Testing Unconfoundedness Assumption Using Auxiliary Variables

Zongwu Cai\textsuperscript{a}, Ying Fang\textsuperscript{b,c}, Ming Lin\textsuperscript{b,c}, Shengfang Tang\textsuperscript{c}

\textsuperscript{a}Department of Economics, University of Kansas, Lawrence, KS 66045, USA.
\textsuperscript{b}Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Sciences, Xiamen University, Xiamen, Fujian 361005, China.
\textsuperscript{c}Department of Statistics, School of Economics, Xiamen University, Xiamen, Fujian 361005, China.

March 4, 2019

Abstract: In this paper, we propose an alternative test procedure for testing the conditional independence assumption which is an important identification condition commonly imposed in the literature of program analysis and policy evaluation. We transform the conditional independence test to a nonparametric conditional moment test using an auxiliary variable which is independent of the treatment assignment variable conditional on potential outcomes and observable covariates. The proposed test statistic is shown to have a limiting normal distribution under null hypotheses of conditional independence. Furthermore, the suggested method is shown to be valid under time series framework and thus the corresponding test statistic and its limiting distribution are also established. Monte Carlo simulations are conducted to examine the finite sample performances of the proposed test statistics. Finally, the proposed test method is applied to test the conditional independence in real examples: the 401(k) participation program and return to college education.

Keywords: Conditional independence; Moment test; Nonparametric estimation; Selection on observable; Treatment effect.

JEL classification: C12; C13; C14; C23

\*The authors acknowledge the financial supports from the National Science Fund of China (NSFC) for Distinguished Scholars (71625001) and the NSFC key projects with grant numbers 71631004 and 71131008.
1 Introduction

It has increased demands for policymakers to evaluate quantitative effects of social and economic programs and policies. A main challenge of identifying causal policy effects lies in the situation that an analyst cannot simultaneously observe a same unit’s outcome when it is exposed to the policy or not. The unobserved part is called the counterfactual outcome, and furthermore, the treatment effect, the terminology used in the literature referring to the causal effect of a policy intervention, is defined as the difference between the observed and its counterfactual outcomes. In the case of ideal experiments, all units are randomly assigned to the so called treated group and control group. Units in the treated group are all exposed to the policy intervention and the ones in the control group are not. The average treatment effect then can be identified by a simple comparison of the observed outcomes between the two groups. However, social scientists are not so lucky to have the capacity to manipulate the experiment design. The identification in a non-experimental setting is mainly dependent on the conditional independence assumption introduced by Rosenbaum and Rubin (1983), which is also called the conditional unconfoundedness assumption or selection on observables in the literature.

The conditional independence refers to the assumption that conditional on observable confounders, the potential outcomes are independent of treatment status. In many applications, a weaker version, the conditional mean independence, is sufficient to identify treatment effects. The conditional independence assumption plays a central role in identifying the average treatment effects. For example, with the conditional independence assumption, Heckman, Ichimura and Todd (1997, 1998) and Hahn (1998) proposed nonparametric regression estimators, Abadie and Imbens (2006, 2016) proposed matching estimators, Hirano, Imbens and Ridder (2003) and Firpo (2007) proposed the inverse probability weighting estimators, and Robins and Rotnitzky (1995) and Lee, Okui and Whang (2017) proposed the doubly robust estimators. For more references, see Imbens (2004) and Imbens and Wooldridge (2009)'s recent surveys in this field. When there exist unobservable confounders which affect both the potential outcomes and treatment status, the conditional independence does not hold anymore and therefore all the aforementioned estimators are generally inconsistent.

Although it is of increased practical interests in testing the conditional independence for quantitatively evaluating policy effects, there are only a few of procedures available in the literature to test the conditional mean independence. Using additional binary instrumental variables, Donald, Hsu and Lieli (2014) proposed a Durbin-Wu-Hausman type statistic to test the conditional mean independence by comparing the local average treatment effect on the treated based on the binary instrumental variable and the average treatment effect on
the treated without using instrumental variables. The test requires the binary instrumental variable to satisfy the so-called one-sided noncompliance condition which implies that there are no always-takers. Recently, Chen, Ji, Zhou and Zhu (2017) proposed another conditional mean independence test without relying on the availability of the binary instrumental variables. Instead, a Kolmogorov-Smirnov type statistic is constructed to compare two estimators, the one which is only valid with conditional mean independence assumption and the other without it. However, the validity of the latter estimator requires that the error terms in both the outcome equation and the selection equation be symmetrically distributed, and the selection equation needs at least one covariate not included in the outcome equation.

This paper proposes an alternative method to test the conditional independence. Instead of the availability of a binary instrumental variable in Donald et al. (2014) and the requirement of symmetrically distributed error terms in Chen et al. (2017), our method relies on the existence of an auxiliary variable which is correlated to potential outcomes but is independent of the treatment status given on potential outcomes and observable covariates. In other words, this auxiliary variable is possible to have an effect on the treatment choice. However, the linkage from the auxiliary variable to the treatment status can be fully captured by potential outcomes and observable covariates. When such auxiliary variables are available, the conditional independence test can be simply implemented by a conditional moment test using nonparametric method. Moreover, compared to Donald et al. (2014) and Chen et al. (2017), our method can be applied to test not only the conditional mean independence but also the conditional independence, the stronger version.

