup, see, e.g., Rosenbaum and Rubin (1983), Hirano, Imbens and Ridder (2003), Firpo (2007), and Chen, Hong and Tarozzi (2008), but also the "local treatment effect" setup, where selection into treatment is possibly endogenous but a binary instrumental variable is available, see, e.g., Abadie (2003), and Frölich and Melly (2013). In this latter case, the IPS aims to balance the covariates among the treated, non-treated, and overall complier subpopulations.

At the practical level, one can think of the IPS as an estimation procedure that attempts to estimate the unknown finite dimensional parameters of a PS model by making the underlying entire covariate distribution of different treatment groups as close to each other as possible. The IPS framework also acknowledges that, in practice, there are different ways to compare covariate distribution functions depending on how covariate distribution balance is measured and the norm chosen. We explicitly consider three natural ways to characterize covariate distribution balance: 1) using the covariates' joint cumulative distribution, 2) their joint characteristic function, or 3) exploiting the Cramér–Wold theorem to focus on the cumulative distribution of the one-dimensional projections of the covariates. In terms of the norm, we focus on Cramér-von Mises-type distances as they can lead to smooth criteria functions that admit a closed-form representation, allowing us to avoid using numerical integration procedures. In fact, our proposed method is easy to use as currently implemented in the new package `IPS` for `R`, available at https://github.com/pedrohcgs/IPS.

The proposed IPS enjoys several appealing properties. First, the IPS procedure guarantees that the unknown PS parameters are globally identified. This is in contrast to the traditional generalized method of moments approach based on finitely many balancing conditions, see, e.g., Hellerstein and Imbens (1999) and Dominguez and Lobato (2004). Second, even though we aim to balance an infinite number of balancing conditions, the IPS estimator does not rely on tuning parameters such as bandwidths. Third, as other inverse probability weighted (IPW) estimators, the IPS does not rely on outcome data and separates the design stage (where one estimates the propensity score) from the analysis stage (where one estimates different treatment effect measures). As advocated by Rubin (2007, 2008), this separation is useful as it simultaneously mimics RCTs and avoids potential data snooping problems. Another direct consequence of this clear separation is that one can use the IPS to estimate a variety of causal effect parameters in a relatively straightforward manner. We illustrate this flexibility by deriving the asymptotic properties of IPW estimators for average, distributional and quantile treatment effects based on the IPS, under both the unconfoundedness and the local treatment effects setups.

Related literature: Our proposal builds on different branches of the econometrics literature. For instance, this paper is related to Shaikh, Simonsen, Vytlacil and Yildiz (2009) and Sant'Anna and Song (2019), who exploit the covariate balancing of the PS to propose specification tests for a given PS model. Here, instead of checking if a given PS estimator balances the covariate distribution among different treatment groups, we propose to estimate the PS unknown parameters by maximizing the covariate balancing. The IPS estimators also build on Dominguez and Lobato (2004) and Escanciano (2006*a*, 2018), who propose generic estimation procedures for finite-dimensional parameters defined via an infinite number of unconditional moment restrictions. Upon characterizing the covariate balancing property of the PS as an infinite number

of unconditional moment restrictions, we are able to adapt their proposals to our causal inference context.

Our proposal is also related to the growing literature on weighting-based covariate balancing methods. Among this branch of the literature, the closest papers to ours are Graham et al. (2012), Imai and Ratkovic (2014), Díaz, Rau and Rivera (2015) and Fan, Imai, Liu, Ning and Yang (2016). An important difference between our proposal and theirs is that all these papers focus exclusively on average treatment effects under unconfoundedness, whereas we show that one can directly use the IPS to estimate a variety of causal parameters of interest such as average, quantile and distributional treatment effects, not only under unconfoundedness but also in settings with endogenous treatment. It is also worth stressing that Graham et al. (2012) and Imai and Ratkovic (2014) propose estimating PS by balancing some specific pre-determined moments of the covariates, and that their procedure requires one to *assume* that the propensity score parameters are uniquely (globally) identified, see, e.g., Assumption 2.1(i) in Graham et al. (2012). In practice, it is hard to verify such important condition, and when such assumption is not satisfied, inference procedures based on their proposal will in general not be valid, see, e.g. Dominguez and Lobato (2004). Our proposed IPS procedure, on the other hand, does not suffer from this drawback as it aims to balance the entire covariate distribution, i.e., our proposal is based on an infinite number of balancing conditions that fully characterize the propensity score.

In a recent working paper, Fan et al. (2016) consider the case where the number of balancing moments grows with the sample size at an appropriate rate. Although this proposal bypass the identification challenge mentioned above (see, e.g., Ai and Chen (2003) and Donald, Imbens and Newey (2003)), to implement their proposal one needs to carefully choose tuning parameters and basis functions such that the resulting balancing moments are guaranteed to be finite. They also exclusively focus on estimating average treatment effects under unconfoundedness assumptions. On the other hand, their proposal does not rely on parametric assumptions like we do. As so, we view our works as complements.

Organization of the paper: Section 2 introduces the 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. In Section 4, we discuss how one can use the IPS in the empirically relevant situation where treatment adoption is endogenous and one has access to a binary instrumental variable. Section 5 illustrates the comparative performance of the proposed method through simulations. Section 6 presents an empirical application. Section 7 concludes. Proofs, as well as additional results, are reported in the Supplementary Appendix[1].

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

1 The Supplementary Appendix is available at https://pedrohcgs.github.io/files/IPS-supplementary.pdf

treatment and untreated, 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 $F_{Y(d)}(y) = \mathbb{P}(Y(d) \leqslant y)$, and $q_{Y(d)}(\tau) = \inf\{y : F_{Y(d)}(y) \geqslant \tau\}$, respectively, where $y \in \mathbb{R}$ and $\tau \in (0,1)$. Henceforth, assume that we have a random sample $\{(Y_i, D_i, \mathbf{X}_i')'\}_{i=1}^n$ from $(Y, D, \mathbf{X}')'$, where $n \geqslant 1$ 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, denote $\mathbb{E}_n[Z] = n^{-1}\sum_{i=1}^n Z_i$.

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 causal 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., Bitler, Gelbach and Hoynes (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 that 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 strategy in policy evaluation to bypass this difficulty is to assume that selection into treatment is based on observable characteristics, and that all individuals have a positive probability of being in either the treatment or the untreated group — the so-called unconfoundedness setup, see, e.g., Rosenbaum and Rubin (1983). Formally, unconfoundedness 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

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

with $1\{\cdot\}$ the indicator function, implying that both DTE(y) and QTE(τ) can also be written as functionals of the observed data; see, e.g., Firpo (2007), and Chen et al. (2008).

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

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

where

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

Estimators for $F_{Y(d)}(y)$, $d \in \{0,1\}$, and DTE(y) are formed using an analogous strategy. For the QTE(τ), one can simply invert the estimator of $F_{Y(d)}(y)$ to estimate $q_{Y(d)}(\tau)$; see, e.g., Firpo (2007) and Chen et al.

(2008). Of course, estimators for other treatment effect measures such as the difference of Theil indexes and/or Gini coefficients can also be formed using a similar strategy, see, e.g., Firpo and Pinto (2016).

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 dimensionality, researchers routinely adopt a parametric approach. A popular choice among practitioners is the logistic model

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

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.,

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

and uses the resulting PS fitted values $p\left(\mathbf{X};\widehat{\boldsymbol{\beta}}_n^{mle}\right)$ 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 as stated in the next Lemma.

Lemma 2.1 *Let $p(\mathbf{X};\boldsymbol{\beta})$ be a parametric model for the unknown propensity score. Then, if the model is correctly specified, for all measurable and integrable function $f(\mathbf{X})$ of the covariates $\mathbf{X}$,*

$$\mathbb{E}\left[\frac{D}{p(\mathbf{X};\boldsymbol{\beta}_0)}f(\mathbf{X})\right] = \mathbb{E}\left[\frac{1-D}{1-p(\mathbf{X};\boldsymbol{\beta}_0)}f(\mathbf{X})\right] = \mathbb{E}\left[f(\mathbf{X})\right] \tag{2.1}$$

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. $f(\mathbf{X}) = \mathbf{X}$),

$$\mathbb{E}\left[\left(\frac{D}{p(\mathbf{X};\boldsymbol{\beta}_0)} - \frac{1-D}{1-p(\mathbf{X};\boldsymbol{\beta}_0)}\right)f(\mathbf{X})\right] = \mathbf{0}. \tag{2.2}$$

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

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

Note that both procedures rely on choosing a finite number of functions $f(\mathbf{X})$, though there is little to no theoretical guidance on how to choose such functions.

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 characterized in Lemma 2.1. 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 that only attempts to balance finitely many covariate moments. In these cases, one must be careful justifying inference procedures as classical tools such as Taylor expansions are harder to justify.

In this paper we aim to estimate the PS parameters $\boldsymbol{\beta}_0$ by taking advantage of all the information contained in the covariate balancing property in Lemma 2.1. As so, our procedure 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 express the infinite number of covariate balancing conditions (2.1) in terms of 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 exploit this characterization, and use the analogy principle to suggest natural estimators 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, (2.1) can be expressed as

$$\mathbb{E}\left[\mathbf{h}\left(D,\mathbf{X};\boldsymbol{\beta}_0\right)|\,\mathbf{X}\right]=\mathbf{0}\ a.s., \tag{2.3}$$

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

$$\varpi_1^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)=\frac{D}{p\left(\mathbf{X};\boldsymbol{\beta}\right)}\Big/\mathbb{E}\left[\frac{D}{p\left(\mathbf{X};\boldsymbol{\beta}\right)}\right],\ \varpi_0^{ps}\left(D,\mathbf{X};\boldsymbol{\beta}\right)=\frac{1-D}{1-p\left(\mathbf{X};\boldsymbol{\beta}\right)}\Big/\mathbb{E}\left[\frac{1-D}{1-p\left(\mathbf{X};\boldsymbol{\beta}\right)}\right].$$

That is, one can express the covariate balancing conditions (2.1) 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 express (2.3) as an infinite number of unconditional covariate balancing restrictions. That is, by appropriately choosing a parametric family of functions $\mathcal{W}=\{w(\mathbf{X};\mathbf{u}):\mathbf{u}\in\Pi\}$, one can equivalently characterize (2.1) as

$$\mathbb{E}\left[\mathbf{h}\left(D,\mathbf{X};\boldsymbol{\beta}_0\right)w(\mathbf{X};\mathbf{u})\right]=\mathbf{0}\ a.e\ in\ \mathbf{u}\in\Pi, \tag{2.4}$$

see, e.g., Lemma 1 of Escanciano (2006*b*) for primitive conditions on the family $\mathcal{W}$ such that the equivalence between (2.3) and (2.4) holds. Choices of weight w satisfying this equivalence include (a) $w(\mathbf{X};\mathbf{u})=1\{\mathbf{X}\leqslant\mathbf{u}\}$, where $\mathbf{u}\in[-\infty,\infty]^k$, $\mathbf{X}\leqslant\mathbf{u}$ is understood coordinate-wise (see, e.g., Stute, 1997, Dominguez and Lobato, 2004, 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, Bierens and Wang, 2012 and Escanciano, 2018), and (c) $w(\mathbf{X};\mathbf{u})=1\{\boldsymbol{\gamma}'\mathbf{X}\leqslant u\}$, where $\mathbf{u}=(\boldsymbol{\gamma},u)\in\mathbb{S}_k\times[-\infty,\infty]$, $\mathbb{S}_k=\left\{\boldsymbol{\gamma}\in\mathbb{R}^k:\|\boldsymbol{\gamma}\|=1\right\}$, and $\|\boldsymbol{\gamma}\|$ is the Euclidean norm of real-valued vector $\boldsymbol{\gamma}$ (see, e.g., Escanciano, 2006*a*). We call (2.4) the "integrated covariate balancing condition" because it uses the integrated (cumulative) measure of covariate balancing.

Finally, let

$$Q_w\left(\boldsymbol{\beta}\right)=\int_{\Pi}\left\|\mathbf{H}_w(\boldsymbol{\beta},\mathbf{u})\right\|^2\Psi(d\mathbf{u}),\quad\boldsymbol{\beta}\in\Theta\subset\mathbb{R}^k, \tag{2.5}$$

where $\mathbf{H}_w(\boldsymbol{\beta},\mathbf{u})=\mathbb{E}\left[\mathbf{h}\left(D,\mathbf{X};\boldsymbol{\beta}\right)w(\mathbf{X};\mathbf{u})\right]$, $\|A\|^2=A^cA$, 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 Π.

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

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}\in\Theta} Q_w(\boldsymbol{\beta}), \tag{2.6}$$

and $\boldsymbol{\beta}_0$ is the unique value such that the covariate balancing condition (2.1) is satisfied.

Lemma 2.2 *Let $\Theta \subset \mathbb{R}^k$ be the parameter space, and assume that (2.1) is satisfied for a unique $\boldsymbol{\beta}_0 \in \Theta$. Then $Q_w(\boldsymbol{\beta}) \geqslant 0$, $\forall \boldsymbol{\beta} \in \Theta$, and $Q_w(\boldsymbol{\beta}_0) = 0$ if and only if the covariate balancing condition (2.1) holds.*

Lemma 2.2 is a global identification result that characterizes $\boldsymbol{\beta}_0$ as the unique minimizer of a population minimum distance function, $Q_w(\boldsymbol{\beta})$. That is, from Lemma 2.2 we have that $\boldsymbol{\beta}_0$ is the unique PS parameter that minimizes the imbalances of all measurable and integrable functions $f(\mathbf{X})$ between the treated, untreated and the combined group. Here, it is worth mentioning that neither Graham et al. (2012) nor Imai and Ratkovic (2014) covariate balancing approach guarantee global identification of the propensity score parameters. Instead, they directly *assume* that the vector of user-selected balancing conditions uniquely identify the propensity score parameters; see, e.g., Assumption 2.1 (i) of Graham et al. (2012). In practice, however, it is hard if not impossible to verify if such condition indeed holds. In cases it does not hold, inference procedures that rely on their proposed propensity score estimator, in general, will not be valid; see, e.g., Dominguez and Lobato (2004). Lemma 2.2 shows that our propose IPS procedure avoids this important drawback.

