Nonparametric Test for Propensity Score

Zixuan Wu

January 19, 2020

1 Introduction

Recently, program evaluation has been a popular topic for econometricians. The goal of program evaluation is to understand the causal effect of a treatment, such as participation in a training program or implementation of a policy, on individuals or on the whole society. To measure this causal effect, we need potential outcomes under two different states: treatment and control. However, these two potential outcomes can't be observed together, i.e. for each individual, we can only observe one outcome at one time. Econometricians have developed different tools to deal with this missing data dilemma. Rosenbaum and Rubin(1983)[8] introduce the notion of propensity score, which is defined as the probability of assignment to treatment conditional on covariates. If the treatment assignment is independent of the potential outcomes conditional on a vector of covariates, then one can obtain unbiased and consistent estimators of different treatment effect measures by adjusting for the propensity score. There are a variety of estimating procedures exploiting this important insight, such as propensity score matching method (PSM), see e.g. Rosenbaum and Rubin(1985)[9], Heckman et al.(1997)[5] and Abadie and Imbens(2016)[1]; inverse probability weighting method (IPW), see e.g.

Rosenbaum(1987)[7], Hirano et al.(2003) and Donald and Hsu(2014); regression methods, see e.g. Hahn(1998)[4] and Firpo(2007); and many others.

Despite their popularity, a main concern of these methods is that the propensity score is usually unknown, and therefore has to be estimated. Given that the covariate space is often of high dimensionality, researchers are more willing to adopt a parametric model for the propensity score since nonparametric methods will suffer from "the curse of dimensionality", which raises the problem of model specification. Propensity score misspecification may lead to misleading treatment effect estimates, see e.g. Frölich(2004), Millimet and Tchernis(2009), Huber et al. (2013) and Busso et al.(2014).

There is a vast amount of literature on testing the specification of a parametric regression model, and they can be divided into two classes: local smoothing method and global smoothing method. The former one makes use of nonparametric estimation of covariates, see e.g. Eubank and Hart(1992), Wooldridge(1992), H*ä*rdle and Mammen(1993), Hong and White(1995), Zheng(1996)[13], Horowitz and Spokoiny(2001), and Guerre and Lavergne (2005). The latter one avoids smoothing estimation by reducing the conditional mean independence to an infinite number of unconditional orthogonality restrictions, see e.g. Bierens and Ploberger(1997)[2], Stute(1997)[12], Stinchcombe and White(1998), Li, Hsiao and Zinn(2003), and Escanciano(2006)[3]. This paper discusses these two classes of test and compare their performances by Monte Carlo simulations. Our discussion are mainly based on Shaikh et al.(2009)[11] and Sant'Anna et al.(2019)[10], which are representative for local smoothing method and global smoothing method respectively.

2 Background

Let D be a binary random variable indicating whether to receive the treatment, i.e. $D = 1$ represents receiving the treatment and $D = 0$ otherwise. Define $Y(1)$ and $Y(0)$ as the potential outcomes under treatment and control, thus the observed outcome is $Y = Y(1) D + Y(0)(1 - D)$. Denote X as a $k \times 1$ vector of available pre-treatment covariates, and the support of X is $\mathcal{X} \subseteq \mathbb{R}^k$. The propensity score is defined as $p(x) = \mathbb{P}(D = 1 \mid X = x)$. We have a random sample $\{\left(Y_i, D_i, X_i'\right)'\}_{i=1}^n$of size $n \geq 1$ from $\left(Y, D, X'\right)'$.

The main goal in program evaluation is to assess the effect of a treatment D on the outcome Y. The most popular parameters of interest include the average treatment effect, $ATE = \mathbb{E}[Y(1) - Y(0)]$, and the average treatment effect on the treated, $ATT = \mathbb{E}[Y(1) - Y(0) \mid D = 1]$. Notice that the potential outcomes $Y(1)$ and $Y(0)$ cannot be jointly observed for the same individual, one of the most popular identification strategies to resolve such difficulty is to assume that selection into treatment is solely based on observable characteristics, the so-called unconfoundedness setup, see e.g. Rosenbaum and Rubin (1983). Formally, we need following assumptions:

Assumption 2.1 $(Y(1), Y(0)) \perp\!\!\!\perp D \mid X$.

Assumption 2.2 $\forall x \in \mathcal{X}, 0 < p(x) < 1$.