The auxiliary variable assumption has been widely used in the literature of dealing with missing data problems as in Zhao and Shao (2015) and Breunig (2017), measurement error problems in Hu and Schennach (2008), and other scenarios. It is not difficult to find such auxiliary variables in many applications. For example, estimating return to college education is one of core issues in labor economics. However, labor economists concern whether attending college education is correlated to some unobservable confounders, such as individual ability, which are also affect future incomes. Chen et al. (2017) revisited this issue using a data set from China Health and Nutrition Survey with rich information about individual characteristics. Age can be considered as a candidate for being an auxiliary variable when we have controls on individual income, working experience and other individual characteristics. This can be further convinced from the data that the average age between the treated group (receiving college education) and the control group (not receiving college education) is indifferent. Another example is to investigate the effect of the participation into the 401(k) programs on financial assets holdings as studied by Abadie (2003), Chernozhukov and Hansen (2004), Ogburn, Rotnitzky and Robins (2015), and Belloni, Chernozhukov, Fernández-Val
and Hansen (2017). The main challenge comes from the unobservable saver heterogeneity coupled with nonrandom selection into participation states. Any estimation without dealing with this concern carefully will be biased upward, tending to overestimating the actual effect of the plan participation. Therefore, it is important to test if participation status is independent of financial assets holdings given some control variables. For this example, one can choose age or family size as auxiliary variable. The readers are referred to Section 5 for detailed analyses of these two examples.

Finally, the proposed methods are extended to the framework of macroeconomic policy evaluation. Recently, there has been increasing interest to identify the causal connection between macroeconomic policy and real economic variables using causal inference technical. See, for example, Angrist and Kuersteiner (2011), Angrist, Jordà and Kuersteiner (2018), Jordà and Taylor (2016), and the references therein. To identify the parameters of interest, such as the causal effects of monetary policy shocks, the average treatment effect of fiscal policy and among others, the selection on observables assumption assumed in Assumption 3.1 which is the conditional independence assumption for time series is assumed to be held in general. Similar to the case of cross-sectional data setting, when auxiliary variable satisfied the requirements mentioned above is available, it is showed that the test for the selection on observables assumption can also be implemented by a conditional moment test using nonparametric procedure. Furthermore, the proposed test statistic is showed to have approximately normal distribution in large samples.

The remainder of the paper is organized as follows. Section 2 introduces the model, discusses sufficient conditions for the testability of conditional independence and formulate our test statistics, and establish their asymptotic properties. In Section 3, we extend the framework of Section 2 to time series setting. Section 4 examines the finite sample performances of the proposed test using Monte Carlo simulations. In Section 5, the proposed method is applied to test the conditional independence in two real examples: the participation into 401(k) program and attending college education, and Section 6 concludes. The proofs of the main results are given in mathematical appendices.

2 Framework and Hypothesis Testing

2.1 Model Setup and Testability

The model is developed within the conventional framework of the Rubin causal model as in Rubin (1974) and Imbens and Rubin (2015). Under such framework, \( Y(1) \) and \( Y(0) \) denote the potential outcomes for a unit receiving or not receiving treatment, respectively. For each unit in the population, let \( D \) denote whether the treatment of interest is received, with \( D = 1 \)
if the unit receives the treatment, otherwise \( D = 0 \). In addition, each unit is also characterized by a vector of covariates denoted by \( X \in \mathbb{R}^d \). The fundamental problem in the treatment effect literature is that exactly one (never both) of the two potential outcomes \( Y(0) \) and \( Y(1) \) are observed for a particular individual. So using the notation above, for each individual, the observed data are only \( (Y, D, X) \) where \( Y = D \cdot Y(1) + (1 - D) \cdot Y(0) \). In the treatment effect literature, the average treatment effect (ATE) is defined by \( \delta = E(Y(1) - Y(0)) \), and the average treatment effect on the treated (ATT) is defined by \( \delta_{D = 1} = E(Y(1) - Y(0) | D = 1) \). Unfortunately, the quantities of interest mentioned above cannot be identified without additional assumptions. One of commonly used assumptions to help identification is the so-called conditional unconfoundedness assumption or conditional independence assumption, which asserts that the potential outcomes \( (Y(0), Y(1)) \) are independent of the treatment assignment variable \( D \) conditional on the vector of covariates \( X \); that is,

\[
(Y(0), Y(1)) \perp D \mid X,
\]

where \( \perp \) indicates statistical independence. However, this assumption may be violated in practice if there exist unobserved confounders which affect both potential outcomes and the treatment assignment variable. Thus, it is desirable to formally propose a procedure to test whether the conditional unconfoundedness assumption holds or not.

We propose a novel method to test the conditional independence assumption. The starting point of our procedure is the availability of a vector of auxiliary variables \( Z \in \mathbb{R}^r \) to satisfy the following assumption.

**Assumption 2.1.** (i) Assume that there exists a vector of continuously distributed variables \( Z \in \mathbb{R}^r \) which are correlated with both \( Y(0) \) and \( Y(1) \) and satisfy the following condition

\[
Z \perp D \mid (Y(0), Y(1), X).
\]

(ii) (Bounded completeness) For each bounded function \( \psi_m(\cdot) \),

\[
E[\psi_m(Y(0), Y(1), X) | X, Z] = 0 \text{ implies that } \psi_m(Y(0), Y(1), X) = 0 \text{ almost surely (a.e.).}
\]

**Remark 2.1.** Assumption 2.1(i) requires whether receiving treatment for a unit is primarily determined by the potential outcomes \( (Y(0), Y(1)) \) and covariates \( X \). The statement is thus that, given \( Y(0), Y(1) \) and \( X \), the treatment assignment variable \( D \) and the auxiliary variable \( Z \) are mutually independent and the information about \( D \) from \( Z \) can then be completely captured by \( (Y(0), Y(1), X) \). This assumption is appropriate when receiving treatment is driven by potential outcomes \( Y(0) \) and \( Y(1) \) and once given the information of \( Y(0), Y(1), \) and \( X, Z \) does not include any additional information on the assignment mechanism. This assumption has been widely adopted in the literature. For example, Breunig (2017) made
a similar assumption under the missing data framework and proposed a test statistic based on series estimators. Other examples include papers by Zhao and Shao (2015), Hu and Schennach (2008), Ramalho and Smith (2013), D’Haultfoeuille (2010), and among others.