Another implication of Lemma 2.2 is that it suggests a natural estimator for $\boldsymbol{\beta}_0$ based on the sample analogue of (2.6), namely,

$$\widehat{\boldsymbol{\beta}}^{ips}_{n,w} = \arg\min_{\boldsymbol{\beta}\in\Theta} Q_{n,w}(\boldsymbol{\beta}), \tag{2.7}$$

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

$$\varpi^{ps}_{n,1}(D,\mathbf{X};\boldsymbol{\beta}) = \frac{D}{p(\mathbf{X};\boldsymbol{\beta})} \Big/ \mathbb{E}_n\left[\frac{D}{p(\mathbf{X};\boldsymbol{\beta})}\right], \tag{2.8}$$

$$\varpi^{ps}_{n,0}(D,\mathbf{X};\boldsymbol{\beta}) = \frac{1-D}{1-p(\mathbf{X};\boldsymbol{\beta})} \Big/ \mathbb{E}_n\left[\frac{1-D}{1-p(\mathbf{X};\boldsymbol{\beta})}\right]. \tag{2.9}$$

We call $\widehat{\boldsymbol{\beta}}^{ips}_{n,w}$ the integrated propensity score estimator of $\boldsymbol{\beta}_0$.

From (2.7), one can conclude that different PS estimators that fully exploit the covariate balancing property (2.1) can be constructed by choosing different w and Ψ_n. In this article, we focus on three different combinations that are intuitive and computationally simple:

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

$$\widehat{\boldsymbol{\beta}}^{ips}_{n,\text{ind}} = \arg\min_{\boldsymbol{\beta}\in\Theta} \int_{[-\infty,\infty]^k} \|\mathbb{E}_n\left[\mathbf{h}_n(D,\mathbf{X};\boldsymbol{\beta})\, 1\{\mathbf{X} \leqslant \mathbf{u}\}\right]\|^2\ F_{n,X}(d\mathbf{u}); \tag{2.10}$$

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

$$\widehat{\boldsymbol{\beta}}^{ips}_{n,\text{proj}} = \arg\min_{\boldsymbol{\beta}\in\Theta} \int_{[-\infty,\infty]\times\mathbb{S}_k} \left\|\mathbb{E}_n\left[\mathbf{h}_n(D,\mathbf{X};\boldsymbol{\beta})\, 1\{\boldsymbol{\gamma}'\mathbf{X} \leqslant u\}\right]\right\|^2\ F_{n,\boldsymbol{\gamma}}(du)d\boldsymbol{\gamma}\,; \tag{2.11}$$

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

normal distribution, leading to the IPS estimator

$$\widehat{\boldsymbol{\beta}}^{ips}_{n,\exp} = \arg\min_{\boldsymbol{\beta}\in\Theta} \int_{\mathbb{R}^k} \left\| \mathbb{E}_n \left[\mathbf{h}_n \left(D, \mathbf{X}; \boldsymbol{\beta} \right) \exp(i\mathbf{u}'\Phi \left(\mathbf{X} \right)) \right] \right\|^2 \frac{\exp\left(-\frac{1}{2}\mathbf{u}'\mathbf{u}\right)}{(2\pi)^{k/2}} d\mathbf{u}. \tag{2.12}$$

The estimators (2.10)-(2.12) build on Dominguez and Lobato (2004), Escanciano (2006*a*), and Escanciano (2018), respectively. Despite the apparent differences, they all aim to minimize covariate distribution imbalances: (2.10) aims to directly minimize imbalances of the joint distribution of covariates; (2.11) exploits the Cramér-Wold theorem and focuses on minimizing imbalances of the distribution of all one-dimensional projections of covariates; and (2.12) focuses on minimizing 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), (2.10)-(2.12) are indeed intrinsically related. Furthermore, we emphasize that neither w nor Ψ_n plays the role of a bandwidth as they do not affect the convergence rate of the IPS estimator.

From the computational perspective, (2.10)-(2.12) are easy to estimate because they do not involve matrix inversion nor nonparametric estimation. In the Supplementary Appendix S3, we show that the (real-valued) objective functions in (2.10)-(2.12) 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 implemented in the new package `IPS` for `R`, available at https://github.com/pedrohcgs/IPS.

Remark 2.1 It is important to stress that the covariate balancing property (2.1) 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, Fortin and Lemieux (1996), and Kline (2011). In addition, as we discuss in Section 4, the same principle can be used to balance the covariate distributions among the treated and non-treated complier subpopulations.

Remark 2.2 It is interesting to compare (2.2) with (2.4) beyond the fact that (2.4) is based on infinitely many balancing conditions whereas (2.2) is not. First, note that (2.4) is based on normalized (or stabilized) weights whereas (2.2) is not. We prefer to use stabilized weights as treatment effect estimators based on them usually have improved finite sample properties (see, e.g., Millimet and Tchernis, 2009 and Busso, Dinardo and McCrary, 2014). Second, note that (2.4) implies a three-way balance (treated, untreated and combined groups), whereas (2.2) only imposes a two-way balance (treated and untreated). We note that (2.2) can lead to relatively smaller/larger PS estimates as a "close to zero" PS estimate in the treated group can be offset by a "close to one" PS estimate in the untreated group. By using (2.4), such a potential drawback is avoided.

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 $\widehat{\boldsymbol{\beta}}^{ips}_{n,w}$. 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. Although our proposal can also be used to

estimate other treatment effects of interest such as those discussed in Firpo and Pinto (2016), 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 $\dot{\mathbf{H}}_w(\boldsymbol{\beta},\mathbf{u}) = \left(\dot{\mathbf{H}}'_{1,w}(\boldsymbol{\beta},\mathbf{u}), \dot{\mathbf{H}}'_{0,w}(\boldsymbol{\beta},\mathbf{u})\right)'$, a $2\times k$ matrix, where, for $d\in\{0,1\}$, $\dot{\mathbf{H}}_{d,w}(\boldsymbol{\beta},\mathbf{u}) = \mathbb{E}\left[\dot{\mathbf{h}}_d(D,\mathbf{X};\boldsymbol{\beta})\, w(\mathbf{X};\mathbf{u})\right]$, with $\dot{\mathbf{h}}_1$ and $\dot{\mathbf{h}}_0$ being the $1\times k$ vectors defined as

$$\dot{\mathbf{h}}_1(D,\mathbf{X};\boldsymbol{\beta}) = -\frac{\varpi_1^{ps}(D,\mathbf{X};\boldsymbol{\beta})}{p(\mathbf{X};\boldsymbol{\beta})}\dot{p}(\mathbf{X};\boldsymbol{\beta})' + \varpi_1^{ps}(D,\mathbf{X};\boldsymbol{\beta})\cdot\mathbb{E}\left[\frac{\varpi_1^{ps}(D,\mathbf{X};\boldsymbol{\beta})}{p(\mathbf{X};\boldsymbol{\beta})}\dot{p}(\mathbf{X};\boldsymbol{\beta})'\right],$$
$$\dot{\mathbf{h}}_0(D,\mathbf{X};\boldsymbol{\beta}) = \frac{\varpi_0^{ps}(D,\mathbf{X};\boldsymbol{\beta})}{1-p(\mathbf{X};\boldsymbol{\beta})}\dot{p}(\mathbf{X};\boldsymbol{\beta})' - \varpi_0^{ps}(D,\mathbf{X};\boldsymbol{\beta})\cdot\mathbb{E}\left[\frac{\varpi_0^{ps}(D,\mathbf{X};\boldsymbol{\beta})}{1-p(\mathbf{X};\boldsymbol{\beta})}\dot{p}(\mathbf{X};\boldsymbol{\beta})'\right],$$

and $\dot{p}(\cdot;\boldsymbol{\beta}) = \partial p(\cdot;\boldsymbol{b})/\partial\boldsymbol{b}|_{\boldsymbol{b}=\boldsymbol{\beta}}$, the $k\times 1$ vector of scores of the PS model $p(\cdot;\boldsymbol{\beta})$. We make the following set of assumptions.

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

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

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

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

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

(iii) $\mathcal{W}_{exp}\equiv\{\mathbf{x}\in\mathcal{X}\mapsto\exp(i\mathbf{u}'\Phi(\mathbf{x}))\,:\,\mathbf{u}\in\Pi\}$*, and* $\Psi_n(\mathbf{u}) = \Psi(\mathbf{u})$*, where* Π *is any 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, we are in a parametric setup. Assumption 2(ii) imposes that the parametric PS is bounded from above and from below. This assumption can be relaxed by assuming that $(D/p(\mathbf{X};\boldsymbol{\beta}),(1-D)/(1-p(\mathbf{X};\boldsymbol{\beta})))'\leqslant\mathbf{b}(\mathbf{X})$ such that $\mathbb{E}\left[\|\mathbf{b}(\mathbf{X})\|^2\right]<\infty$. Assumptions 2$(iii)$-$(iv)$ impose additional smoothness conditions on the PS, whereas Assumption 2(v) (together with Assumption 3) 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 uniformly bounded by an integrable function.

Assumption 3 restricts our attention to the IPS estimators (2.10)-(2.12). As mentioned before, we focus on such estimators because of their computational simplicity and transparency. Nonetheless, other types of IPS estimators can also be formed, provided that the weighting function w and integrating measure Ψ_n satisfy some high-level regularity conditions.

The next theorem characterizes the asymptotic properties of the IPS estimators $\widehat{\boldsymbol{\beta}}^{ips}_{n,w}$. Define the $k \times k$ (real-valued) matrix

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

and the $k \times 1$ (real-valued) vector

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

Theorem 3.1 *Under Assumptions 2 - 3, as $n \to \infty$,*

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

Furthermore, provided that the matrix $C_{w,\Psi}$ is positive definite,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{ips}_{n,w} - \boldsymbol{\beta}_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_{w,\Psi}\left(D_i, \mathbf{X}_i; \boldsymbol{\beta}_0\right) + o_p\left(1\right), \tag{3.1}$$

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

From Theorem 3.1, we conclude that the proposed IPS estimator is consistent, admits an asymptotic linear representation with influence function $l_{w,\Psi}\left(D, \mathbf{X}; \boldsymbol{\beta}_0\right)$, and converges to a normal distribution. The asymptotic linear representation (3.1) plays a major role in establishing the asymptotic properties of causal parameters such as average, distributional, and quantile treatment effects; see Section 3.2.

Remark 3.1 Although the results in Theorem 3.1 focus on the case where the propensity score is correctly specified, it is not difficult to show that the IPS estimators are still consistent when the model is locally misspecified, i.e., when $\mathbb{E}\left[\mathbf{h}\left(D, \mathbf{X}; \boldsymbol{\beta}_0\right) | \mathbf{X}\right] = n^{-1/2} \cdot \mathbf{s}\left(\mathbf{X}\right)$ a.s., for some integrable function $\mathbf{s}\left(\mathbf{X}\right)$. In this case, $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{ips}_{n,w} - \boldsymbol{\beta}_0\right)$ would still be asymptotically normal, with a mean given by

$$-C^{-1}_{w,\Psi} \cdot \int_{\Pi} \left(\dot{\mathbf{H}}_w(\boldsymbol{\beta}_0, \mathbf{u})^c \, \mathbf{S}_w\left(\mathbf{u}\right) + \dot{\mathbf{H}}'_w(\boldsymbol{\beta}_0, \mathbf{u}) \left(\mathbf{S}_w\left(\mathbf{u}\right)'\right)^c \right) \Psi(d\mathbf{u}),$$

where $\mathbf{S}_w\left(\mathbf{u}\right) = \mathbb{E}\left[\mathbf{s}\left(\mathbf{X}\right) w(\mathbf{X}; \mathbf{u})\right]$, and variance given by $\Omega^{ips}_{w,\Psi}$; see, e.g., Remark 1 in Escanciano (2006*a*), and Propositions 3 and 4 in Dominguez and Lobato (2015). Based on these results, one can compute the local bias of IPW estimators for different causal parameters. We omit such derivations for the sake of brevity.

3.2 Estimating treatment effects under unconfoundedness

In this section, we illustrate how one can estimate and make asymptotically valid inference about average, distributional, and quantile treatment effects under the unconfoundedness assumption 1 using IPW estimators based on the IPS estimator $\widehat{\boldsymbol{\beta}}^{ips}_{n,w}$.

Based on the discussion in Section 2.1, the IPW estimators for ATE, DTE and QTE are respectively:

$$\widehat{ATE}_n^{ips} = \mathbb{E}_n\left[\left(\varpi_{n,1}^{ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right) - \varpi_{n,0}^{ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right)\right)Y\right], \tag{3.2}$$

$$\widehat{DTE}_n^{ips}(y) = \mathbb{E}_n\left[\left(\varpi_{n,1}^{ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right) - \varpi_{n,0}^{ps}\left(D,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{ips}\right)\right)1\{Y \leqslant y\}\right], \tag{3.3}$$

$$\widehat{QTE}_n^{ips}(\tau) = \widehat{q}_{n,Y(1)}^{ips}(\tau) - \widehat{q}_{n,Y(0)}^{ips}(\tau), \tag{3.4}$$

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

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

with $\rho_\tau(a) = a\cdot(\tau - 1\{a \leqslant 0\})$ the check function as in Koenker and Bassett (1978), and the weights $\varpi_{n,1}^{ps}$ and $\varpi_{n,0}^{ps}$ are as in (2.8) and (2.9).