As shown by Rosenbaum (1987), under Assumptions 2.1–2.2, ATE and ATT are identified by

$$ATE = \mathbb{E}\left[\left(\frac{D}{p(X)} - \frac{1 - D}{1 - p(X)}\right)Y\right] \quad and \quad ATT = \frac{\mathbb{E}\left[\left(D - \frac{p(X)(1-D)}{1-p(X)}\right)Y\right]}{\mathbb{E}[D]} \tag{1}$$

Thus we can estimate the propensity score first, and then use the sample analogy to estimate ATE and ATT:

$$\widehat{ATE}_n = \frac{1}{n}\sum_{i=1}^{n}\left[(\frac{D_i}{\hat{p}(X_i)} - \frac{1-D_i}{1-\hat{p}(X_i)})Y_i\right] \text{ and } \widehat{ATT}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[(D_i - \frac{\hat{p}(X_i)(1-D_i)}{1-\hat{p}(X_i)})Y_i\right]}{\frac{1}{n}\sum_{i=1}^{n} D_i} \tag{2}$$

To ensure that such estimators are well-defined, we need to assess the overlap between the distribution of the propensity score among treatment and control groups. Following Heckman et al. (1998), we can compare kernel density estimates of the propensity score among treated and control samples to determine the common support region. Although kernel density estimators are popular, they involve choosing tuning parameters such as bandwidths and often suffer from boundary bias. Of course such inconveniences can be easily avoided if one focuses on CDFs instead of densities. And we discuss specification tests for propensity score models based on kernel density estimators and CDFs respectively.

3 Shaikh et al.(2009) test

Assume that the propensity score $p(X)$ has a density with respect to Lebesgue measure and denote it as $f(p)$. Denote $f_1(p)$ and $f_0(p)$ as the density of the propensity score for treatment group and control group respectively.

Lemma 3.1 *Let* $\alpha = \frac{\mathbb{P}(D=0)}{\mathbb{P}(D=1)}$ *and and assume* $0 < \mathbb{P}(D=0) < 1$. *Then for all* $0 < p(X) = p < 1$, *we have that*

$$f_1(p) = \alpha\frac{p}{1-p}f_0(p) \tag{3}$$

Lemma 3.1 implies that if the parametric model for $p(X)$ is correctly specified, then

we can expect (3) to hold approximately when estimated.

Lemma 3.2 *Let* $\alpha = \frac{\mathbb{P}(D=0)}{\mathbb{P}(D=1)}$ *and assume* $0 < \mathbb{P}(D=0) < 1$. *Let* Q *be a random variable with density w.r.t. Lebesgue measure and* $Q \in [0,1]$, *denote* $g_1(q)$ *and* $g_0(q)$ *as the density of* Q *conditional on* $D=1$ *and* $D=0$ *respectively. Then we have*

$$g_1(q) = \alpha \frac{q}{1-q} g_0(q) \tag{4}$$

for all $q \in (0,1)$ *if and only if*

$$\mathbb{E}[D - Q \mid Q = q] = 0 \tag{5}$$

Lemma 3.2 implies that we can test whether the parametric model for $p(X)$ is correctly specified by means of testing whether the sample analogy of (5) holds. The obvious advantage of (5) is that it only involves low dimensional conditional expectations, which avoids "the curse of dimensionality" when exploiting nonparametric estimation methods.

Now let us demonstrate our test in a formal way. We would like to test whether there exists $\theta_0 \in \Theta$ such that $\mathbb{E}[D \mid Q(X, \theta_0)] = Q(X, \theta_0)$ with probability 1. And our null and alternative hypotheses are given by

$$H_0 : \exists \theta_0 \in \Theta \quad s.t. \quad \mathbb{P}\{\mathbb{E}[D \mid Q(X, \theta_0)] = Q(X, \theta_0)\} = 1$$

$$H_1 : \mathbb{P}\{\mathbb{E}[D \mid Q(X, \theta)] = Q(X, \theta)\} < 1 \quad \forall \theta \in \Theta$$

By analogy with the test statistic in Zheng (1996), we consider testing based on

$$\hat{V}_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{h} K\left(\frac{Q\left(X_i, \hat{\theta}_n\right) - Q\left(X_j, \hat{\theta}_n\right)}{h} \right) \hat{\varepsilon}_i \hat{\varepsilon}_j \tag{6}$$

where $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimator which minimizes $\mathbb{E}\left[(Y - Q(X, \theta))^2\right]$, $K(\cdot)$ is a kernel function and h is the bandwidth parameter, and $\hat{\varepsilon}_i := D_i - Q\left(X_i, \hat{\theta}_n\right)$.