Assumption 2.1(ii) is normally referred to as the bounded completeness in $Z$ of the conditional distribution of $(Y(0),Y(1))$ conditional on $(X,Z)$. This type of completeness condition is weaker than the completeness in $Z$ of the conditional distribution of $(Y(0),Y(1))$ conditional on $(X,Z)$ (which is defined as: for any measurable functions $\phi(\cdot)$ with $E|\phi(Y(0),Y(1),X)| < \infty$, $E(\phi(Y(0),Y(1),X)|X,Z) = 0$ implies $\phi(Y(0),Y(1),X) = 0$ almost surely). There are many families of distributions that are bounded complete and sufficient conditions for bounded completeness can be found in Mattner (1993) or D’Haultfoeuille (2011) and among others. Assumption 2.1(ii) has also been largely used in econometrics as identification assumptions, see, for example, Darolles, Fan, Florens and Renault (2011), Newey and Powell (2003), Blundell, Chen and Kristensen (2007), Hu and Schennach (2008), D’Haultfoeuille (2011), Hoderlein, Nesheim and Simoni (2017), Breunig (2017) and Breunig, Mammen and Simoni (2018).

For example, Assumption 2.1 holds true in the following model. Suppose that there exist maps $\lambda(\cdot)$ and $\pi(\cdot)$ from, respectively, $\mathbb{R}^{d+r}$ and $\mathbb{R}^2$ to $\mathbb{R}^2$, such that

$$(Y(0),Y(1)) = \pi(\lambda(X,Z) + \varepsilon), \quad \text{with} \quad Z \perp \varepsilon|X,$$

and

$$D = h(Y(0),Y(1),X,\zeta), \quad \text{with} \quad \zeta \perp (Z,\varepsilon) |X,$$

under a large support condition of $\lambda(X,Z)$ in the sense of Assumption A2 in D’Haultfoeuille (2011), regularity assumptions for $\varepsilon$, and if the conditional characteristic function $\varphi_{\varepsilon|X}(t|x)$ of $\varepsilon$ conditional on $X$ is infinitely often differentiable and does not vanish on the real line. For this model defined above, one can check the conditions of Theorem 2.1(i) in D’Haultfoeuille (2011) and it is easy to obtain that Assumption 2.1 is indeed satisfied; see D’Haultfoeuille (2011) for more details. It is a standard assumption in the deconvolution literature that the characteristic function is non-vanishing and there are many distributions satisfying this assumption. Examples include the normal, $\chi^2$, Student, Gamma, and double exponential distributions, while the uniform and the triangular distributions are the only common distributions to violate this restriction. In the example of return to college education, $\zeta$ is the unobservable individual ability. Age can be regarded as a valid auxiliary variable when it is not correlated with unobservable ability conditional on incomes, working experience and other covariates.

The following result states that the conditional unconfoundedness assumption is testable
under Assumption 2.1.

**Lemma 2.1.** (i) Under Assumption 2.1(i), the conditional unconfoundedness assumption implies that $E(D - E(D|X), Z) = 0$; that is, $E(D|X, Z) = E(D|X)$.

(ii) If both Assumptions 2.1(i) and (ii) are satisfied, the conditional unconfoundedness assumption is equivalent to $E(D|X, Z) = E(D|X)$.

**Proof.** It is easy to see that under Assumption 2.1(i), one can obtain $Z \perp D \mid (X, Y(0), Y(1))$. So by the similar arguments of Theorem 2.1 in Breunig (2017), one can show that this lemma is true and the details are thus omitted.

**Remark 2.2.** Lemma 2.1(i) implies that, under Assumption 2.1(i), to test whether the conditional unconfoundedness assumption is true or not, one can test whether the auxiliary variable $Z$ has explanatory power for the mean of the treatment assignment variable $D$ given covariates $X$. Furthermore, if Assumption 2.1(ii) is also satisfied, Lemma 2.1(ii) implies that the conditional unconfoundedness assumption is equivalent to $E(D|X, Z) = E(D|X)$ a.e., which implies that one can construct a test which is necessary and sufficient for testing the conditional unconfoundedness assumption.

**Remark 2.3.** If the auxiliary variable $Z$ only satisfies the condition $Z \perp D \mid (X, Y(0))$ instead of Assumption 2.1(i), following the same idea of Lemma 2.1, one can construct a test to test whether the potential outcome $Y(0)$ is independent of the treatment variable $D$ conditional on $X$, which is sufficient to identify the average treatment effect for the treated and quantile treatment effect for the treated and their conditional versions.

### 2.2 Test Statistic

Based on the argument in the previous section, it shows that the test of the conditional independence can be transformed into a test of the insignificance of the auxiliary variable. Thus, one can test the conditional independence using the following conditional moment test:

$$H_0 : E(D|X, Z) = E(D|X) \ a.e. \ versus \ H_1 : E(D|X, Z) \neq E(D|X) \quad (2.1)$$

on a set with positive measure. Based on Lemma 2.1(i), if the null hypothesis $H_0$ is rejected, the conditional independence assumption should also be rejected, so it is meaningful to test $H_0$ in practice. Moreover, as pointed out in Remark 2.2, if both Assumption 2.1(i) and Assumption 2.1(ii) are satisfied, our test becomes if and only if.

Various procedures have been proposed to test the hypothesis stated in (2.1), see, for example, to name just a few, Bierens (1982), Bierens et al. (1997), Aït-Sahalia, Bickel and Stoker (2001), Fan and Li (1996), and Li (1999)). Following Fan and Li (1996) and Li
(1999), we adopt the kernel estimation method to construct the test statistic under the null hypothesis. To this end, we first introduce some notations.