To derive the asymptotic properties of (3.2)-(3.4), we need to make an additional assumption about the underlying distributions of the potential outcomes $Y(1)$ and $Y(0)$.

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

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

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

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 only plays a role in the analysis of quantile treatment effects. In principle, Assumption 4(iii) can be relaxed at the cost of using more complex arguments, see Chernozhukov, Fernández-Val, Melly and Wüthrich (2019) for details.

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

$$\psi_{w,\Psi}^{ate}(Y,D,\mathbf{X}) = g^{ate}(Y,D,\mathbf{X}) - l_{w,\Psi}(D,\mathbf{X};\boldsymbol{\beta}_0)'\cdot\mathbf{G}_{\boldsymbol{\beta}}^{ate}, \tag{3.5}$$

$$\psi_{w,\Psi}^{dte}(Y,D,\mathbf{X};y) = g^{dte}(Y,D,\mathbf{X};y) - l_{w,\Psi}(D,\mathbf{X};\boldsymbol{\beta}_0)'\cdot\mathbf{G}_{\boldsymbol{\beta}}^{dte}(y), \tag{3.6}$$

$$\psi_{w,\Psi}^{qte}(Y,D,\mathbf{X};\tau) = -\left(g^{qte}(Y,D,\mathbf{X};\tau) - l_{w,\Psi}(D,\mathbf{X};\boldsymbol{\beta}_0)'\cdot\mathbf{G}_{\boldsymbol{\beta}}^{qte}(\tau)\right) \tag{3.7}$$

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

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

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

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

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 and QTE 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 $\Omega_{w,\Psi,\tau}^{qte} = \mathbb{E}\left[\psi_{w,\Psi}^{qte}(Y,D,\mathbf{X};\tau)^2\right]$.

Theorem 3.2 *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 3.2 indicates that one can use our proposed IPS estimator to estimate a variety of causal parameters that are able to highlight treatment effect heterogeneity[2]. Furthermore, Theorem 3.2 also suggests that 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 arguments. We omit the details for the sake of brevity.

Remark 3.2 In Supplementary Appendix S8, we show that results analogous to Theorem 3.2 also hold for the average, distributional and quantile treatment effect on the treated. These treatment effects parameters can have higher policy relevancy in setups where the policy intervention is directed at individuals with certain characteristics, e.g., when a clinical treatment is directed to units with a specific symptoms; see e.g., Heckman, Ichimura and Todd (1997).

4 The IPS when treatment is endogenous

In many important applications, the assumption that treatment adoption is exogenous may be too restrictive. For instance, when individuals do not comply with their treatment assignment, or more generally when they sort into treatment based on expected gains, Assumption 1 is likely to be violated. Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) point out that when this is the case and a binary instrument (Z) for the selection into treatment is available, one can only nonparametrically identify treatment effect measures for the subpopulation of compliers, that is, individuals who comply with their actual assignment of treatment, and would have complied with the alternative assignment. As shown by Abadie (2003), Frölich (2007), and Frölich and Melly (2013), the instrument propensity score $q(\mathbf{X}) \equiv \mathbb{P}(Z = 1|X)$ plays a prominent role in this local treatment effect (LTE) setup. In this section, we show that one can use the IPS approach

2 Although the results stated in Theorem 3.2 for distribution and quantile treatment effects are pointwise, in Appendix S6 we prove their uniform counterparts using empirical process techniques. We omit the details in the main text only to avoid additional cumbersome notation. We refer interested readers to the proof of Theorem 3.2 in Appendix S6 for additional details.

to estimate the instrument propensity score $q(\mathbf{X})$, by maximizing covariate distribution balancing among different instrument-by-treatment subgroups.

Before providing the details about how we apply the IPS approach to estimate $q(\mathbf{X})$ under the LTE setup, we introduce a brief description of the LTE setup. Let Z be a binary instrumental variable Z for the treatment assignment. Denote $D(0)$ and $D(1)$ the value that D would have taken if Z is equal to zero or one, respectively. The realized treatment is $D = ZD(1) + (1-Z)D(0)$. Thus, the observed sample in the LTE setup consists of independent and identically distributed copies $\left\{\left(Y_i, D_i, Z_i, \mathbf{X}_i'\right)'\right\}_{i=1}^{n}$. To identify the average, distributional and quantile treatment effects for the compliers, we follow Abadie (2003) and make the following assumption.

Assumption 5 *(i)* $(Y(0), Y(1), D(0), D(1)) \perp\!\!\!\perp Z|\mathbf{X}$*; (ii) for some* $\varepsilon > 0$, $\varepsilon \leqslant q(\mathbf{X}) \leqslant 1-\varepsilon$ *a.s. and* $\mathbb{P}(D(1)=1|\mathbf{X}) > \mathbb{P}(D(0)=1|\mathbf{X})$ *a.s.; and (iii)* $\mathbb{P}(D(1) \geqslant D(0)|\mathbf{X}) = 1$ *a.s.*.

Assumption 5(i) imposes that, once we condition on $\mathbf{X}$, Z is "as good as randomly assigned". Assumption 5(ii) imposes a common support condition, and guarantees that, conditional on $\mathbf{X}$, Z is a relevant instrument for D. Finally, Assumption 5(iii) is a monotonicity condition that rules out the existence of defiers.

From Abadie (2003) and Frölich and Melly (2013), we have that under Assumption 5, the average, distributional and quantile treatment effects for compliers are nonparametrically identified, i.e.,

$$\begin{aligned}
LATE &\equiv \mathbb{E}[Y(1) - Y(0)|\mathcal{C}] = \mathbb{E}\left[\varpi_1^{lte}(D, Z, \mathbf{X}; q) \cdot Y\right] - \mathbb{E}\left[\varpi_0^{lte}(D, Z, \mathbf{X}; q) \cdot Y\right], \\
LDTE(y) &\equiv \mathbb{P}(Y(1) \leqslant y|\mathcal{C}) - \mathbb{P}(Y(0) \leqslant y|\mathcal{C}) = F_{\varpi_1^{lte}.Y}(y) - F_{\varpi_0^{lte}.Y}(y), \\
LQTE(\tau) &\equiv q_{Y(1)|\mathcal{C}}(\tau) - q_{Y(0)|\mathcal{C}}(\tau) = F^{-1}_{\varpi_1^{lte}.Y}(\tau) - F^{-1}_{\varpi_0^{lte}.Y}(\tau),
\end{aligned}$$

where $\mathcal{C}$ denotes the complier subpopulation, and, for $d \in \{0, 1\}$,

$$\varpi_d^{lte}(D, Z, \mathbf{X}; q) = \frac{1\{D=d\}}{\kappa_d(q)}\left(\frac{Z}{q(\mathbf{X})} - \frac{(1-Z)}{1-q(\mathbf{X})}\right), \tag{4.1}$$

$$F_{\varpi_d^{lte}.Y}(y) = \mathbb{E}\left[\varpi_d^{lte}(D, Z, \mathbf{X}; q) \cdot 1\{Y \leqslant y\}\right], \tag{4.2}$$

and

$$\kappa_d(q) \equiv \mathbb{E}\left[\frac{1\{D=d\}Z}{q(\mathbf{X})} - \frac{1\{D=d\}(1-Z)}{1-q(\mathbf{X})}\right],$$

and $F^{-1}_{\varpi_d^{lte}.Y}(\tau) = \inf\left\{y : F_{\varpi_d^{lte}.Y}(y) \geqslant \tau\right\}$. From the above results, it is clear that the instrument PS plays a prominent role in the LTE setup, and that once we have an estimator for q available, it is relatively straightforward to construct estimators for the LATE, LDTE, and LQTE.

To estimate the instrument PS $q(\cdot)$, we adopt a parametric approach, i.e., we assume that $q(\mathbf{X}) = q\left(\mathbf{X}; \boldsymbol{\beta}_0^{lte}\right)$, where q is known up to the finite-dimensional parameters $\boldsymbol{\beta}_0^{lte}$. Here, as we are interested in treatment effects for the (latent) subpopulation of compliers, we will attempt to estimate $\boldsymbol{\beta}_0^{lte}$ by maximizing the covariate distribution balance among compliers. To do so, we build on Theorem 3.1 of Abadie (2003), which establishes that, for every measurable and integrable function $f(\mathbf{X})$ of the covariates $\mathbf{X}$,

$$\begin{aligned}
\mathbb{E}\left[\varpi_1^{lte}\left(D, Z, \mathbf{X}; \boldsymbol{\beta}_0^{lte}\right) \cdot f(\mathbf{X})\right] &= \mathbb{E}\left[\varpi^{lte}\left(D, Z, \mathbf{X}; \boldsymbol{\beta}_0^{lte}\right) \cdot f(\mathbf{X})\right], \\
\mathbb{E}\left[\varpi_0^{lte}\left(D, Z, \mathbf{X}; \boldsymbol{\beta}_0^{lte}\right) \cdot f(\mathbf{X})\right] &= \mathbb{E}\left[\varpi^{lte}\left(D, Z, \mathbf{X}; \boldsymbol{\beta}_0^{lte}\right) \cdot f(\mathbf{X})\right],
\end{aligned} \tag{4.3}$$

where $\varpi_d^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}_0^{lte}\right)$ is defined as in (4.1) but with $\boldsymbol{\beta}_0^{lte}$ playing the role of q, as we assume that q is a parametric model, and

$$\varpi^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)=\frac{1}{\kappa\left(\boldsymbol{\beta}\right)}\left(1-\frac{\left(1-D\right)Z}{q\left(\mathbf{X};\boldsymbol{\beta}\right)}-\frac{D\left(1-Z\right)}{1-q\left(\mathbf{X};\boldsymbol{\beta}\right)}\right),$$

with

$$\kappa\left(\boldsymbol{\beta}\right)\equiv\mathbb{E}\left[1-\frac{\left(1-D\right)Z}{q\left(\mathbf{X};\boldsymbol{\beta}\right)}-\frac{D\left(1-Z\right)}{1-q\left(\mathbf{X};\boldsymbol{\beta}\right)}\right].$$

As noted in Theorem 3.1 of Abadie (2003), under Assumption 5, $\mathbb{E}\left[\varpi^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}_0^{lte}\right)\cdot f\left(\mathbf{X}\right)\right]=\mathbb{E}\left[f\left(\mathbf{X}\right)|\mathcal{C}\right]$, implying that (4.3) are indeed balancing conditions for the complier subpopulation.

Next and analogously to the discussion in Section 2.2, we rewrite (4.3) as

$$\mathbf{H}_w^{lte}(\boldsymbol{\beta}_0^{lte},\mathbf{u})=\mathbf{0}\ a.e\ in\ \mathbf{u}\in\Pi, \tag{4.4}$$

where $\mathbf{H}_w^{lte}(\boldsymbol{\beta},\mathbf{u})=\mathbb{E}\left[\mathbf{h}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)w(\mathbf{X};\mathbf{u})\right]$, with $\mathbf{h}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)=(h_1^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right),$ $h_0^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right))'$, and, for $d\in\{0,1\}$, $h_d^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)=\varpi_d^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)-\varpi^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)$.

Based on (4.4), we then show in Lemma S7.1 in the Supplementary Appendix that $\boldsymbol{\beta}_0^{lte}$ is globally identified, i.e., $\boldsymbol{\beta}_0^{lte}$ is the unique minimizer of the population minimum distance criteria $Q_w^{lte}(\boldsymbol{\beta})=\int_\Pi\left\|\mathbf{H}_w^{lte}(\boldsymbol{\beta},\mathbf{u})\right\|^2\Psi(d\mathbf{u})$. Thus, like in the case where treatment is exogenous, we can fully exploit the balancing conditions (4.3) and estimate $\boldsymbol{\beta}_0^{lte}$ by

$$\widehat{\boldsymbol{\beta}}_{n,w}^{lips}=\arg\min_{\boldsymbol{\beta}\in\Theta}\int_\Pi\left\|\mathbf{H}_{n,w}^{lte}(\boldsymbol{\beta},\mathbf{u})\right\|^2\Psi_n(d\mathbf{u}), \tag{4.5}$$

where Ψ_n is a uniformly consistent estimator of Ψ, $\mathbf{H}_{n,w}^{lte}(\boldsymbol{\beta},\mathbf{u})=\mathbb{E}_n\left[\mathbf{h}_n^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)w(\mathbf{X};\mathbf{u})\right]$, $\mathbf{h}_n^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)=(h_{n,1}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right),\ h_{n,0}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right))'$, $h_{n,d}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)=\varpi_{n,d}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)-\varpi_n^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)$, and

$$\begin{aligned}
\varpi_{n,d}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)&=\frac{1\left\{D=d\right\}}{\kappa_{n,d}\left(\boldsymbol{\beta}\right)}\left(\frac{Z}{q\left(\mathbf{X};\boldsymbol{\beta}\right)}-\frac{\left(1-Z\right)}{1-q\left(\mathbf{X};\boldsymbol{\beta}\right)}\right)\\
\kappa_{n,d}\left(\boldsymbol{\beta}\right)&=\mathbb{E}_n\left[1\left\{D=d\right\}\left(\frac{Z}{q\left(\mathbf{X};\boldsymbol{\beta}\right)}-\frac{1-Z}{1-q\left(\mathbf{X};\boldsymbol{\beta}\right)}\right)\right]\\
\varpi_n^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)&=\frac{1}{\kappa_n\left(\boldsymbol{\beta}\right)}\left(1-\frac{\left(1-D\right)Z}{q\left(\mathbf{X};\boldsymbol{\beta}\right)}-\frac{D\left(1-Z\right)}{1-q\left(\mathbf{X};\boldsymbol{\beta}\right)}\right),\\
\kappa_n\left(\boldsymbol{\beta}\right)&=\mathbb{E}_n\left[1-\frac{\left(1-D\right)Z}{q\left(\mathbf{X};\boldsymbol{\beta}\right)}-\frac{D\left(1-Z\right)}{1-q\left(\mathbf{X};\boldsymbol{\beta}\right)}\right].
\end{aligned} \tag{4.6}$$

As before, we focus our attention on the three weighting functions described in Assumption 3. We call (4.5) the local integrated propensity score (LIPS) estimator.