We need following assumptions to study the asymptotic properties of our test.

Assumption 3.1 *(D_i, X_i), $i = 1, ..., n$ is an i.i.d. sequence of random variables on $\{0,1\} \times \mathbb{R}^k$.*

Assumption 3.2 *Θ is a compact subset of $\mathbb{R}^k$.*

Assumption 3.3 *$Q : supp(X_i) \times \Theta \to [0,1]$ satisfies:*

(a) $Q(X_i, \theta)$ has a continuous density $f(x, \theta)$ w.r.t. Lebesque measure for all θ in a neighborhood of θ_0, where $\theta_0 = \arg\min_{\theta \in \Theta} \mathbb{E}\left[(D_i - Q(X_i, \theta))^2\right]$.

(b) $Q(x, \theta)$ is Lipschitz continuous w.r.t.θ in the sense that for all $\theta \in \Theta$ and $\theta' \in \Theta$, $\left|Q(x, \theta) - Q\left(x, \theta'\right)\right| \leq G(x) \left\|\theta - \theta'\right\|$, where $\mathbb{E}\left[G^{4+\delta}(X_i)\right] < \infty$ for some $\delta > 0$.

Assumption 3.4 *$K : \mathbb{R} \to \mathbb{R}$ is bounded, Lipschitz continuous, symmetric and satisfies:*

(a) $\int K(u)\, du = 1$.

(b) $\int |K(u)|\, du < \infty$.

(c) $\int |uK(u)|\, du < \infty$.

(d) $\int |u^2 K(u)|\, du < \infty$.

Assumption 3.5 *$\hat{\theta}_n$ satisfies $\left\|\hat{\theta}_n - \theta_0\right\| = O_p\left(\frac{1}{\sqrt{n}}\right)$.*

Assumption 3.6 *The bandwidth sequence satisfies $0 < h = h_n \to 0$ and $nh^4 \to \infty$.*

Then we demonstrate the behaviour of our test statistic under the null hypothesis.

Theorem 3.1 *Suppose Assumptions 4.1-4.6 hold. If* $\mathbb{E}[D_i \mid X_i] = Q(X_i, \theta_0)$ *with probability 1, then* $\hat{V}_n$ *in (6) satisfies* $n\sqrt{h}\hat{V}_n \to N(0, \Sigma)$*, where* $\Sigma = 2\iint q_1^2 (1-q_1)^2 K^2(u) f^2(q_1) du dq_1$*. Moreover,* Σ *may be consistently estimated by*

$$\hat{\Sigma}_n = \sum_{1 \le i,j \le n : i \ne j} \frac{2\hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2}{n(n-1)h} K^2 \left(\frac{Q\left(X_i, \hat{\theta}_n\right) - Q\left(X_j, \hat{\theta}_n\right)}{h} \right) \tag{7}$$

4 Sant'Anna et al.(2019) test

Assume that the propensity score $p(X)$ has a density with respect to a dominating measure, and that the density is bounded away from zero and infinity uniformly over its support.

Lemma 4.1 *Let* $\alpha = \frac{\mathbb{P}(D=0)}{\mathbb{P}(D=1)}$*/ and assume that* $0 < \mathbb{P}(D=1) < 1$*. If* $0 < p(X) < 1$ *a.s., then* $\forall u \in [0,1]$*,*

$$\mathbb{E}[I\{p(X) \le u\} \mid D=1] = \alpha \mathbb{E}\left[\frac{p(X)}{1-p(X)} I\{p(X) \le u\} \mid D=0 \right] \tag{8}$$

Furthermore, (8) holds if and only if $\forall u \in [0,1]$

$$\mathbb{E}[(D - p(X)) I\{p(X) \le u\}] = 0 \tag{9}$$

Lemma 4.1 implies that if the propensity score is correctly specified, the sample analogue of (8) should hold. Thus, we can test whether the parametric model for $p(X)$

is correctly specified based on

$$H_0 : \forall u \in \Pi, \exists \theta_0 \in \Theta, s.t. \mathbb{E}\left[\left(D - q\left(X, \theta_0\right)\right) I\{q\left(X, \theta_0\right) \leq u\}\right] = 0 \tag{10}$$

where $\Theta \subset \mathbb{R}^k$, $\Pi = [0,1]$, and $q\left(X, \theta\right) : \mathcal{X} \times \Theta \mapsto [0,1]$ is a family of parametric functions dominated by the finite dimensional parameter θ. Common specifications for $q\left(X, \theta\right)$ include the Probit, $\phi\left(X'\theta\right)$ and the Logit, $\Lambda\left(X'\theta\right)$, where $\phi\left(\cdot\right)$ and $\Lambda\left(\cdot\right)$ are the normal and logistic link functions.