Let $W = (X', Z')' \in \mathbb{R}^p$, where $X$ is of dimension $d$ and $Z$ is of dimension $r$, $d + r = p$, and \{\(Y_i, D_i, X_i, Z_i\)\}_{i=1}^n be a set of $n$ independent and identically distributional (iid) observations on $(Y, D, X, Z)$. Define $\varepsilon = D - m(X)$, where $m(X) = E(D|X)$. Then, the null hypothesis testing problem formulated in (2.1) can be rewritten as

$$H_0 : E(\varepsilon W) = 0 \text{ a.e. versus } H_1 : E(\varepsilon |W) \neq 0$$

(2.2)
on a set with positive measure. Note that $T = E[\varepsilon E(\varepsilon|W)] = E\{[E(\varepsilon|W)]^2\} \geq 0$ and the equality holds if and only if $H_0$ is true. Hence, $T$ can be served as a proper candidate for consistent testing $H_0$ and we may use the sample analogue of $T$ to form a test as

$$T_n^* = \frac{1}{n} \sum_{i=1}^n \varepsilon_i E(\varepsilon_i | W_i).$$

However, this test statistic is infeasible because $\varepsilon_i$ and $E(\varepsilon_i | W_i)$ are not observed directly but they can be estimated by some standard nonparametric techniques. To be specific, to obtain a feasible test statistic, we first estimate $\varepsilon_i$ and $E(\varepsilon_i | W_i)$ nonparametrically and then plug the corresponding estimates into the test statistic $T_n^*$ to obtain a feasible version. In order to avoid the random denominator problem, we follow the standard procedure to adopt a density weighted version of $T$, which is $T_n^{**} = \frac{1}{n} \sum_{i=1}^n \left[\varepsilon_i f(X_i)\right] E[\varepsilon_i f(X_i)|W_i] f_W(W_i)$, where $f(\cdot)$ is the density function of $X$ and $f_W(\cdot)$ is the density function of $W$.

Define a leave-one-out kernel estimator of $E(D_i|X_i)$ as

$$\hat{D}_i = \frac{1}{(n-1)h_1^d} \sum_{j \neq i, j=1}^n K_1\left(\frac{X_j - X_i}{h}\right) D_j / \hat{f}(X_i),$$

where

$$\hat{f}(X_i) = \frac{1}{(n-1)h_1^d} \sum_{j \neq i, j=1}^n K_1\left(\frac{X_j - X_i}{h}\right),$$

is the leave-one-out kernel estimator of $f(X_i)$ with $K_1(\cdot)$ being a kernel and $h_1$ denoting the bandwidth. Then, a kernel-based sample analogue of $T$ is given by

$$T_n = \frac{1}{n(n-1)h^p} \sum_{i=1}^n \sum_{j \neq i, j=1}^n (\hat{\varepsilon}_i \hat{f}(X_i))(\hat{\varepsilon}_j \hat{f}(X_j)) K_{ij},$$

where $\hat{\varepsilon}_i = D_i - \hat{D}_i$ is the nonparametric residual estimator and $K_{ij} = K((W_j - W_i)/h)$ with $K(\cdot)$ being a kernel and $h$ denoting the bandwidth.
Before establishing the asymptotic distribution of the test statistic \( T_n \) under \( H_0 \), the following definitions and assumptions are provided. We first introduce two definitions which are also used in Robinson (1988) and Fan and Li (1996). The first one defines a class of higher order kernels, we do not need a higher kernel and the second one defines a class of smooth functions.

**Definition 1.** \( \mathcal{I}_\lambda, \lambda \geq 1 \), is the class of even functions \( \kappa: \mathbb{R} \to \mathbb{R} \) satisfying

\[
\int_{\mathbb{R}} u^i \kappa(u) du = \delta_{i0} \quad (i = 0, 1, \ldots, \lambda - 1),
\]

and

\[
\kappa(u) = O((1 + |u|^{\lambda+1+\delta})^{-1}), \quad \text{for some } \delta > 0,
\]

where \( \delta_{ij} \) is the Kronecker’s delta.

**Definition 2.** \( \varphi^*_\alpha, \alpha > 0, \vartheta > 0 \) is the class of functions \( g: \mathbb{R}^p \to \mathbb{R} \) satisfying: \( g \) is \((l-1)\)-times partially differentiable, for \( l - 1 \leq \vartheta \leq l \); for some \( \rho > 0 \), \( \sup_{y \in \phi_{z \rho}} |g(y) - g(z) - Q_g(y, z) + y - z| < 0 \) for all \( z \), where \( \phi_{z \rho} = \{ y : |y - z| < \rho \} \); \( Q_g = 0 \) when \( l = 1 \); \( Q_g \) is a \((l-1)\)th degree homogeneous polynomial in \( y - z \) with coefficients the partial derivatives of \( g \) at \( z \) of orders 1 through \( m - 1 \) when \( l > 1 \); and \( g(z) \), its partial derivatives of order \( l - 1 \) and less, and \( D_g(z) \), have finite \( \alpha \)th moments.

Then, we provide two sets of assumptions:

**Assumption 2.2.** (i) \( f(\cdot) \in \varphi^*_\nu, m(x) = E(D|X = x) \in \varphi^*_l, \) and \( f_W(w) \in \varphi^*_\nu \) for some \( \nu \geq 2 \) and \( \nu > 0 \).

(ii) Let \( \kappa_1(\cdot) \) be a \( \nu \)-th order kernel and let \( \kappa(\cdot) \) be a nonnegative second order kernel.

(iii) The error \( \varepsilon = D - m(X) \) satisfies \( E(\varepsilon^4) < \infty \). The conditional variance function \( \sigma^2(w) = E(\varepsilon^2|W = w) \) and \( \mu_4(w) = E(\varepsilon^4|W = w) \) are continuous. In addition, \( f_W(w)\sigma^2(w) \) and \( f_W(w)\mu_4(w) \) are bounded on \( \mathbb{R}^p \).