In what follows, we derive the asymptotic properties of the instrument IPS estimator $\widehat{\boldsymbol{\beta}}_{n,w}^{lips}$. Let the score of $\mathbf{H}_w^{lte}(\boldsymbol{\beta},\mathbf{u})$ be defined as $\dot{\mathbf{H}}_w^{lte}(\boldsymbol{\beta},\mathbf{u})=\left(\dot{\mathbf{H}}_{1,w}^{lte\prime}(\boldsymbol{\beta},\mathbf{u}),\dot{\mathbf{H}}_{0,w}^{lte\prime}(\boldsymbol{\beta},\mathbf{u})\right)'$ where, for $d\in\{0,1\}$, $\dot{\mathbf{H}}_{d,w}^{lte}(\boldsymbol{\beta},\mathbf{u})=\mathbb{E}\left[\dot{\mathbf{h}}_d^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)w(\mathbf{X};\mathbf{u})\right]$,

$$\dot{\mathbf{h}}_d^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)=\dot{\boldsymbol{\varpi}}_d^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)-\dot{\boldsymbol{\varpi}}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right),$$

with

$$\dot{\boldsymbol{\varpi}}_d^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)=-\frac{1\left\{D=d\right\}}{\kappa_d\left(\boldsymbol{\beta}\right)}\left(\frac{Z}{q\left(\mathbf{X};\boldsymbol{\beta}\right)^2}+\frac{\left(1-Z\right)}{\left(1-q\left(\mathbf{X};\boldsymbol{\beta}\right)\right)^2}\right)\cdot\dot{q}\left(\mathbf{X};\boldsymbol{\beta}\right)'$$

$$+\varpi_d^{lte}(D,Z,\mathbf{X};\boldsymbol{\beta})\cdot\mathbb{E}\left[\frac{1\{D=d\}}{\kappa_d(\boldsymbol{\beta})}\left(\frac{Z}{q(\mathbf{X};\boldsymbol{\beta})^2}+\frac{(1-Z)}{(1-q(\mathbf{X};\boldsymbol{\beta}))^2}\right)\cdot\dot{q}(\mathbf{X};\boldsymbol{\beta})'\right],$$

and

$$\dot{\boldsymbol{\varpi}}^{lte}(D,Z,\mathbf{X};\boldsymbol{\beta})=-\frac{1}{\kappa(\boldsymbol{\beta})}\left(\frac{D(1-Z)}{(1-q(\mathbf{X};\boldsymbol{\beta}))^2}-\frac{(1-D)Z}{q(\mathbf{X};\boldsymbol{\beta})^2}\right)\cdot\dot{q}(\mathbf{X};\boldsymbol{\beta})'$$
$$+\varpi^{lte}(D,Z,\mathbf{X};\boldsymbol{\beta})\cdot\mathbb{E}\left[\frac{1}{\kappa(\boldsymbol{\beta})}\left(\frac{D(1-Z)}{(1-q(\mathbf{X};\boldsymbol{\beta}))^2}-\frac{(1-D)Z}{q(\mathbf{X};\boldsymbol{\beta})^2}\right)\cdot\dot{q}(\mathbf{X};\boldsymbol{\beta})'\right],$$

and $\dot{q}(\cdot;\boldsymbol{\beta})=\partial q(\cdot;\boldsymbol{b})/\partial\boldsymbol{b}|_{\boldsymbol{b}=\boldsymbol{\beta}}$. We make the following set of assumptions, which are the analogue of Assumption 2.

Assumption 6 *(i) $q(\mathbf{x})=q(\mathbf{x};\boldsymbol{\beta}_0^{lte})$, where $\boldsymbol{\beta}_0^{lte}$ is an interior point of a compact set $\Theta\subset\mathbb{R}^k$; (ii) for some $\delta>0$, $\delta\leqslant q(\mathbf{x};\boldsymbol{\beta})\leqslant 1-\delta$ for all $\mathbf{x}\in\mathcal{X}$, $\boldsymbol{\beta}\in int(\Theta)$; (iii) with probability one, $q(\mathbf{X};\boldsymbol{\beta})$ is continuous at each $\boldsymbol{\beta}\in\Theta$; (iv) with probability one, $q(\mathbf{X};\boldsymbol{\beta})$ is continuously differentiable in a neighborhood of $\boldsymbol{\beta}_0^{lte}$, $\Theta_0^{lte}\subset\Theta$; (v) for $d\in\{0,1\}$*

$$\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in\Theta_0^{lte}}\left\|1\{D=d\}\left(\frac{Z}{q(\mathbf{X};\boldsymbol{\beta})^2}+\frac{(1-Z)}{(1-q(\mathbf{X};\boldsymbol{\beta}))^2}\right)\cdot\dot{q}(\mathbf{X};\boldsymbol{\beta})\right\|\right]<\infty.$$

The next theorem characterizes the asymptotic properties of the instrument IPS estimators $\widehat{\boldsymbol{\beta}}_{n,w}^{lips}$. Define the $k\times k$ (real-valued) matrix

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

and the $k\times 1$ (real-valued) vector

$$l_{w,\Psi}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}_0^{lte}\right)=-\left(C_{w,\Psi}^{lte}\right)^{-1}\cdot\int_\Pi\left(\dot{\mathbf{H}}_w^{lte}(\boldsymbol{\beta}_0^{lte},\mathbf{u})^c\,w(\mathbf{X};\mathbf{u})+\dot{\mathbf{H}}_w^{lte}(\boldsymbol{\beta}_0^{lte},\mathbf{u})'w(\mathbf{X};\mathbf{u})^c\right)\Psi(d\mathbf{u})$$
$$\cdot\,\mathbf{h}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}_0^{lte}\right). \quad (4.7)$$

Theorem 4.1 *Under Assumptions 3, 5, and 6, as $n\to\infty$,*

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

Furthermore, provided that the matrix $C_{w,\Psi}^{lte}$ is positive definite,

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

and

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

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

With the results of Theorem 4.1 at hand, we can estimate the LATE, LDTE, and LQTE by using the instrument IPS estimators:

$$\widehat{LATE}_n^{lips}=\mathbb{E}_n\left[\left(\varpi_{n,1}^{lte}\left(D,Z,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{lips}\right)-\varpi_{n,0}^{lte}\left(D,Z,\mathbf{X};\widehat{\boldsymbol{\beta}}_{n,w}^{lips}\right)\right)Y\right], \quad (4.8)$$

$$\widehat{LDTE}_n^{lips}(y)=\widehat{F}_{n,\varpi_1^{lte}\cdot Y}^r(y)-\widehat{F}_{n,\varpi_0^{lte}\cdot Y}^r(y), \quad (4.9)$$

$$\widehat{LQTE}_n^{lips}(\tau)=\widehat{F}_{n,\varpi_1^{lte}\cdot Y}^{r,-1}(\tau)-\widehat{F}_{n,\varpi_0^{lte}\cdot Y}^{r,-1}(\tau), \quad (4.10)$$

where, for $d \in \{0,1\}$, $\widehat{F}^{r}_{n,\varpi^{lte}_{d}\cdot Y}(\cdot)$ denotes the rearrangement of $\widehat{F}_{n,\ \varpi^{lte}_{d}\cdot Y}(\cdot)$,

$$\widehat{F}_{n,\varpi^{lte}_{d}\cdot Y}(\cdot) = \mathbb{E}_n\left[\varpi^{lte}_{n,d}\left(D,Z,\mathbf{X};\widehat{\boldsymbol{\beta}}^{lips}_{n,w}\right)1\{Y \leqslant \cdot\}\right],$$

if $\widehat{F}_{n,\varpi^{lte}_{d}\cdot Y}$ is not monotone, see, e.g., Chernozhukov, Fernández-Val and Galichon (2010), and Wüthrich (2019)[3]. Importantly, these rearrangements do not change the asymptotic properties of the estimators.

To derive the asymptotic properties of (4.8)-(4.10), we impose the following regularity conditions, which are the analogue of Assumption 4.

Assumption 7 *For $d \in \{0,1\}$, $(i)\,\mathbb{E}\left[Y(d)^2\,|\mathcal{C}\right] < M$ for some $0 < M < \infty$, (ii)*

$$\mathbb{E}\left[\sup_{\beta\in\Theta^{lte}_{0}}\left\|1\{D=d\}\left(Y(d)-\mathbb{E}[Y(d)|\,\mathcal{C}]\right)\left(\frac{Z}{q(\mathbf{X};\boldsymbol{\beta})^2}+\frac{(1-Z)}{(1-q(\mathbf{X};\boldsymbol{\beta}))^2}\right)\cdot\dot{q}(\mathbf{X};\boldsymbol{\beta})\right\|\right] < \infty,$$

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

Theorem 4.2 *Under Assumptions 3, 5 -7, for each $y \in \mathbb{R}$, $\tau \in [\varepsilon, 1-\varepsilon]$, we have that, as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{LATE}^{lips}_{n} - LATE\right) \xrightarrow{d} N\left(0,\Omega^{late}_{w,\Psi}\right),$$
$$\sqrt{n}\left(\widehat{LDTE}^{lips}_{n} - LDTE\right)(y) \xrightarrow{d} N\left(0,\Omega^{ldte}_{w,\Psi,y}\right),$$
$$\sqrt{n}\left(\widehat{LQTE}^{lips}_{n} - LQTE\right)(\tau) \xrightarrow{d} N\left(0,\Omega^{lqte}_{w,\Psi,\tau}\right),$$

where $\Omega^{late}_{w,\Psi}$, $\Omega^{ldte}_{w,\Psi,y}$ and $\Omega^{lqte}_{w,\Psi,\tau}$ are defined in the proof of Theorem 4.2 in Appendix S7.

Remark 4.1 Although the results stated in Theorem 4.2 for local distribution and quantile treatment effects are pointwise, in Appendix S7 we prove their uniform counterparts using empirical process techniques. We omit the details in the main text only to avoid additional cumbersome notation. We refer interested readers to the proof of Theorem 4.2 in Appendix S7 for additional details.

Remark 4.2 For brevity, we focused on the unconditional LATE, LDTE and LQTE causal parameters. However, one can readily use the instrument IPS discussed in this section to estimate other conditional treatment effect measures, such as the conditional local quantile treatment effects introduced by Abadie, Angrist and Imbens (2002), and the local average response functions introduced by Abadie (2003). Given the results in Theorem 4.1, establishing the asymptotic properties of these conditional treatment effect measures is relatively straightforward.

Remark 4.3 We note that under Assumption 5, when one fixes $f(\mathbf{X}) = \mathbf{X}$ and subtracts the second equality in (4.3) from the first equality in (4.3), one has that, after some straightforward manipulation,

$$\mathbb{E}\left[\left(\frac{Z}{q(\mathbf{X};\boldsymbol{\beta}^{lte}_{0})}-\frac{1-Z}{1-q(\mathbf{X};\boldsymbol{\beta}^{lte}_{0})}\right)\mathbf{X}\right] = \mathbf{0}.$$

3 Lack of monotonicity may appear in finite samples because the weights $w^{lte}_{n,d}$ can be negative. This poses problems for the inversion of the weighted cumulative distribution functions to obtain the quantile functions. On the other hand, under Assumption 5, the population weights w^{lte}_{d} are non-negative, implying that these potential problems disappear, asymptotically. As discussed in detail in Chernozhukov et al. (2010), we can bypass such challenges by monotonizing $\widehat{F}_{n,\ w^{lte}_{d}\cdot Y}$ via rearrangements.

Thus, by substituting D and $p(\mathbf{X};\boldsymbol{\beta}_0)$ in (2.2) with Z and $q\left(\mathbf{X};\boldsymbol{\beta}_0^{lte}\right)$, one can, in principle, use Imai and Ratkovic (2014)'s covariate balancing propensity score procedure to estimate the instrument propensity score. However (and analogous to the discussion in Section 2), such a procedure would only partly exploit Theorem 3.1 of Abadie (2003), which is in contrast with our proposed LIPS procedure. As a consequence, the LIPS estimation procedure can lead to estimators with improved finite-sample properties; we illustrate this point via Monte Carlo simulations in the next section.

5 Monte Carlo simulations

We present two sets of Monte Carlo experiments to study the finite sample properties of our proposed treatment effect estimators based on the IPS. In Section 5.1, we conduct simulations in a stylized design largely based on Kang and Schafer (2007). In Section 5.2, we present a set of simulations that is calibrated to our empirical application analyzing the effect of 401(k) retirement plans on asset accumulation. We consider both unconfounded and endogenous treatment setups.

Under unconfoundedness, we compare the performance of different IPW estimators for the ATE and the QTE(τ), $\tau \in \{0.10, 0.25, 0.5, 0.75, 0.9\}$, when one estimates the PS using our proposed IPS estimators (2.10)-(2.12), the classical maximum likelihood (ML) approach, Imai and Ratkovic (2014)'s just-identified covariate balancing propensity score (CBPS) as in (2.2) with $f(\mathbf{X}) = \mathbf{X}$, and Imai and Ratkovic (2014)'s over-identified CBPS (2.2) with $f(\mathbf{X}) = \left(\mathbf{X}', \dot{p}(\mathbf{X};\boldsymbol{\beta})'\right)'$, i.e., on top of balancing the means, one also makes use of the likelihood score equation. In all cases, we consider a logistic PS model. All treatment effect estimators use stabilized weights (2.8) and (2.9).