4.1 Projection-based specification tests

Recall that $\varepsilon_i\left(\theta\right) = D_i - q\left(X_i, \theta\right)$. For all $u \in \Pi$, define

$$\mathcal{P}_n I\{q\left(X, \theta\right) \leq u\} = I\{q(X, \theta) \leq u\} - g'\left(X, \theta\right) \Delta_n^{-1}\left(\theta\right) G_n\left(u, \theta\right) \tag{11}$$

where $g\left(x, \theta\right) = \partial q\left(x, \theta\right) / \partial \theta$ is the score function of $q\left(x, \theta\right)$,

$G_n\left(u, \theta\right) = \frac{1}{n} \sum_{i=1}^n g\left(X_i, \theta\right) I\{q\left(X_i, \theta\right) \leq u\}$ and $\Delta_n\left(\theta\right) = \frac{1}{n} \sum_{i=1}^n g\left(X_i, \theta\right) g'\left(X_i, \theta\right)$

And our test statistics are based on continuous functionals of the projection-based empirical process $\hat{R}_n^p\left(u\right)$,

$$\hat{R}_n^p\left(u\right) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i\left(\hat{\theta}_n\right) \mathcal{P}_n I\{q\left(X_i, \hat{\theta}_n\right) \leq u\} \tag{12}$$

where $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimator for θ_0 under H_0. Two popular examples of such functionals are Cramé-von-Mises and Kolmogorov-Smirnov functionals,

$$CvM_n = \int_\Pi \left|\hat{R}_n^p\left(u\right)\right|^2 F_n\left(du\right) = \frac{1}{n} \sum_{i=1}^n \left[\hat{R}_n^p\left(q\left(X_i, \hat{\theta}_n\right)\right)\right]^2 \tag{13}$$

$$KS_n = \sup_{u \in \Pi} \left| \hat{R}_n^p(u) \right| \tag{14}$$

where $F_n(u) = \frac{1}{n}\sum_{i=1}^n I\{q\left(X_i, \hat{\theta}_n\right) \leq u\}$ is the empirical distribution function of $q\left(X_i, \hat{\theta}_n\right)$.

A main advantage of considering the projection is that we can easily prove that the projected process $\hat{R}_n(u)$ is insensitive to the choice of the estimator $\hat{\theta}_n$. On the contrary, if we consider the umprojected process $\hat{R}_n(u)$,

$$\hat{R}_n(u) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i\left(\hat{\theta}_n\right) I\{q\left(X_i, \hat{\theta}_n\right) \leq u\} \tag{15}$$

We can prove that

$$\hat{R}_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) I\{q(X_i, \theta_0) \leq u\} - \sqrt{n}\left(\hat{\theta}_n - \theta_0\right)' \mathbb{E}\left[g(X, \theta_0) I\{q(X, \theta_0) \leq u\}\right] + o_p(1) \tag{16}$$

And (16) implies that the effect of replacing θ_0 by $\hat{\theta}_n$ is non-negligible, thus the asymptotic null distributions of tests based on (15) are sensitive to the choice of the estimator $\hat{\theta}_n$.

4.2 Asymptotic theory

The asymptotic null distributions of tests based on (12) are the limiting distributions of continuous functionals of $\hat{R}_n^p$ under H_0. For a generic set $\mathcal{G}$, let $\ell^\infty(\mathcal{G})$ be the Banach space of all uniformly bounded real functions on $\mathcal{G}$, equipped with the uniform metric $\|f\|_{\mathcal{G}} \equiv sup_{z \in \mathcal{G}} |f(z)|$.

We study the weak convergence of $\hat{R}_n^p(u)$ and its related processed as elements of $\ell^\infty(\Pi)$, where $\Pi \equiv [0, 1]$. We assume the following regularity conditions. Let Θ_0 be an arbitrarily small neighborhood around θ_0 such that $\Theta_0 \subset \Theta$. For any $d_1 \times d_2$ matrix $A = (a_{ij})$, let $\|A\|$ denote its Euclidean norm, i.e. $\|A\| = [tr(AA')]^{1/2}$.