**Assumption 2.3.** As \( n \to \infty, h_1 \to 0, h \to 0, nh_1^d \to \infty, nh^p \to \infty, nh^{p/2}h_1^{2\nu} \to 0 \) and \( h^p/h_1^{2d} \to 0 \).

These assumptions are quite standard and can be seen in many nonparametric test literatures. With Assumptions 2.2 and 2.3, the asymptotic distribution of the test statistic \( T_n \) under \( H_0 \) can be derived, which is formally summarized in the following theorem with its detailed proof given in Appendix A.
Theorem 2.1. Suppose Assumptions 2.2 and 2.3 are satisfied. Then, we have

(1) Under $H_0$, $\tilde{T}_n = \frac{\sqrt{h_p^2 T_n}}{\sqrt{2\sigma_T}} \xrightarrow{d} N(0,1)$, where

$$\hat{\sigma}_T^2 = \frac{1}{n(n-1)h_p^2} \sum_{i=1}^{n} \sum_{j \neq i} \left( \hat{\varepsilon}_i \hat{f}(X_i) \right)^2 \left( \hat{\varepsilon}_j \hat{f}(X_j) \right)^2 K_{ij} \cdot \left( \int K^2(v)dv \right),$$

is a consistent estimator of $\sigma_T^2$ given by

$$\sigma_T^2 = E \left( f^4(X) f_W(W) \sigma^4(W) \right) \left( \int K^2(v)dv \right).$$

(2) Under $H_1$, $P(\tilde{T}_n > Q_n) \rightarrow 1$ for any non-stochastic sequence $Q_n = o(nh_p^2)$.

Theorem 2.1(2) follows from the fact that under $H_1$, $T_n \overset{p}{\rightarrow} E \left[ f_W(W) f^2(X) (E(D|W) - E(D|X))^2 \right] > 0$ and $\sigma_T^2 = O_p(1)$. The proofs of these are straightforward and are thus omitted. Based on Theorem 2.1(1), we can have the following one-sided asymptotic test for $H_0$: rejecting $H_0$ at the significance level $\alpha_0$ if $\tilde{T}_n > c$ where $c$ is the upper $\alpha_0$-percentile of the standard normal distribution.

The theorem shows that the test statistic $\tilde{T}_n$ has asymptotic standard normal distribution under the null hypothesis. However, Monte Carlo simulations reported in Li (1999) and Lavergne and Vuong (2000) reveal that the normal approximation has substantial finite sample bias. Moreover, the test statistic $T_n$ depends on two sets of smoothing parameters $h_1$ and $h_i$, and might be sensitive to the choice of the smoothing parameters. Therefore, Lavergne and Vuong (2000) developed a modified version of the test $T_n$ which has better finite sample performance than the test $T_n$, while Li (1999) adopted this idea to suggest a new modified test and further showed that the modified test has the same asymptotic distribution as $T_n$. Alternatively, one also can use Bootstrapping method to better approximate the null distribution of $T_n$; see Li and Racine (2007) for further details for Bootstrapping method. In this paper, we also adopt the idea of Lavergne and Vuong (2000) and Li (1999) to suggest a modified test statistic. To this send, by substituting $\hat{D}_i$ and $\hat{f}(X_i)$ into the expression of $T_n$ and doing a simplification, then, we can obtain

$$T_n = \frac{1}{n(n-1)^3 h_p h_1^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i} \sum_{l \neq j} (\hat{D}_i - \hat{D}_k)(\hat{D}_j - \hat{D}_l)$$

$$\times K_1 \left( \frac{X_i - X_k}{h_1} \right) K_1 \left( \frac{X_j - X_l}{h_1} \right) K \left( \frac{W_i - W_j}{h} \right).$$

As pointed out in Li (1999), the terms containing squares of the error terms may cause finite sample bias for the test statistic $T_n$, hence we can subtract these terms from $T_n$ and replace $n(n-1)^3$ by $n(n-1)(n-2)(n-3)$ to obtain a new test with possibly smaller finite sample
bias. The modified test is given by
\[ J_n = \frac{1}{n(n-1)(n-2)(n-3)} \left[ n(n-1)^3 T_n - n(n-1)(n-2) A_n - 2n(n-1)(n-2) B_n \right], \]
where
\[ A_n = \frac{1}{n(n-1)(n-2)h_1^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i, j} (D_i - D_k)(D_j - D_k) \times K_1 \left( \frac{X_i - X_k}{h_1} \right) K_1 \left( \frac{X_j - X_k}{h_1} \right) K \left( \frac{W_i - W_j}{h} \right), \]
and
\[ B_n = \frac{1}{n(n-1)(n-2)h_1^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i, j} (D_i - D_j)(D_j - D_k) \times K_1 \left( \frac{X_i - X_j}{h_1} \right) K_1 \left( \frac{X_j - X_k}{h_1} \right) K \left( \frac{W_i - W_j}{h} \right). \]

The following theorem shows that the test statistic \( J_n \) indeed has the same asymptotic distribution as \( T_n \), which is stated here with its proof relegated to Appendix A.

**Theorem 2.2.** Under Assumptions 2.2 and 2.3, we have
(1) Under \( H_0 \), \( \tilde{J}_n = \frac{nh^{p/2} J_n}{\sqrt{\tilde{\sigma}_T^2}} \xrightarrow{d} \mathcal{N}(0, 1) \), where \( \tilde{\sigma}_T^2 \) is the same as defined in Theorem 2.1.
(2) Under \( H_1 \), \( P(\tilde{J}_n > Q_n) \rightarrow 1 \) for any non-stochastic sequence \( Q_n = o(nh^{p/2}) \).