When treatment take-up is endogenous and a binary instrument Z is available, we compare the performance of different IPW estimators for the LATE and the LQTE(τ), $\tau \in \{0.10, 0.25, 0.5, 0.75, 0.9\}$, when one estimates the instrument PS $q(\cdot)$ using our proposed local IPS estimator (4.5) with exponential, indicator, and projection-based weights, the classical ML approach, Imai and Ratkovic (2014)'s just-identified and over-identified CBPS with Z playing the role of D. In all cases, we consider a logistic instrument PS model.

For each design, we conduct $1,000$ Monte Carlo simulations. We compare various IPW estimators in terms of average bias, root mean square error (RMSE), relative mean square error (relMSE), empirical 95% coverage probability, the median length of a 95% confidence interval, and the asymptotic relative efficiency (ARE)[4]. For the relative measures of performance, relMSE and ARE, we treat estimators based on the over-identified CBPS as the benchmark. The confidence intervals are based on the normal approximations in Theorems 3.2 and 4.2, with the asymptotic variances being estimated by their sample analogues — the impact of the different estimation methods used is reflected on the adjustment term for the estimation effect, i.e., the asymptotic linear representation of the different PS estimators. For the variance of QTE (LQTE) estimators, 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`. We use the `CBPS` package in `R` to estimate both CBPS estimators. Due to space constraints, the tables

4 For any parameter η of a distribution F, and for estimators $\widehat{\eta}_1$ and $\widehat{\eta}_2$ approximately $N(\eta, V_1/n)$ and $N(\eta, V_2/n)$, respectively, the asymptotic relative efficiency of $\widehat{\eta}_2$ with respect to $\widehat{\eta}_1$ is given by V_1/V_2; see, e.g., Section 8.2 in van der Vaart (1998).

containing all the results of the simulations are deferred to the Supplementary Appendices S1 and S2.

5.1 Stylized simulation

5.1.1 Unconfoundedness setup

Our stylized 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 4×4 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

$$Y(1) = 210 + m(\mathbf{X}) + \varepsilon(1), \qquad Y(0) = 200 - m(\mathbf{X}) + \varepsilon(0),$$

where $m(\mathbf{X}) = 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4$, $\varepsilon(1)$ and $\varepsilon(0)$ are independent $N(0, 1)$ random variables. The ATE and the QTE(τ) are equal to 10, for all $\tau \in (0, 1)$.

We consider two different scenarios to assess the sensitivity of the proposed estimators not only under correctly specified PS models but also under misspecified PS models that are "nearly correct". In the first experiment, we observed data on $\left(Y_i, D_i, \mathbf{X}_i'\right)$, and, therefore, all IPW estimators are correctly specified. In the second experiment, we observed data on $\left(Y_i, D_i, \mathbf{W}_i'\right)$, where $\mathbf{W} = (W_1, W_2, W_3, W_4)'$ with $W_1 = \exp(X_1/2)$, $W_2 = X_2/(1 + \exp(X_1))$, $W_3 = (X_1X_3/25 + 0.6)^3$, and $W_4 = (X_2 + X_4 + 20)^2$. In this second scenario, the IPW estimators for ATE and QTE(τ) are misspecified. We consider sample size n equal to 500 [5]. All available covariates ($\mathbf{X}$ or $\mathbf{W}$) enter the PS model linearly.

Table S1.1 displays the simulation results for both scenarios. When the PS model is correctly specified, all estimators perform well in terms of bias and coverage probability, i.e., all estimators are essentially unbiased and their associated confidence intervals have correct coverage. Comparing ML-based with CBPS-based estimators, we note that IPW estimators based on ML tend to have higher MSE, longer confidence intervals, and lower ARE. Thus, it is clear that CBPS-based IPW estimators can improve upon those based on ML. However, our simulation results under correct specification suggest that we can further improve the performance of the CBPS estimator by fully exploiting the covariate balancing property of the propensity score. For instance, the relative mean square error of the ATE estimators based on the IPS tends to be at least 10% smaller than those based on the CBPS; one can also see improvements when focusing on QTE, especially for $\tau \geqslant 0.5$. The gains in terms of ARE also tend to be large. For example, the ARE of the ATE estimator based on the IPS with projection weight function with respect to the one based on the over-identified CBPS is 1.26. This implies that the ATE estimator based on the over-identified CBPS would require $1.26 \times n$ observations to perform equivalently to the ATE estimator based on IPS with projection weight. IPS estimators based on the exponential weight also tend to dominate CBPS estimators in terms of mean square errors and ARE. Finally, we note that IPW estimators based on the IPS with the indicator function tend to give slightly larger confidence intervals than when using other IPS estimators, perhaps

5 Simulation results with $n = 200$ and $n = 1000$ are available in the Supplementary Appendix.

because there are multiple covariates (four in our simulation design), implying that many $1\{\mathbf{X}_i \leqslant \mathbf{u}\}$ are equal to zero when $\mathbf{u}$ is evaluated at the sample observations. In practice, we recommend practitioners to favor the other considered weighting functions.

When the PS model is misspecified, our Monte Carlo results suggest that the potential gains of using the IPS can also be 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 just-identified CBPS improve on ML, though under-coverage is still an unresolved issue. Estimators based on the over-identified CBPS tend to have better coverage than those based on the just-identified CBPS, but under-coverage of QTE(0.75) and QTE(0.90) is still severe, perhaps because of the large biases. Finally, in this DGP, our proposed IPS estimators can further improve upon CBPS, though this is not always the case. In terms of RMSE, the gains of adopting the IPS estimator compared to the over-identified CBPS are noticeable when one focus on ATE and the QTE with $\tau \geqslant 0.50$. When one focus on $QTE(0.10)$ or $QTE(0.25)$, though, the over-identified CBPS estimators tends to dominate the IPS ones in term of RMSE. In terms of inference, we note that, in this DGP, all considered treatment effect estimators based on IPS with the projection weighting function seems to control size; all other estimators do not share this property in this misspecified DGP.

How can the IPS estimators improve the performance of MLE and CBPS-type estimators? To tackle this question, we consider the following six measures of overall covariate distribution imbalances:

$$
\begin{aligned}
KS_{bal}(\boldsymbol{\beta}) &= \sup_{i:1,\ldots,n} \|DistImb(\mathbf{X}_i^*, \boldsymbol{\beta})\|, \qquad RCvM_{bal}(\boldsymbol{\beta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} DistImb(\mathbf{X}_i^*, \boldsymbol{\beta})^2}, \\
KS_{bal1}(\boldsymbol{\beta}) &= \sup_{i:1,\ldots,n} \|DistImb_1(\mathbf{X}_i^*, \boldsymbol{\beta})\|, \qquad RCvM_{bal1}(\boldsymbol{\beta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} DistImb_1(\mathbf{X}_i^*, \boldsymbol{\beta})^2} \\
KS_{bal0}(\boldsymbol{\beta}) &= \sup_{i:1,\ldots,n} \|DistImb_0(\mathbf{X}_i^*, \boldsymbol{\beta})\|, \qquad RCvM_{bal0}(\boldsymbol{\beta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} DistImb_0(\mathbf{X}_i^*, \boldsymbol{\beta})^2},
\end{aligned}
\tag{5.1}
$$

where

$$
\begin{aligned}
DistImb(\mathbf{x}, \boldsymbol{\beta}) &= \mathbb{E}_n\left[\left(\varpi_{n,1}^{ps}(D, \mathbf{X}^*; \boldsymbol{\beta}) - \varpi_{n,0}^{ps}(D, \mathbf{X}^*; \boldsymbol{\beta})\right) 1\{\mathbf{X}^* \leqslant \mathbf{x}\}\right], \\
DistImb_1(\mathbf{x}, \boldsymbol{\beta}) &= \mathbb{E}_n\left[\left(\varpi_{n,1}^{ps}(D, \mathbf{X}^*; \boldsymbol{\beta}) - 1\right) 1\{\mathbf{X}^* \leqslant \mathbf{x}\}\right], \\
DistImb_0(\mathbf{x}, \boldsymbol{\beta}) &= \mathbb{E}_n\left[\left(\varpi_{n,0}^{ps}(D, \mathbf{X}^*; \boldsymbol{\beta}) - 1\right) 1\{\mathbf{X}^* \leqslant \mathbf{x}\}\right]),
\end{aligned}
$$

and $\varpi_{n,1}^{ps}(D, \mathbf{X}; \boldsymbol{\beta})$ and $\varpi_{n,0}^{ps}(D, \mathbf{X}; \boldsymbol{\beta})$ are as defined in (2.8) and (2.9), respectively, and $\mathbf{X}^* = \mathbf{X}$ when the PS model is correctly specified, and $\mathbf{X}^* = \mathbf{W}$ when the PS is misspecified. Notice that all distributional covariate imbalance measures in (5.1) are measured in percentage points, and, according to Lemma 2.1, should be close to zero when the PS model is correctly specified. Here, note that the KS and $RCvM$ covariate distributional imbalance metrics are related to the Kolmogorov-Smirnov and Cramér-von Mises test statistics that are popular in specification testing.

Table S1.2 presents these six distributional balance measures when one estimates the PS parameter β using different estimation procedures. When the PS model is correctly specified, we can see that the IPS

estimators always improve upon the CBPS and ML estimators. Among the IPS estimators, the one based on indicator function is the one who shows smallest CDF imbalances, which should be no surprise as its objective function is designed to achieve that. The performance of IPS with exponential or projection weighting function are very similar. When the PS is misspecified, we see that CBPS-based estimators tend to achieve better covariate balancing than the ML-based estimator, and that the IPS estimators with indicator and exponential weighting function further improve upon them. These distributional imbalances tend to be larger when one uses the IPS estimator with projection weighting function, though. This is not in conflict with our theory since these distributional imbalances are not expected to be close to zero under model misspecification.

All in all, these simulation results reveals that one can indeed enjoy gains in precision by using our proposed IPS estimator.

5.1.2 Local Treatment Effect Setup

We now consider the setup where treatment is endogenous but one has access to a binary instrument Z, as described in Section 4. As in the unconfoundedness case, we consider sample size n equal to 500 [6], and our PS models incorporate all available covariates in a linear fashion.

The simulation design is similar to the one in Section 5.1.1. Let $\mathbf{X}$, $\mathbf{W}$, $Y(1)$, and $Y(0)$ be defined as before. The true instrument PS is given by

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

the instrument Z is generated as $Z = 1\{q(\mathbf{X}) > U_1\}$, where U_1 follows a uniform $(0,1)$ distribution. The potential treatments $D(1)$ and $D(0)$ are generated as $D(1) = 1\{p^*(Y(1) - Y(0)) > U_2\}$ and $D(0) = 0$, where U_2 follows a uniform $(0,1)$ distribution, and

$$p^*(Y(1) - Y(0)) = \frac{\exp(2 + 0.05 \cdot (Y(1) - Y(0)))}{1 + \exp(2 + 0.05 \cdot (Y(1) - Y(0)))}.$$

Finally, the realized treatment is $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$, and the realized outcome is $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$. The LATE, LQTE($\tau$), $\tau = \{0.10, 0.25, 0.50, 0.75, 0.90\}$ are approximately equal to 39.25, 42.93, 36.93, 34.40, 36.93, and 42.93, respectively. This design is consistent with a generalized Roy model, under which individuals with higher treatment effects are more likely to be treated if they are eligible for treatment. We also emphasize that, given the one-sided non-compliance, LATE and LQTE are equal to the ATT and QTT in this scenario.

As before, we consider two scenarios. On the first one, we observed data on $\left(Y_i, D_i, Z_i, \mathbf{X}_i'\right)$, and, therefore, all IPW estimators are correctly specified. In the second scenario, we observed data on $\left(Y_i, D_i, Z_i, \mathbf{W}_i'\right)$, and all considered IPW estimators for LATE and LQTE are misspecified.

Table S1.3 displays the simulation results for both scenarios. When the instrument PS model is correctly specified, all estimators perform well in terms of bias and coverage probability, except the estimators based on the LIPS estimator (4.5) with the indicator weighting function — the bias of the local treatment effect estimators based on LIPS with indicator function is non-negligible when $n = 500$, and such biases distort

6 Simulation results with $n = 200$ and $n = 1000$ are available in the Supplementary Appendix.

the confidence intervals. In additional simulations, we note that the bias associated with estimators based on the LIPS with the indicator weighting function converges to zero when sample size grows, though the rate of convergence is rather slow. As such, we recommend that, in practice, one should favor the other PS estimators with respect to the LIPS with the indicator weighting function.

Like in the unconfoundedness setup, we note that IPW estimators based on ML tend to have higher mean square error, longer confidence intervals, and lower ARE than the IPW estimators based on the just-identified CBPS estimator; the performance of the over-identified CBPS is worse than the just-identified CBPS when one focus on LATE and LQTE with $\tau \leqslant 0.5$. The results in Table S1.3 also show that, when the instrument propensity score is correctly specified, the LIPS estimators with the exponential or projection weighting function tend to outperform the other methods, particularly when estimating the LATE, LQTE(0.75), and LQTE(0.90). These gains can be very pronounced. For example, the ARE of the LATE estimator based on the IPS with projection weight function with respect to the one based on the over-identified CBPS is 1.91. This implies that the LATE estimator based on the over-identified CBPS would require sample size almost twice as large to perform equivalently to the LATE estimator based on LIPS with projection weight. Using the exponential weight function also leads to large gains in precision without sacrificing coverage probability.