Assumption 4.1 *(i) The parameter space Θ is a compact subset of $\mathcal{R}^k$; (ii) The true*

parameter θ_0 *belongs to the interior of* Θ*; (iii)* $\left\|\hat{\theta}_n - \theta_0\right\| = O_p\left(n^{-1/2}\right)$.

Assumption 4.2 *The parametric propensity score function* $q(x,\theta)$ *is twice continuously differentiable in* Θ_0 *for each* $x \in \mathcal{X}$*, with its first derivative* $g(x,\theta) = \partial g(x,\theta)/\partial\theta = (g_1(x,\theta), ..., g_k(x,\theta))'$ *satisfying* $\mathbb{E}[sup_{\theta\in\Theta_0} \|g(X,\theta)\|] < \infty$ *and its second derivative satisfying* $\mathbb{E}[sup_{\theta\in\Theta_0} \|\partial g(X,\theta)/\partial\theta\|] < \infty$*. Furthermore, the matrix* $\Delta(\theta) \equiv \mathbb{E}[g(X,\theta)\, g'(X,\theta)]$ *is nonsingular in* Θ_0*.*

Assumption 4.3 *The function* $F_\theta(u) = \mathbb{P}(q(X,\theta) \leq u)$ *satisfies* $sup_{u\in\Pi} |F_{\theta_1}(u) - F_{\theta_2}(u)| \leq C\|\theta_1 - \theta_2\|$*, where* C *is a bounded positive number, not depending on* θ_1 *and* θ_2*.*

Theorem 4.1 *Let Assumptions 4.1-4.3 hold. Then, under* H_0*, we have that*

$$\sup_{u\in\Pi} \left|\hat{R}_n^p(u) - R_{n0}^p(u)\right| = o_p(1) \text{ and } \hat{R}_n^p(u) \Rightarrow R_\infty^p$$

where R_∞^p *denotes a Gaussian process with mean zero and covariance structure given by*

$$K^p(u_1,u_2) = \mathbb{E}[q(X,\theta_0)(1 - q(X,\theta_0))\mathcal{P}I\{q(X,\theta_0) \leq u_1\}\mathcal{P}I\{q(X,\theta_0) \leq u_2\}]$$

Theorem 1 and the continuous mapping theorem yield the asymptotic null distributions of continuous functionals of $\hat{R}_n^p(u)$, including the test statistics CvM_n and KS_n given in (13) and (14) respectively.

Theorem 4.2 *Under the assumptions of Theorem 1 and* H_0*, for any continuous functional* $\Gamma(\cdot)$ *from* $\ell^\infty(\Pi)$ *to* $\mathbb{R}$*, we have* $\Gamma\left(\hat{R}_n^p\right) \xrightarrow{d} \Gamma(R_\infty^p)$*.*
Futhermore, $CvM_n \xrightarrow{d} CvM_\infty := \int_\Pi |R_\infty^p(u)|^2 dF_{\theta_0}(u)$*, where* $F_{\theta_0}(u) = \mathbb{P}(q(X,\theta_0) \leq u)$ *denotes the cumulative distribution function of* $q(X,\theta_0)$ *and* $KS_n \xrightarrow{d} KS_\infty := \sup_{u\in\Pi} |R_\infty^p(u)|$*.*

Notice that the covariance structure $K^p(\cdot,\cdot)$ is rather complicated, in order to cumpute critical values, we exploit a multiplier bootstrap method.

More precisely, in order to estimate the critical values, we propose to approximate the asymptotic behavior of $\hat{R}_n^p(u)$ by that of

$$\hat{R}_n^{p*}(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \left(\hat{\theta}_n\right) \mathcal{P}_n I\{q\left(X_i, \hat{\theta}_n\right) \leq u\} V_i \tag{17}$$

where $\{V_i\}_{i=1}^n$ is a sequence of *i.i.d.* random variables with zero mean, unit variance and bounded support, independent of the original sample $\{\left(D_i, X_i'\right)'\}_{i=1}^n$. We use *i.i.d.* Bernoulli variates $\{V_i\}$ proposed by Mammen(1993)[6]:

$$V_i = \begin{cases} (1-\sqrt{5})/2, & p = (\sqrt{5}+1)/2\sqrt{5} \\ (\sqrt{5}+1)/2, & p = (\sqrt{5}-1)/2\sqrt{5} \end{cases}$$