### 3 Extension to Time Series Setting

In this section we extend the proposed procedures to the macroeconomic policy evaluation framework. It is well known that identifying the causal connection between macroeconomic policy and real economic variables is one of the most important and widely studied questions in macroeconomics. Several procedures, for example, dynamic stochastic general equilibrium modeling (DSGE) and vector autoregressive (VAR) models, have been adopted to analyze the macroeconomic policy effects and most of these procedures depend heavily on some structural specifications of the entire economic system. Therefore, the validity of these results relies on how precise the assumed economic models are. Recently, some authors, such as Angrist and Kuersteiner (2011), Angrist et al. (2018), Bojinov and Shephard (2018), Jordà and Taylor (2016), Kuersteiner, Phillips and Villamizar-Villegas (2018), and among others, considered alternative methods to make causal inferences for macroeconomic policy effects using the extended versions of the aforementioned treatment effect approaches. Indeed, Angrist and Kuersteiner (2011) was the first paper to make a strong link between linear impulse response
function (IRF) and the dynamic (time series) treatment effect. Moreover, compared to DSGE models, the identification and estimation of dynamic treatment effects requires no need to specify the structural process of $Y_t$ and only focuses on the policymaking process, alleviating the crucial model misspecification problem faced in the other main macroeconomic models and providing a more flexible tool for the analysis and evaluation of macroeconomic causal relationships. For details, readers are referred to the paper by Angrist and Kuersteiner (2011) and the survey paper by Liu, Cai, Fang and Lin (2019).

3.1 Model Setup and Test Statistic

Causal effects in macroeconomic policy evaluation framework are also defined using the notion as in Rubin (1974) for potential outcomes, but it is complicated by the fact that potential outcomes are determined not just by current policy actions but also by past actions, lagged outcomes, and covariates in this setting. To capture macroeconomic dynamics, following Angrist and Kuersteiner (2011), we suppose that the economy can be described by the observed vector stochastic process $\chi_t = (Y_t, F_t, D_t)$, where $Y_t$ is a vector of outcome variables, $D_t$ is a policy variable that takes values $d_0, \ldots, d_J$ and $F_t$ is a vector of other exogenous and (lagged) endogenous variables. Let $\psi$ be the policy regime, which takes values in a parameter space $\Psi$. The following definition of potential outcomes takes from Angrist and Kuersteiner (2011) and Angrist et al. (2018).

Definition 3. For fixed $t, l$ and $\psi$, potential outcomes $\{Y_{t,l}^{\psi}(d); d \in D\}$ are defined as the set of values the observed outcome variable $Y_{t+1}$ would take on if $D_t = d$, with $d \in D = \{d_0, \ldots, d_J\}$.

For notational simplicity, without loss of generality, we assume that the observed outcome variable $Y_{t+1}$ is of dimension one throughout the paper. Denote $Y_{t,L} = (Y_{t+1}, \ldots, Y_{t+L})'$ and define the vector of potential outcomes up to horizon $L$ by $Y_{t,L}^{\psi}(d) = (Y_{t,1}^{\psi}(d), \ldots, Y_{t,L}^{\psi}(d))$, the observed outcomes then can be written as

$$Y_{t,L} = \sum_{d \in D} Y_{t,L}^{\psi}(d) I\{D_t = d\},$$

where $I\{\cdot\}$ is the indicator function. Therefore, based on these definitions, one can define the dynamic average responses to policy $d_k$ relative to the benchmark policy $d_0$, given by

$$E\left[Y_{t,L}^{\psi}(d_k) - Y_{t,L}^{\psi}(d_0)\right] := \delta_k.$$

Thus, the collection of all possible policy effects is $\delta = (\delta_1, \ldots, \delta_J)'$. Since the potential outcomes for counterfactual policy choices are unobserved, $\delta_k, k = 1, \ldots, J$, can not be
estimated directly. To identify $\delta_k, k = 1, \cdots, J$, similar to the cross-section treatment effects literature, Angrist et al. (2018) adopted the following selection on observables assumption, which was proposed in Angrist and Kuersteiner (2011).

Assumption 3.1. Selection on observables (conditional independence):

$$Y_{t,l}^{\psi}(d_k) \perp D_t | X_t \text{ for all } l \geq 0 \text{ and for all } d_k, \text{ with } \psi \text{ fixed}; \psi \in \Psi,$$

where $X_t$ is a vector of predetermined variables that predict $D_t$.

This assumption was proposed by Angrist and Kuersteiner (2011) to test the causal effects of monetary policy shocks; see Angrist and Kuersteiner (2011) for further details. The selection on observables assumption implies that potential outcomes are independent of the policy variables conditional on some appropriate predetermined variables. Under this assumption, the parameters of interest $\delta_k, k = 1, \cdots, J$, can be identified as

$$\delta_k = E\left\{E\left[Y_{t,l}^{\psi}(d_k) - Y_{t,l}^{\psi}(d_0)\middle| X_t\right]\right\} = E\left\{E\left[Y_{t,l}|D_t = d_k, X_t\right] - E\left[Y_{t,l}|D_t = d_0, X_t\right]\right\}.$$

Furthermore, based on the identified results above, Angrist et al. (2018) proposed the method based on inverse probability weighting (IPW) to estimate $\delta_k$; see Angrist et al. (2018) for more details. As mentioned in Section 2, the selection on observables assumption may not be held in practice if there are some unobserved confounders which affect both the potential outcomes $Y_{t,l}^{\psi}(d_k)$ and treatment variable $D_t$. Therefore, our target in this section is test whether Assumption 3.1 is true or not. Similar to Section 2, the starting point of the procedure is also the availability of a vector of auxiliary variable $Z_t \in \mathbb{R}^r$ satisfying the following assumption.