When the instrument PS model is misspecified, our Monte Carlo results suggest that using the LIPS can also be attractive. In this setup, we note that estimators based on ML tend to have higher biases, RMSE, and misleading confidence intervals. Local treatment effect estimators based on the (instrumented) CBPS improve upon those based on ML, with the just-identified CBPS estimator performing better than the over-identified CBPS (at least in terms of RMSE). However, under-coverage tends to remain severe, unless one focus on LQTE with $\tau \leqslant 0.25$. Our simulation results also suggest that our proposed LIPS estimators can lead to local treatment effect estimators with better statistical properties than those based on the (instrumented) CBPS. The RMSE improvements are very pronounced when one focus on LATE and LQTE with $\tau \geqslant 0.50$. Furthermore, our LIPS estimators tend to have size closer to the nominal level than the other PS estimators considered.

Next, we want to better understand the ability of these different estimators to minimize covariate distributional imbalances among compliers. Towards this end, we consider the following six distributional imbalance measures:

$$
\begin{aligned}
KS^{lte}_{bal}(\boldsymbol{\beta}) &= \sup_{i:1,\ldots,n} \|DistImb^{lte}(\mathbf{X}^*_i,\boldsymbol{\beta})\|, \qquad RCvM^{lte}_{bal}(\boldsymbol{\beta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} DistImb^{lte}(\mathbf{X}^*_i,\boldsymbol{\beta})^2}, \\
KS^{lte}_{bal1}(\boldsymbol{\beta}) &= \sup_{i:1,\ldots,n} \|DistImb^{lte}_1(\mathbf{X}^*_i,\boldsymbol{\beta})\|, \qquad RCvM^{lte}_{bal1}(\boldsymbol{\beta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} DistImb^{lte}_1(\mathbf{X}^*_i,\boldsymbol{\beta})^2} \qquad (5.2) \\
KS^{lte}_{bal0}(\boldsymbol{\beta}) &= \sup_{i:1,\ldots,n} \|DistImb^{lte}_0(\mathbf{X}^*_i,\boldsymbol{\beta})\|, \qquad RCvM^{lte}_{bal0}(\boldsymbol{\beta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} DistImb^{lte}_0(\mathbf{X}^*_i,\boldsymbol{\beta})^2},
\end{aligned}
$$

where

$$
DistImb^{lte}(\mathbf{x},\boldsymbol{\beta}) = \mathbb{E}_n\left[\left(\varpi^{lte}_{n,1}\left(D,Z,\mathbf{X}^*;\boldsymbol{\beta}\right) - \varpi^{lte}_{n,0}\left(D,\mathbf{X}^*;\boldsymbol{\beta}\right)\right) 1\left\{\mathbf{X}^* \leqslant \mathbf{x}\right\}\right],
$$

$$DistImb_1^{lte}(\mathbf{x},\boldsymbol{\beta}) = \mathbb{E}_n\left[\left(\varpi_{n,1}^{lte}\left(D,Z,\mathbf{X}^*;\boldsymbol{\beta}\right) - \varpi_n^{lte}\left(D,Z,\mathbf{X}^*;\boldsymbol{\beta}\right)\right)1\left\{\mathbf{X}^* \leqslant \mathbf{x}\right\}\right],$$
$$DistImb_0^{lte}(\mathbf{x},\boldsymbol{\beta}) = \mathbb{E}_n\left[\left(\varpi_{n,0}^{lte}\left(D,Z,\mathbf{X}^*;\boldsymbol{\beta}\right) - \varpi_n^{lte}\left(D,Z,\mathbf{X}^*;\boldsymbol{\beta}\right)\right)1\left\{\mathbf{X}^* \leqslant \mathbf{x}\right\}\right]),$$

and $\varpi_{n,d}^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)$, $d \in \{0,1\}$, and $\varpi_n^{lte}\left(D,Z,\mathbf{X};\boldsymbol{\beta}\right)$ are as defined in the system (4.6), and $\mathbf{X}^* = \mathbf{X}$ when the model is correctly specified, and $\mathbf{X}^* = \mathbf{W}$ when the model is misspecified. Notice that, according to Theorem 3.1 of Abadie (2003) and (4.3), all six distributional covariate imbalance should be close to zero when the instrument PS model is correctly specified.

Table S1.4 presents these six distributional balances measures when one estimates the instrument PS parameter β using different estimations procedures. When the instrument PS model is correctly specified, we can see that the LIPS estimator improve upon the CBPS and ML estimators. Among the LIPS estimators, the one with indicator function is the one who shows smallest distributional imbalances, which, again, should be no surprise as its objective function is designed to achieve that. The performance of LIPS with exponential or projection weighting function are similar, though the one with projection function tends to perform better. When the instrument PS is misspecified, we see that CBPS-based estimators tend to achieve better covariate balancing than the ML-based estimator, and that the IPS estimator with indicator and exponential weighting function further improve upon them. Just like in the unconfoundedness setup, these distributional imbalances tend to be larger when one use the IPS estimator with projection weighting function, even though the results in Table S1.3 favors this estimator. Like we mentioned before, this is not in conflict with our theory: these distributional imbalances are not guaranteed to be close to zero under model misspecification.

Overall, our stylized Monte Carlo simulations illustrate that our proposed LIPS estimators can lead to treatment effect estimators with improved finite sample properties. Our simulation results also point out that treatment effect estimators based on the IPS and LIPS estimators with either exponential or projection weighting functions tend to perform better than when one uses the indicator weighting function. As such, we recommend that, in practice, one should favor these weighting functions with respect to the indicator weighting function, especially when the dimension of the covariates included in the PS model is moderate or high[7].

5.2 Empirically calibrated simulation

To evaluate the performance of our proposed methods in a empirically driven setting, we conduct simulations calibrated to our empirical application—the effect of 401(k) retirement plans on asset accumulation—in Section 6 (Benjamin, 2003; Abadie, 2003; Chernozhukov and Hansen, 2004; Wüthrich, 2019). The outcome of interest Y_i is the net financial assets for household i. We consider two setups. In the first one, we want to assess the effect of 401(k) *eligibility* on accumulated assets (intention-to-treat/unconfoundedness setup). In the second one, we assess the effect of 401(k) *participation* on accumulated assets (local treatment setup). On top of the outcome Y, treatment eligibility Z, and treatment participation D, we also observe 10 additional covariates: income, log-income, age, family size, years of education, dummies for homeownership, marital

7 In unreported additional simulations, we also have found that the numerical performance of IPS_{ind} and $LIPS_{ind}$ is sometimes sensitive to initial values used in the optimization procedure when the number of included covariates is moderate. We argue that this is an additional reason to favor the other weighting functions with respect to the indicator one.

status, two-earner status, defined benefit pension status, and individual retirement account participation status. See Section 6 for more background on the application.

5.2.1 Intention to Treat

We start by analyzing the effect of 401(k) *eligibility* on accumulated assets. Here, we assume that, conditional on the vector of observed covariates, eligibility to 401(k) is as-good-as-random.

We calibrate our simulations as follows. We set the treated and untreated potential outcomes as

$$Y(1) = \frac{g_{out}(\mathbf{X})'\beta_1^{itt}}{s_{out}} + \varepsilon(1), \qquad Y(0) = \frac{g_{out}(\mathbf{X})'\beta_0^{itt}}{s_{out}} + \varepsilon(0), \tag{5.3}$$

where $g_{out}(\mathbf{X})$ includes an intercept and all 60 two-way interactions of the available covariates (income, log-income, age, family size, years of education, dummies for homeownership, marital status, two-earner status, defined benefit pension status, and individual retirement account participation status), β_d^{itt}, $d \in \{0,1\}$, is equal to the estimated OLS coefficients by regressing net financial assets on the two-way interactions using all the observed data with $Z = d$, and $\varepsilon(d)$ are independent $N(0,1)$ random variables. Covariates are randomly drawn from the empirical distribution of covariates in the original data. We set $s_{out} = 1,000$ so our outcome is measured in thousands of dollars. The ATE (or ITT) and the QTE(τ), $\tau = \{0.10, 0.25, 0.50, 0.75, 0.90\}$ are approximately equal to 8.55, 1.19, 4.03, 7.55, 12.87, and 14.57, respectively.

The true PS is given by

$$p(\mathbf{X}) = \frac{\exp\left(g_{ps}(\mathbf{X})'\beta_{ps}\right)}{1 + \exp\left(g_{ps}(\mathbf{X})'\beta_{ps}\right)}, \tag{5.4}$$

where $g_{ps}(\mathbf{X})$ is a vector including an intercept, all covariates in a linear fashion and the square of income, log-income, age, family size and years of education, the 16×1 vector β_{ps} is equal to the estimated maximum-likelihood coefficient of the logistic regression of Z on these covariates using all available data, and the treatment eligibility Z is generated as $Z = 1\{p(\mathbf{X}) > U\}$, where U follows a uniform $(0,1)$ distribution. In this section, we take $D = Z$ as we are interested in the intention-to-treat effects.

Like in the stylized simulation setup, we consider two different scenarios to assess the sensitivity of the proposed estimators not only under correctly specified PS models but also under misspecified PS models that are "nearly correct". In the first setup, we assume that researchers include the correct set of 16 covariates into the PS model. In the second setup, we consider the case that researchers do not include the squared covariates, and, therefore, the resulting PS model is misspecified. We set sample size n equal to $1,000$.

Table S2.1 displays the simulation results for both scenarios. When the PS model is correctly specified, we note that the over-identified CBPS estimators tend to have a non-negligible bias, while the just-identified CBPS estimators tend to under-cover the true causal parameters of interest—this under-coverage is particularly severe when focusing on the ATE and QTE(0.75). The IPW estimators based on maximum likelihood tend to perform well in terms of bias, RMSE, and coverage probability, except when $\tau = 0.75$. In this latter case, the estimator based on ML tends to under-cover the true QTE(0.75). Among the IPS-based estimators, the one that uses indicator function seems to be the least precise, both in terms of RMSE and length of 95% confidence intervals. This is in line with our simulation results using the stylized setup. As so, we once more reiterate that practitioners should favor the other IPS estimators in detriment of the one that used the indicator

function. Both IPS estimators with exponential and projection weighting functions tend to improve upon the over-identified CBPS in terms of MSE and ARE when one focuses on QTE's. The IPW estimator using IPS with exponential weights leads to tighter confidence intervals than the other IPS estimators, too. The gains in asymptotic relative efficiency are only pronounced when focusing on QTE's. In fact, when focusing on the ATE, in this correctly specified DGP it is hard to improve upon the IPW estimator based on the MLE.

When the propensity score is misspecified, several interesting patterns arise. First, we notice that the RMSE for the ATE estimator based on ML is substantially higher than the RMSE of the other considered estimators. In this particular case, the IPS estimator with projection weighting function has the smallest RMSE, followed by the over-identified CBPS. In terms of inference, the performance of the IPS estimators with projection function is substantially better than the other considered estimators. When one focuses on QTE's, we notice that all proposed IPS estimators tend to perform better than the over-identified CBPS estimator, except in the case with $\tau = 0.90$. We also notice that no estimator seems to control the the probability of type I errors across all considered quantiles. Given that the PS models are misspecified, this is not a big surprise. Finally, our simulation results reveal that the reduction in MSE by adopting IPS with exponential weights tends to be large, especially for $\tau = 0.25, 0.50, 0.75$.

Table S2.2 presents different distributional balance measures for different propensity score estimators. When the PS model is correctly specified, we can see that the IPS estimator always improve upon the CBPS and ML estimators. Among the IPS estimators, the one with indicator function is the one who performs "best", though this should be no surprise. Excluding the IPS with indicator weight function, the IPS with exponential weighting function seems to dominate the others. When the PS is misspecified, we see that CBPS-based estimators do not always improve upon the ML-based estimator. In fact, the over-identified CBPS estimator tend to perform worst than the other PS estimators in these covariate distributional imbalances measures. All IPS estimators seems to lead to improved distributional balance, though this is not always guaranteed by our theory.

5.2.2 Local Treatment Effects

In this section, we analyze the effect of 401(k) *participation* on accumulated assets. Here, we assume that, conditionally on the vector of observed covariates, eligibility to 401(k) is as-good-as-random, and use it as an instrument for 401(k) participation (treatment).

We calibrate our simulation as follows. We set the treated and untreated potential outcomes exactly as in (5.3), and the instrument propensity score $q(\mathbf{X})$ exactly as in (5.4). The instrument, eligibility to 401(k), Z is generated as $Z = 1\{q(\mathbf{X}) > U_1\}$, where U_1 follows a uniform $(0,1)$ distribution. The potential treatments $D(1)$ and $D(0)$—401(k) participation if eligible or not eligible–are generated as $D(1) = 1\{p^*(Y(1) - Y(0)) > U_2\}$ and $D(0) = 0$, where U_2 follows a uniform $(0,1)$ distribution, and

$$p^*(Y(1) - Y(0)) = \frac{\exp\left(\beta_0^{lte} + \beta_1^{lte} \cdot (Y(1) - Y(0))\right)}{1 + \exp\left(\beta_0^{lte} + \beta_1^{lte} \cdot (Y(1) - Y(0))\right)},$$

where β_0^{lte} and β_1^{lte} were calibrated by first simulating $Y(1) - Y(0)$ using all the data, and then using maximum likelihood to estimate the logit regression of treatment participation on an intercept and this difference, using

only data of eligible households.[8] Finally, the realized 401(k) participation is $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$, and the realized outcome (net financial household) is $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$. The LATE and LQTE($\tau$), $\tau = \{0.10, 0.25, 0.50, 0.75, 0.90\}$ are approximately equal to 9.04, 1.37, 4.18, 7.76, 13.35, and 15.23, respectively. Like in the stylized setup, LATE and LQTE are equal to the ATT and QTT as we have one-sided non-compliance..