The bootstrapped version of our test statistics $\Gamma\left(\hat{R}_n^p\right)$ is simply given by $\Gamma\left(\hat{R}_n^{p*}\right)$. For instance,

$$CvM_n^* = \tfrac{1}{n}\textstyle\sum_{i=1}^n \left[\hat{R}_n^{p*}\left(q\left(X_i, \hat{\theta}_n\right)\right)\right]^2$$
$$KS_n^* = \sup_{u \in \Pi} \left|\hat{R}_n^{p*}(u)\right|$$

The asymptotic critical values are then estimated by

$$c_{n,\alpha}^{\Gamma\,*} \equiv inf\{c_\alpha \in [0,\infty) : \lim_{n\to\infty} \mathbb{P}_n^*\{\Gamma\left(\hat{R}_n^{p*}\right) > c_\alpha\} = \alpha\}$$

where $\mathbb{P}_n^*$ means bootstrap probability. In practice, $c_{n,\alpha}^{\Gamma\,*}$ is approximated as accurately as desired by $\left(\Gamma\left(\hat{R}_n^{p*}\right)\right)_{B(1-\alpha)}$, the $B(1-\alpha)$-th order statistic from B replicates $\{\left(\Gamma\left(\hat{R}_n^{p*}\right)\right)_l\}_{l=1}^B$ of $\Gamma\left(\hat{R}_n^{p*}\right)$.

Theorem 4.3 *Assume Assumptions 4.1-4.3. Then,* $\hat{R}_n^{p*} \underset{*}{\Rightarrow} R_\infty^p$ *a.s.*

where R_∞^p *is the Gaussian process defined in Theorem 4.1, and* $\underset{*}{\Rightarrow}$ *denotes the weak convergence under the bootstrap law. Additionally, for any continuous functional* $\Gamma(\cdot)$ *from* $\ell^\infty(\Pi)$ *to* $\mathbb{R}$*, we have* $\Gamma\left(\hat{R}_n^{p*}\right) \underset{*}{\xrightarrow{d}} \Gamma\left(R_\infty^p\right)$ *a.s. under the bootstrap law.*

5 Monte Carlo simulation

We conduct a series of Monte Carlo experiments to compare previous tests:

(i) Shaikh et al.'s(2009) test, $T_n(h_n) = \sqrt{\frac{n-1}{n}} \frac{n h_n^{1/2} \hat{V}_n(h_n)}{\sqrt{\hat{\Sigma}_n(h_n)}}$, where $\hat{V}_n(h_n)$ and $\hat{\Sigma}_n(h_n)$ are given by (6) and (7) respectively;

(ii) Sant'Anna et al.(2019) test, CvM_n and KS_n given in (13) and (14) respectively.

We consider sample sizes n equal to 100, 200, 400, 600, 800 and 1000. For each design, we consider 1000 Monte Carlo experiments. For Shaikh et al.'s(2009) test, we use the standard normal kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$, and the bandwidth sequence h_n is chosen to be equal to $cn^{-1/8}$ for c = 0.01, 0.05, 0.10 and 0.15.

5.1 Simulation 1

We first consider the following data generating processes:

DGP1. $D^* = \frac{(X_1+X_2)}{3} - \varepsilon$;

DGP2. $D^* = -1 + \frac{(X_1+X_2+X_1X_2)}{3} - \varepsilon$;

DGP3. $D^* = -0.2 + \frac{\left(X_1^2-X_2^2\right)}{2} - \varepsilon$;

DGP4. $D^* = \frac{(0.1+X_1/3)}{\exp((X_1+X_2)/3)} - \varepsilon$;

DGP5. $D^* = \frac{(-0.8+(X_1+X_2+X_1X_2)/3)}{\exp(0.2+(X_1+X_2)/3)} - \varepsilon$.

For each DGP, $D = I\{D^* > 0\}$, $\varepsilon \perp\!\!\!\perp (X_1, X_2)$, where $X_1 = Z_1$, $X_2 = (Z_1 + Z_2)/\sqrt{2}$, and Z_1, Z_2, ε are independent standard normal random variables.