Assumption 3.2. Suppose that there is a vector of continuously distributed variable $Z_t \in \mathbb{R}^r$ which is correlated with the potential outcomes $Y_{t,l}^{\psi}(d_k)$ and satisfies the following condition

$$Z_t \perp D_t | (Y_{t,l}^{\psi}(d_k), X_t) \text{ for all } l \geq 0 \text{ and for all } d_k, \text{ with } \psi \text{ fixed}; \psi \in \Psi.$$

This assumption is similar to Assumption 2.1(i) in Section 2 and it requires that receiving treatment or not is mainly driven by the potential outcomes. In fact, it may be more easy to find the appropriate auxiliary variable $Z_t$ in time series setting. For example, the lagged variable of the outcome variable $Y_t$ may be used as the proper candidates for the auxiliary variables $Z_t$. Similar to Lemma 2.1 in Section 2, under this assumption, we can obtain the following Lemma.

Lemma 3.1. Under Assumption 3.2, Assumption 3.1 implies that $E(D_t|X_t, Z_t) = E(D_t|X_t)$. 

12
Proof. This Lemma can be proved by the similar arguments to Lemma 2.1 and the details are thus omitted.

Lemma 3.1 implies that to test the selection on observables assumption being true or not, one can test whether $E(D_t|X_t, Z_t) = E(D_t|X_t)$ is true or not in practice. Formally, we consider the following testing hypothesis: $H_0 : E(D_t|X_t, Z_t) = E(D_t|X_t)$ a.e. versus $H_1 : E(D_t|X_t, Z_t) \neq E(D_t|X_t)$ on a set with positive measure. Again, denote $W_t = (X'_t, Z'_t)' \in \mathbb{R}^p$, where $X_t \in \mathbb{R}^d$, $Z_t \in \mathbb{R}^r$ and $p = d + r$ and define $\omega_t = D_t - E(D_t|X_t)$. Similar to the previous section, we propose the following test statistic

$$S_n = \frac{1}{n(n-1)h^p} \sum_{t=1}^{n} \sum_{s \neq t} (\hat{\omega}_t \hat{f}(X_t))(\hat{\omega}_s \hat{f}(X_s)) K_{st},$$

where $\hat{\omega}_t$, $\hat{f}(X_t)$ and $K_{st}$ are defined by a similar manner as in Section 2.

### 3.2 Asymptotic Distribution of $S_n$

In order to establish the asymptotic distribution of the test statistic $S_n$ proposed above, we make the following assumptions.

**Assumption 3.3.** (i) Assume that the process $\{D_t, X_t, Z_t\}_{t=1}^{n}$ is strictly stationary and absolutely regular process with the mixing coefficient $\beta_t \leq C_\beta r^t$ defined by

$$\beta_t = \sup_{s \in \mathbb{N}} E \left[ \sup_{A \in \mathcal{F}_{s+t}^t} \left\{ |P(A|\mathcal{F}_{s}^t) - P(A)| \right\} \right]$$

for all $s, t \geq 1$, where $0 < C_\beta < \infty$ and $0 < \rho < 1$ are constants, and $\mathcal{F}_{s+t}^t$ denotes the $\sigma$-field generated by $\{(X_i, D_i, Z_i) : i \leq t \leq j\}$.

(ii) Assume that $\omega_t = D_t - E(D_t|X_t)$ satisfies for all $t \geq 1$

$$E[\omega_t|\Omega_{t-1}] = 0,$$

where $\Omega_t = \sigma\{(X_{s+1}, D_s) : s \leq t\}$ is a sequence of $\sigma$-field generated by $\{(X_{s+1}, D_s) : s \leq t\}$.

(iii) In addition, assume that

$$E[|\omega_t^{4+\xi}|] < \infty \quad \text{and} \quad E\left[ |\omega_t^{i_1} \omega_t^{i_2} \cdots \omega_t^{i_l}|^{1+\eta} \right] < \infty,$$

for some arbitrarily small $\xi > 0$ and $\eta > 0$, where $2 \leq l \leq 4$ is an integer, $0 \leq i_j \leq 4$ and $\sum_{j=1}^{l} i_j \leq 8$. 

13
Assumption 3.4. (i) Let $\sigma^2(w) = E[\omega^2|W_t = w]$ and $\mu_4(w) = E[\omega^4|W_t = w]$. Assume that $\sigma^2(w)$ and $\mu_4(w)$ satisfy some Lipschitz conditions:

$$|\sigma^2(u + v) - \sigma^2(u)| \leq G(u)||v|| \quad \text{and} \quad |\mu_4(u + v) - \mu_4(u)| \leq G(u)||v||$$

with $E\left[\left|G(W_t)\right|^{2+\zeta}\right] < \infty$ for some small $\zeta > 0$, where $|| \cdot ||$ denotes the Euclidean norm.

(ii) Let $f_{\tau_1,\tau_2,\ldots,\tau_l}(\cdot)$ be the joint probability density of $(W_{1+\tau_1}, \ldots, W_{1+\tau_l})(1 \leq l \leq 4)$. Assume that $f_{\tau_1,\tau_2,\ldots,\tau_l}(\cdot)$ exists, is bounded and satisfies the following Lipschitz condition:

$$|f_{\tau_1,\tau_2,\ldots,\tau_l}(w_1 + v_1, \ldots, w_l + v_l) - f_{\tau_1,\tau_2,\ldots,\tau_l}(w_1, \ldots, w_l)| \leq G_{\tau_1,\tau_2,\ldots,\tau_l}(w_1, \ldots, w_l)||v||,$$

where $G_{\tau_1,\tau_2,\ldots,\tau_l}(w_1, \ldots, w_l)$ is integrable and satisfies the following conditions

$$\int G_{\tau_1,\tau_2,\ldots,\tau_l}(w_1, \ldots, w_l)||w||^{2\zeta}dw < M < \infty,$$

and

$$\int G_{\tau_1,\tau_2,\ldots,\tau_l}(w_1, \ldots, w_l)f_{\tau_1,\tau_2,\ldots,\tau_l}(w_1, \ldots, w_l)dx < M < \infty,$$

for some $\zeta > 1$ and constant $M > 0$.