Like in the ITT setup, we consider two different scenarios. In the first setup, we assume that researchers include the correct set of 16 covariates into the instrument PS model. In the second setup, we consider the case that researchers do not include the squared covariates, and, therefore, the resulting instrument PS model is misspecified. We consider sample size n equal to $1,000$.

Table S2.3 displays the simulation results for both scenarios. When the instrument PS is correctly specified, we notice that most IPW estimators are approximately unbiased, with the exception of the LIPS estimator with indicator weighting function and the over-identified CBPS estimators; these bias are particularly pronounced for the LATE and and for LQTE with $\tau = 0.75, 0.90$. In terms of MSE, coverage probabilities, and length of confidence intervals, we notice that the PS estimator based on ML tends to perform best when one focuses on the LATE. When one focus on LQTE, though, our simulation results reveal that the LIPS with either exponential or projection weighting function tend to lead to important gains in precision without sacrificing correct coverage probability. For example, when one focuses on $LQTE(0.50)$, we see that our proposed LIPS estimator with exponential weighting function has MSE that is approximately 20% smaller than the over-identified CBPS estimator, has coverage probability that is close to the nominal level, and shorter confidence intervals, too. In fact, the over-identified CBPS estimator would need a sample size 1.65 larger to perform equivalently to the LIPS with exponential weighting function. We also notice that the just-identified CBPS and the ML PS estimators tend to under-cover the true treatment effects—these under-coverage problems are specially severe when $\tau = 0.10$ or $\tau = 0.75$. Our proposed LIPS estimators based on exponential or projection weighting functions do not suffer from these practically important problems.

When the PS is misspecified, we first notice that, when one is interested in the LATE, the ML instrument PS estimator has a substantially higher MSE than any other estimators considered. We notice that the IPW estimator based on the over-identified CBPS estimator substantially improves the MSE of the LATE. Interestingly, we also see that our proposed LIPS with projection weighting function further improves upon the over-identified CBPS not only in terms of MSE but also in terms of coverage probability, length of confidence intervals, and asymptotic relative efficiency. In sum, when one is focused on the LATE, our simulation results show that the gains of adopting the LIPS with projection weight are potentially big in the presence of "modest" instrument PS misspecifications. When one focuses their attention on the LQTE, we first notice that the LIPS estimators tend to substantially improve upon the CBPS estimators in terms of bias and MSE, except when $\tau = 0.90$. The improviments in MSE can be large. in MSE LIPS estimators, the ones with exponential weighting function tend to perform best in this misspecified DGP. For instance, the MSE of the $LQTE(0.50)$ estimator based on the LIPS with exponential weighting function is 64% smaller than the one associated with the over-identified CBPS estimator.

Next, we analyze different covariate distributional balance measures for different propensity score esti-

8 β_0^{lte} and β_1^{lte} are approximately equal to 0.812 and 0.005, respectively.

mators. The results are displayed in Table S2.4. When the instrument PS model is correctly specified, we can see that the LIPS estimator with exponential weight tends to improve upon the CBPS and ML estimators. The same is true for the LIPS estimator with indicator function, though this is expected. We see that the LIPS estimator with projection weight function tends to improve upon the ML estimator but performs slightly worse than the CBPS estimators. When the PS is misspecified, though, we see that the ML estimator tends to perform best (after, of course, the LIPS with indicator function), even though the simulation results in Table S2.3 reveal that our LIPS estimators tend to dominate ML estimators in this particular case. Like we mentioned in our stylized simulation setups, this is not in conflict with our theory: these covariate distributional imbalances are not guaranteed to be close to zero under model misspecification.

Overall, the results of our calibrated Monte Carlo simulations agree with the ones in our stylized setup. That is, our results illustrate that our proposed IPS/LIPS estimators can lead to treatment effect estimators with improved finite sample properties. Just like in the stylized setup, our simulation results also point out that treatment effect estimators based on the IPS and LIPS estimators with either exponential or projection weighting functions tend to perform better than when one uses the indicator weighting function, and we recommend practitioners to favor them.

6 Effect of 401(k) retirement plans on asset accumulation

As discussed in Benjamin (2003), Abadie (2003), Chernozhukov and Hansen (2004), Belloni, Chernozhukov, Fernández-Val and Hansen (2017) and many others, tax-deferred retirement plans have been popular in the US since the 1980s. A main goal of these programs is to increase individual saving for retirement. Amongst the most popular tax-deferred programs is the 401(k) plan. Interestingly, 401(k) plans are provided by employers, and, therefore, only workers in firms that offer such programs are eligible. On the other hand, we emphasize that *eligible* employees choose whether to participate (i.e., make a contribution) or not, making the evaluation of the effectiveness of 401(k) plans on accumulated assets more challenging as a result of endogeneity concerns — individuals who participate in 401(k) programs have stronger preferences for savings and would have saved more even in the absence of these programs.

To bypass the endogeneity challenge, Benjamin (2003) uses data from the 1991 Survey of Income and Program Participation (SIPP) and compares households that are eligible with those who are non-eligible for 401(k) plans to assess the effect of *eligibility* on accumulated assets. He argues that since 401(k) eligibility is determined by the employers, household preference for savings plays a negligible role in determining eligibility once one controls for observed household characteristics. Using PS matching, Benjamin (2003) finds evidence that 401(k) eligibility has a positive effect on asset accumulation.

Abadie (2003), Chernozhukov and Hansen (2004) and Wüthrich (2019), on the other hand, study the effect of 401(k) *participation* on asset accumulation, using 401(k) eligibility as an instrument for the actual participation status. Similarly to Benjamin (2003), they argue that 401(k) eligibility is exogenous after controlling for a vector of observed household characteristics. Abadie (2003), using a semiparametric IPW estimator for the LATE, finds that the effect of 401(k) participation on net financial assets is significant and positive. Chernozhukov and Hansen (2004) and Wüthrich (2019), using an IV quantile regression model,

also find positive and significant effects of 401(k) participation on net financial assets.

In what follows, we apply the methodology discussed in Sections 3.2 and 4 to study the effects of eligibility and participation in 401(k) programs on saving behavior. As suggested by Benjamin (2003), Abadie (2003), and Chernozhukov and Hansen (2004), eligibility is assumed to be exogenous after controlling for covariates. Also note that, because only eligible individuals can enroll in 401(k) plans, the monotonicity condition in Assumption 5(iii) holds trivially, and the LATE and LQTE estimators presented in Section 4 approximate the average and quantile treatment effect for the treated (i.e., for 401(k) participants).

We use the same dataset as Benjamin (2003), Chernozhukov and Hansen (2004) and Wüthrich (2019). The data consists of a sample of 9,910 households from the 1991 SIPP[9]. The outcomes of interest are net financial assets, and total wealth. For the (instrument) propensity score estimation, we adopt a logistic specification, and use all two-way interactions between income, log-income, age, family size, years of education, dummies for homeownership, marital status, two-earner status, defined benefit pension status, and individual retirement account participation status. To assess the reliability of this parametric PS model, we apply the specification test of Sant'Anna and Song (2019) with 1,000 bootstrap draws, and fail to reject the null of the propensity score model being correctly specified at the 10% level.

Panel A (Panel B) of Table 1 shows the point estimates and standard errors (in parentheses) for the effect of 401(k) eligibility (participation among compliers) on net financial assets and total wealth. We present IPW estimators for the ATE, QTE(τ), LATE, LQTE(τ), $\tau = 0.10, 0.25, 0.50, 0.75, 0.90$, using the same PS estimation methods as in the simulation exercise in Section 5, except the IPS and LIPS estimators based on the indicator weighting function, as they tend to be numerically unstable when the dimension of covariates is moderate. We also report two measures of covariate distributional imbalance for each PS estimator.

The results in Panel A suggest that 401(k) eligibility has a positive and significant average impact on both net financial assets and total wealth and that the effect is more pronounced at the higher quantiles. When one compares the treatment effect measures across different PS estimation methods, we see that the results tend to be similar for net financial assets; for total wealth, we note that estimators based on the over-identified CBPS estimator suggest much larger effects of 401(k) eligibility at higher quantiles than those based on our proposed IPS estimators; see Figure 1 for a more detailed comparison between the QTE estimates based on the IPS with projection weighting function, over-identified CBPS (the default in the `CBPS R` package), and those based on ML. Interestingly, although we adopt a parsimonious specification, our IPW results for net financial assets are similar to those in Belloni et al. (2017), who consider lasso-based high-dimensional specifications based on a doubly-robust procedure[10]. For instance, their estimated ATE of 401(k) plan eligibility on net financial assets is 7,848 with a standard error of 1,317, while our estimated ATE based on IPS with projection weighting function is 7,784 with a standard error of 1,605; QTE estimates based on over-identified CBPS and IPS are also close to those in Belloni et al. (2017). In terms of covariate

9 The original data have 9,915 households, but we follow Benjamin (2003) and delete the five observations with zero or negative income. Descriptive statistics are available in Table 1 in Benjamin (2003) and in Tables 1 and 2 in Chernozhukov and Hansen (2004).

10 Their procedure requires one to model not only the propensity score but also the outcome equation for treated and untreated units. As so, they use "more information" than IPW procedures. Belloni et al. (2017) focus on net financial assets as a measure of wealth.

distributional imbalances, note that our proposed IPS estimators improve upon both CBPS estimators; the IPS estimator with exponential weight function also improves upon the MLE.

Panel A: Effects of 401(k) plan eligibility on wealth

	Outcome: Net Financial Assets					Outcome: Total Wealth				
	MLE	$CBPS_{just}$	$CBPS_{over}$	IPS_{exp}	IPS_{proj}	MLE	$CBPS_{just}$	$CBPS_{over}$	IPS_{exp}	IPS_{proj}
ATE	8,138	8,190	8,820	8,218	7,784	6,049	5,997	7,906	6,589	5,395
	(1,135)	(1,150)	(1,362)	(1,376)	(1,605)	(1,823)	(1,811)	(2,486)	(2,201)	(2,792)
QTE(0.10)	1,113	1,200	1,160	1,150	1,050	399	400	400	400	375
	(264)	(254)	(259)	(258)	(283)	(578)	(558)	(601)	(557)	(577)
QTE(0.25)	996	996	1,000	1,000	996	3,024	2,917	3,425	2,993	2,950
	(229)	(228)	(237)	(225)	(231)	(611)	(591)	(789)	(593)	(617)
QTE(0.50)	4,447	4,200	4,559	4,350	4,300	7,402	7,419	9,027	7,615	7,419
	(278)	(259)	(331)	(276)	(309)	(1,162)	(1,111)	(1,580)	(1,143)	(1,157)
QTE(0.75)	13,065	12,995	13,980	13,339	12,859	9,131	8,871	13,050	10,419	8,665
	(931)	(922)	(1,166)	(964)	(1,025)	(2,833)	(2,786)	(3,742)	(2,972)	(3,158)
QTE(0.90)	21,249	21,053	23,441	21,890	20,899	15,857	15,979	18,500	17,547	15,504
	(2,223)	(2,247)	(2,934)	(2,399)	(2,782)	(5,806)	(5,829)	(6,662)	(6,167)	(7,818)
KS_{bal}	2.26	2.76	2.63	1.98	2.31	2.26	2.76	2.63	1.98	2.31
$RCvM_{bal}$	0.56	0.56	0.65	0.51	0.57	0.56	0.56	0.65	0.51	0.57

Panel B: Effects of 401(k) plan participation on wealth

	Outcome: Net Financial Assets					Outcome: Total Wealth				
	MLE	$CBPS_{just}$	$CBPS_{over}$	$LIPS_{exp}$	$LIPS_{proj}$	MLE	$CBPS_{just}$	$CBPS_{over}$	$LIPS_{exp}$	$LIPS_{proj}$
LATE	11,674	11,700	12,767	12,108	11,176	8,706	8,568	11,590	9,923	7,740
	(1,621)	(1,640)	(1,929)	(1,929)	(2,250)	(2,609)	(2,587)	(3,532)	(3,094)	(3,872)
LQTE(0.10)	2,554	2,624	2,642	2,624	2,449	1,220	1,109	1,471	1,120	1,045
	(506)	(486)	(486)	(495)	(528)	(815)	(807)	(827)	(786)	(833)
LQTE(0.25)	1,618	1,536	1,753	1,589	1,529	5,226	4,853	6,204	5,200	5,003
	(266)	(265)	(284)	(260)	(266)	(879)	(839)	(1,122)	(828)	(857)
LQTE(0.50)	7,285	7,041	7,849	7,341	7,197	10,187	9,925	12,701	10,730	10,026
	(533)	(515)	(653)	(518)	(524)	(1,279)	(1,232)	(1,696)	(1,249)	(1,316)
LQTE(0.75)	19,939	19,589	21,772	20,325	19,410	14,061	13,200	19,909	16,353	13,041
	(1,034)	(1,015)	(1,319)	(1,065)	(1,136)	(1,054)	(1,037)	(1,342)	(1,087)	(1,159)
LQTE(0.90)	28,501	28,450	31,798	31,200	27,919	19,200	18,908	24,402	22,400	17,172
	(728)	(715)	(869)	(749)	(839)	(779)	(779)	(898)	(808)	(998)
KS_{bal}^{lte}	3.23	3.94	3.71	2.90	3.41	3.23	3.94	3.71	2.90	3.41
$RCvM_{bal}^{lte}$	0.80	0.81	0.92	0.73	0.84	0.80	0.81	0.92	0.73	0.84