Let $X = (1, X_1, X_2)'$, the null hypothesis is

$$H_0 : \exists \theta_0 = (\beta_0, \beta_1, \beta_2,)' \in \Theta, \text{s.t. } \mathbb{E}\left[D \mid \phi\left(X'\theta_0\right)\right] = \phi\left(X'\theta_0\right) \text{ a.s.}$$

where $\phi(\cdot)$ is the cumulative function of the standard normal distribution, and θ_0 is estimated by maximum likelihood method. Obviously, DGP1 falls under H_0, whereas DGP2–DGP5 fall under H_1.

dgp	n	CvM	KS	Shaikh(0.01)	Shaikh(0.05)	Shaikh(0.10)	Shaikh(0.15)
1	100	0.057	0.052	0.033	0.023	0.008	0.004
	200	0.054	0.058	0.054	0.02	0.011	0.003
	400	0.061	0.053	0.045	0.021	0.015	0.005
	600	0.061	0.057	0.052	0.022	0.015	0.009
	800	0.06	0.056	0.04	0.027	0.017	0.011
	1000	0.062	0.054	0.052	0.025	0.012	0.007
2	100	0.269	0.246	0.062	0.073	0.069	0.055
	200	0.602	0.526	0.091	0.19	0.211	0.191
	400	0.931	0.862	0.215	0.459	0.519	0.526
	600	0.987	0.961	0.318	0.668	0.751	0.763
	800	1	0.994	0.509	0.844	0.901	0.906
	1000	1	0.999	0.662	0.939	0.963	0.968
3	100	0.35	0.279	0.061	0.059	0.018	0
	200	0.544	0.464	0.165	0.132	0.046	0.011
	400	0.714	0.658	0.417	0.319	0.107	0.018
	600	0.778	0.735	0.579	0.378	0.147	0.038
	800	0.805	0.778	0.65	0.437	0.157	0.034
	1000	0.83	0.797	0.685	0.472	0.154	0.04
4	100	0.166	0.14	0.043	0.03	0.02	0.01
	200	0.353	0.282	0.061	0.065	0.05	0.031
	400	0.639	0.523	0.088	0.192	0.21	0.182
	600	0.853	0.77	0.138	0.348	0.425	0.399
	800	0.935	0.886	0.193	0.508	0.612	0.604
	1000	0.981	0.964	0.302	0.691	0.8	0.808
5	100	0.136	0.111	0.058	0.03	0.013	0.01
	200	0.182	0.16	0.046	0.039	0.036	0.028
	400	0.357	0.315	0.058	0.08	0.084	0.073
	600	0.503	0.456	0.08	0.136	0.155	0.145
	800	0.609	0.578	0.085	0.185	0.234	0.223
	1000	0.691	0.635	0.109	0.249	0.31	0.319

Figure 1: Simulation 1

The simulation result is given by Figure 1. Obviously, we use DGP1 to check the size of the test while DGP2–DGP5 are used to check the power of the test. For Sant'Anna et al.(2019) test, the simulated size is close to the nominal level($\alpha = 0.05$), and the power increases as sample size increases. And For Shaikh et al.'s(2009) test, the performance is not satisfactory. The simulated size does well only under $c = 0.01$. And when we focus on the power of the test, the optimal choice of tuning parameter c varies as sample size and DGP change, which suggests that Shaikh et al.'s(2009) test is very sensitive to the tuning parameters.

dgp	n	CvM	KS	Shaikh(0.01)	Shaikh(0.05)	Shaikh(0.10)	Shaikh(0.15)
6	100	0.079	0.077	0.048	0.019	0.006	0.001
	200	0.07	0.066	0.043	0.018	0.011	0.007
	400	0.058	0.064	0.051	0.025	0.012	0.008
	600	0.056	0.059	0.046	0.024	0.007	0.004
	800	0.044	0.048	0.039	0.018	0.007	0.005
	1000	0.05	0.058	0.05	0.018	0.005	0.003
7	100	0.079	0.092	0.059	0.018	0.005	0.001
	200	0.102	0.104	0.066	0.014	0.004	0.001
	400	0.132	0.133	0.042	0.017	0.007	0.004
	600	0.163	0.163	0.044	0.025	0.016	0.008
	800	0.247	0.231	0.044	0.046	0.039	0.019
	1000	0.305	0.27	0.054	0.065	0.05	0.031
8	100	0.087	0.1	0.067	0.023	0.006	0.003
	200	0.135	0.131	0.039	0.024	0.008	0.003
	400	0.255	0.234	0.049	0.042	0.034	0.018
	600	0.42	0.372	0.071	0.099	0.088	0.05
	800	0.588	0.54	0.115	0.185	0.172	0.128
	1000	0.661	0.599	0.127	0.245	0.246	0.186
9	100	0.088	0.095	0.045	0.019	0.005	0.001
	200	0.127	0.121	0.042	0.016	0.012	0.005
	400	0.237	0.198	0.039	0.03	0.029	0.021
	600	0.375	0.31	0.059	0.046	0.044	0.034
	800	0.467	0.392	0.051	0.065	0.07	0.066
	1000	0.587	0.486	0.053	0.089	0.108	0.108
10	100	0.08	0.093	0.048	0.016	0.001	0
	200	0.077	0.082	0.037	0.021	0.009	0.003
	400	0.128	0.117	0.042	0.019	0.014	0.01
	600	0.221	0.166	0.061	0.034	0.03	0.014
	800	0.309	0.226	0.054	0.051	0.05	0.033
	1000	0.404	0.31	0.064	0.094	0.09	0.054