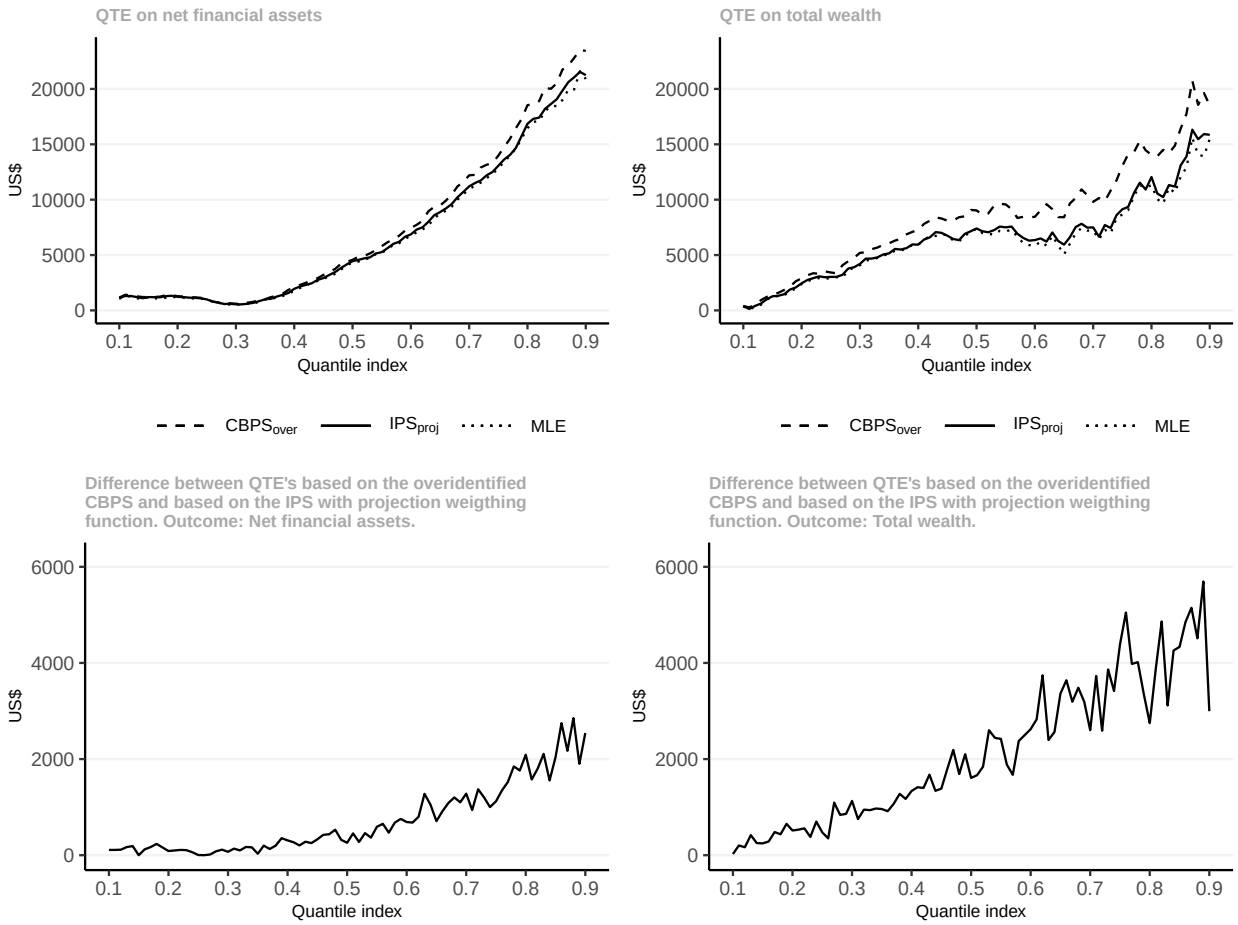
Note: Same data used by Benjamin (2003) and Chernozhukov and Hansen (2004). The propensity score model is based on a logistic link function. Standard errors in parentheses. The estimators in Panel A are the same as those we describe in Table S1.1, whereas those in Panel B are the same as those described in Table S1.3. The distributional imbalance measures are as defined in (5.1) and (5.2).

Table 1: Effects of 401(k) plan on different measures of wealth

The results in Panel B paint a similar picture as those in Panel A: 401(k) participation tends to have a positive and significant average impact on both measures of wealth, and the effect is more pronounced at the right tail of the wealth measures. As we illustrated in Figure 2, there are quantitative differences between the LQTE estimates based on different PS estimation methods, with those based on the over-identified instrument CBPS suggesting much larger effects than the other estimation methods, though the shape of the LQTE function is similar across specifications. Like in Panel A, our results for net financial assets are also similar to those in Belloni et al. (2017). Their reported LATE estimate based on high-dimensional sparse models is 11,267 (standard error of 1,890), while our IPS LATE estimate with projection weighting function is 11,176 (standard error of 2,250). In terms of LQTE, Belloni et al. (2017) estimates tend to be similar to IPS estimates with exponential weighting function, except for quantiles above $\tau = 0.85$, where their estimates

tend to be higher than any of the considered IPW procedures. In terms of covariate distributional imbalances, Panel B also reveals that our proposed LIPS estimators tend to improve upon both CBPS estimators; the LIPS estimator with exponential weight function perform best according to this metric.

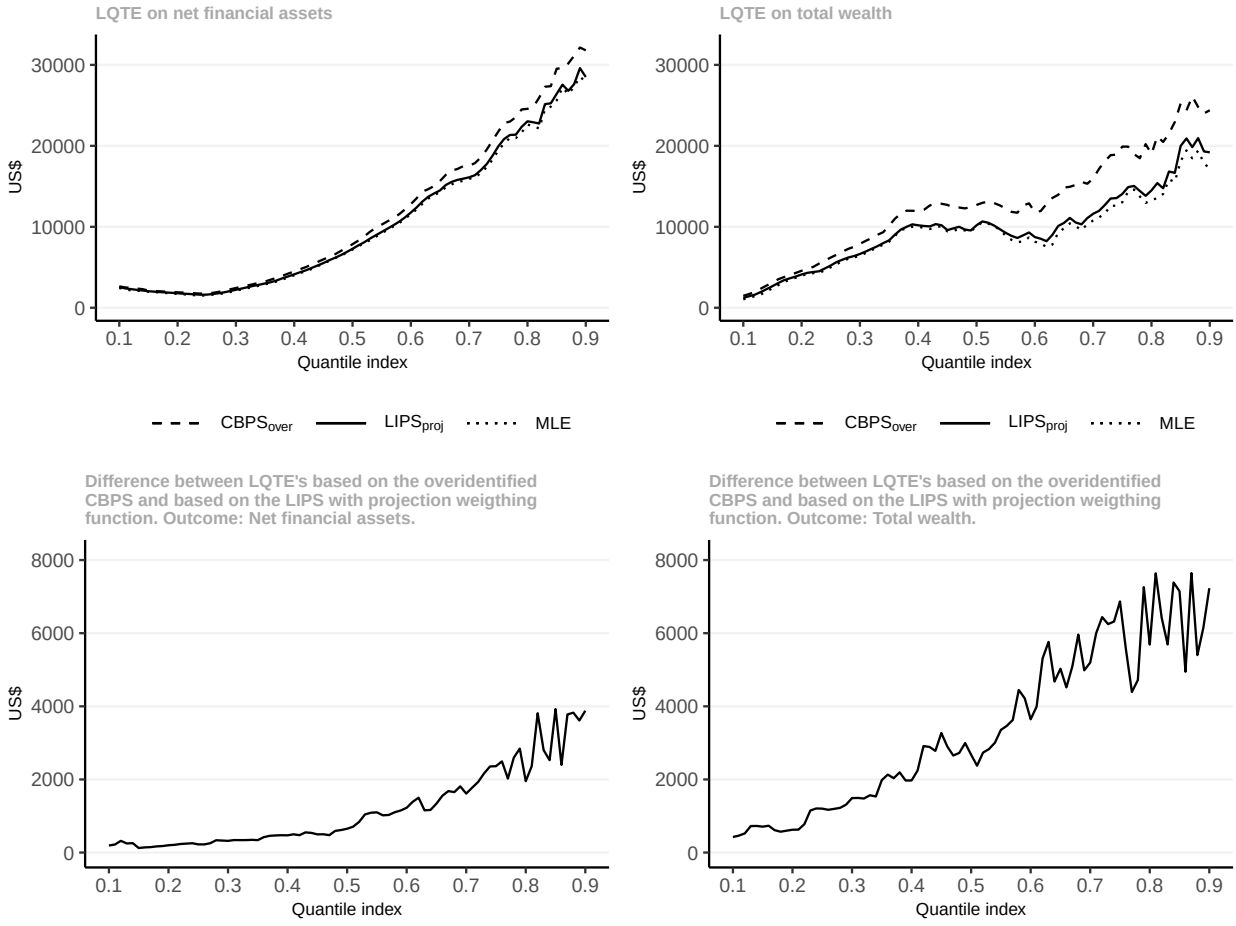
Figure 1: Estimated quantile treatment effects of 401(k) eligibility on different wealth measures.



The results in Panel B paint a similar picture as those in Panel A: 401(k) participation tends to have a positive and significant average impact on both measures of wealth, and the effect is more pronounced at the right tail of the wealth measures. As we illustrated in Figure 2, there are quantitative differences between the LQTE estimates based on different PS estimation methods, with those based on the over-identified instrument CBPS suggesting much larger effects than the other estimation methods, though the shape of the LQTE function is similar across specifications. Like in Panel A, our results for net financial assets are also similar to those in Belloni et al. (2017). Their reported LATE estimate based on high-dimensional sparse models is 11,267 (standard error of 1,890), while our IPS LATE estimate with projection weighting function is 11,176 (standard error of 2,250). In terms of LQTE, Belloni et al. (2017) estimates tend to be similar to IPS estimates with exponential weighting function, except for quantiles above $\tau = 0.85$, where their estimates tend to be higher than any of the considered IPW procedures. In terms of covariate distributional imbalances, Panel B also reveals that our proposed LIPS estimators tend to improve upon both CBPS estimators; the

LIPS estimator with exponential weight function perform best according to this metric.

Figure 2: Estimated local quantile treatment effects of 401(k) participation on different wealth measures.



7 Conclusion

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. Importantly, we have shown that our framework can accommodate the empirically relevant situation under which treatment allocation is endogenous. We derived the large sample properties of average, distributional and quantile treatment effect estimator based on the proposed integrated propensity scores, and illustrated its attractive properties via a Monte Carlo study and an empirical application.

Although this paper devoted most of its attention to forming IPW-type treatment effect estimators, we note that sometimes researchers are willing to consider an outcome regression model, on top of the propensity

score model. In such cases, we stress that one can easily combine our IPS estimation procedure with such outcome regression model to form doubly-robust, locally efficient treatment effect estimators, see, e.g., Słoczyński and Wooldridge (2018) and references therein. Perhaps even better, one can use the integrated moment approach adopted in this paper to estimate not only the propensity score, but also the outcome regression model. We leave the detailed discussion of such procedure for future research.

References

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

Abadie, A., Angrist, J. D., and Imbens, G. W. (2002), "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70(1), 91–117.

Ai, C., and Chen, X. (2003), "Efficient estimation of models with conditional moment restrictions containin unknown functions," *Econometrica*, 71(6), 1795–1843.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 91(434), 444–455.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017), "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85(1), 233–298.

Benjamin, D. J. (2003), "Does 401(k) eligibility increase saving? Evidence from propensity score subclassification," *Journal of Public Economics*, 87(5-6), 1259–1290.

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

Bierens, H. J., and Wang, L. (2012), "Integrated Conditional Moment Tests for Parametric Conditional Distributions," *Econometric Theory*, 28(2), 328–362.

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.

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.

Chen, X., Hong, H., and Tarozzi, A. (2008), "Semiparametric efficiency in GMM models with auxiliary data," *The Annals of Statistics*, 36(2), 808–843.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010), "Quantile and Probability Curves Without Crossing," *Econometrica*, 78(3), 1093–1125.

Chernozhukov, V., Fernández-Val, I., Melly, B., and Wüthrich, K. (2019), "Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes," *Journal of the American Statistical Association*, pp. 1–24.

Chernozhukov, V., and Hansen, C. (2004), "The impact of 401(k) participation on the wealth distribution: an instrumental quantile regression analysis," *The Review of Economics and Statistics*, 86(3), 735–751.

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

Díaz, J., Rau, T., and Rivera, J. (2015), "A Matching Estimator Based on a Bilevel Optimization Problem," *Review of Economics and Statistics*, 97(4), 803–812.

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.

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

Donald, S. G., Imbens, G. W., and Newey, W. K. (2003), "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics*, 117(1), 55–93.

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.

Fan, J., Imai, K., Liu, H., Ning, Y., and Yang, X. (2016), "Improving Covariate Balancing Propensity Score : A Doubly Robust and Efficient Approach," *Mimeo*, pp. 1–47.

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. (2007), "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139(1), 35–75.

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., 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.

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(4), 605–654.

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

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. 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.

Millimet, D. L., and Tchernis, R. (2009), "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies," *Journal of Business & Economic Statistics*, 27(3), 397–415.

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.

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., and Song, X. (2019), "Specification tests for the propensity score," *Journal of Econometrics*, 210(2), 379–404.

Shaikh, A. M., Simonsen, M., Vytlacil, E. J., and Yildiz, N. (2009), "A specification test for the propensity score using its distribution conditional on participation," *Journal of Econometrics*, 151(1), 33–46.

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.

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

Wüthrich, K. (2019), "A closed-form estimator for quantile treatment effects with endogeneity," *Journal of Econometrics*, 210(2), 219–235.

Zhao, Q. (2019), "Covariate balancing propensity score by tailored loss functions," *The Annals of Statistics*, 47(2), 965–993.

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.