Figure 2: Simulation 2

5.2 Simulation 2

We push forward the dimensionality of the covariates to see how these tests perform with 10 continuous covariates. We consider the following five DGPs:

DGP6. $D^* = -\frac{\sum_{j=1}^{10} X_j}{6} - \varepsilon$;

DGP7. $D^* = -1 - \frac{\sum_{j=1}^{10} X_j}{10} + \frac{X_1 X_2}{2} - \varepsilon$;

DGP8. $D^* = -1 - \frac{\sum_{j=1}^{10} X_j}{10} + \frac{X_1 \sum_{k=2}^{5} X_k}{4} - \varepsilon$;

DGP9. $D^* = -1.5 - \frac{\sum_{j=1}^{10} X_j}{6} + \frac{\sum_{k=1}^{10} X_k^2}{10} - \varepsilon$;

DGP10. $D^* = \frac{-0.1+0.1\sum_{j=1}^{5} X_j}{\exp\left(-0.2\sum_{k=1}^{10} X_j\right)} - \varepsilon$.

The simulation result is given by Figure 2. Obviously, we use DGP6 to check the

size of the test while DGP7–DGP10 are used to check the power of the test. We first focus on the size of these tests. The Sant'Anna et al.(2019) test is oversized when the sample size is relatively small($n = 100$). As sample size increases, the simulated size is closed to its nominal level. And Shaikh et al.'s(2009) test still performances well under $c = 0.01$. As for the test power, it increases as sample size increases for all these tests. It seems that CvM test outperforms KS test, and the optimal choice of tuning parameters for Shaikh et al.'s(2009) test still varies a lot.

6 Conclusion

In this paper, we introduce two different methods for the test of propensity score, based on Shaikh et al.(2009) and Sant'Anna et al.(2019) respectively. The former one, exploiting kernel density estimation, requires the local smoothing of the data and the choice of tuning parameters such as bandwidth, whereas the latter one avoids such limitations. We conduct a series of Monte Carlo simulations to compare these two different tests, and the results suggest that Sant'Anna et al.(2019) test does outperform Shaikh et al.'s(2009) test, no matter for test size or test power. And Sant'Anna et al.(2019) test does not need to select tuning parameter, which is of great convenience.

References

[1] Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

[2] Herman J Bierens and Werner Ploberger. Asymptotic theory of integrated conditional moment tests. *Econometrica: Journal of the Econometric Society*, pages

1129–1151, 1997.

[3] J Carlos Escanciano. A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6):1030–1051, 2006.

[4] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

[5] James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654, 1997.

[6] Enno Mammen et al. Bootstrap and wild bootstrap for high dimensional linear models. *The annals of statistics*, 21(1):255–285, 1993.

[7] Paul R Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

[8] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[9] Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[10] Pedro HC Sant'Anna and Xiaojun Song. Specification tests for the propensity score. *Journal of Econometrics*, 210(2):379–404, 2019.

[11] Azeem M Shaikh, Marianne Simonsen, Edward J Vytlacil, and Nese Yildiz. A specification test for the propensity score using its distribution conditional on participation. *Journal of Econometrics*, 151(1):33–46, 2009.

[12] Winfried Stute. Nonparametric model checks for regression. *The Annals of Statistics*, pages 613–641, 1997.

[13] John Xu Zheng. A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, 75(2):263–289, 1996.