Assumption 3.5. (i) The density function $f(x)$ and $f_W(w)$ of $X_t$ and $W_t$ satisfy $f(x) \in \varphi^{\infty}$, $f_W(w) \in \varphi^{\infty}$, respectively, and $\gamma(x) = E(D_t|X_t = x) \in \varphi^{\nu+\eta}$ for some integer $\nu \geq 2$.

(ii) $f(x)$, $f_W(w)$ and $\gamma(x)$ all satisfy some Lipschitz conditions:

$$|g(u + v) - g(u)| \leq G(u)||v||,$$

where $G(u)$ has finite $(2 + \eta')$-th moment for some small $\eta' > 0$ and $g(\cdot) = f(\cdot), f_W(\cdot)$ or $\gamma(\cdot)$.

(iii) The product kernel is used for both $K(\cdot)$ and $K_1(\cdot)$. Let $k(\cdot)$ and $k_1(\cdot)$ be their corresponding univariate kernel, then $k_1(\cdot) \in \mathcal{S}_\nu$, $k(\cdot)$ is non-negative and $k(\cdot) \in \mathcal{S}_2$.

(iv) $h_1 \to 0$, $h = O(n^{-\alpha})$ for some $0 < \alpha < \frac{7p}{8}$. Finally, as $n \to \infty$, $h^p/h_1^{2d} = o(1)$, $nh^{p/2}h_1^{2\nu} = o(1)$.

These assumptions are commonly used in nonparametric literatures and the detailed discussions can be found in Li (1999). With Assumptions 3.3-3.5, one can derive the asymptotic distribution of the test statistic $S_n$, which is formally summarized in the following theorem and its The detailed proof given in Appendix B.

**Theorem 3.1.** Assume that Assumptions 3.3-3.5 are satisfied. Then we have
(1) Under $H_0$, $\tilde{S}_n = \frac{nh^{p/2}S_n}{\sqrt{2\sigma_s}} \xrightarrow{d} \mathcal{N}(0,1)$, where

$$\tilde{\sigma}_s^2 = \frac{1}{n(n-1)h^p} \sum_{t=1}^{n} \sum_{s \neq t} \hat{\omega}_t^2 \hat{f}^2(X_t) \hat{\omega}_s^2 \hat{f}^2(X_s) K_{ts}^2,$$

is a consistent estimator of $\sigma_s^2$ given by

$$\sigma_s^2 = E\left[f^4(X_t)f_W(W_t)\sigma^4(W_t)\right] \left( \int K^2(u)du \right).$$

(2) Under $H_1$, $P(\tilde{S}_n > Q_n) \rightarrow 1$ for any non-stochastic sequence $Q_n = o(nh^{p/2})$.

Again, as in Section 2, if we subtract these terms containing squares of the error terms from $S_n$ and replace $n(n-1)^3$ by $n(n-1)(n-2)(n-3)$, then we can obtain a modified test statistic, given by

$$M_n = \frac{1}{n(n-1)(n-2)(n-3)} \left[ n(n-1)^3 S_n - n(n-1)(n-2) A'_n - 2n(n-1)(n-2) B'_n \right],$$

where

$$A'_n = \frac{1}{n(n-1)(n-2)h^p h_1^{2d}} \sum_{t=1}^{n} \sum_{s \neq t} \sum_{l \neq t, l \neq s} \left( D_t - D_l \right) \left( D_s - D_l \right)$$

$$\times K_1 \left( \frac{X_t - X_l}{h_1} \right) K_1 \left( \frac{X_s - X_l}{h_1} \right) K \left( \frac{W_t - W_s}{h} \right),$$

and

$$B'_n = \frac{1}{n(n-1)(n-2)h^p h_1^{2d}} \sum_{t=1}^{n} \sum_{s \neq t} \sum_{l \neq t, l \neq s} \left( D_t - D_s \right) \left( D_s - D_l \right)$$

$$\times K_1 \left( \frac{X_t - X_s}{h_1} \right) K_1 \left( \frac{X_s - X_l}{h_1} \right) K \left( \frac{W_t - W_s}{h} \right).$$

The following theorem shows that both $M_n$ and $S_n$ share the exactly same asymptotic distribution and the detailed proof is relegated to Appendix B.

**Theorem 3.2.** Under Assumptions 3.3-3.5, we have

(1) Under $H_0$, $\tilde{M}_n = \frac{nh^{p/2}M_n}{\sqrt{2\sigma_s}} \xrightarrow{d} \mathcal{N}(0,1)$, where $\tilde{\sigma}_s^2$ is the same as defined in Theorem 3.1.

(2) Under $H_1$, $P(\tilde{M}_n > Q_n) \rightarrow 1$ for any non-stochastic sequence $Q_n = o(nh^{p/2})$.

## 4 Monte Carlo Studies

In this section, we examine the finite-sample performance of the nonparametric tests of $J_n$ and $M_n$ through following two Monte Carlo simulation results with the first example for
the iid setting and the second one for time series content.

**Example 1.** To study the size and power properties of the test statistic $J_n$, the following data generating processes (DGP) is used:

$$Z \sim \mathcal{N}(0, 1), \quad \xi \sim \mathcal{N}(0, 1), \quad X = \gamma Z + \sqrt{1 - \gamma^2} \xi,$$

$$Y(1) = \rho Z + \gamma_1 X + \epsilon_1, \quad Y(0) = \rho Z + \gamma_0 X + \epsilon_0,$$

and

$$D = I\left\{\frac{\mu}{2} (Y(0) + Y(1)) + \sqrt{1 - \mu^2/2} X > U\right\}, \quad U \sim unif(0, 1),$$

where $I\{\cdot\}$ denotes indicator function, $Z$, $\xi$, $\epsilon_1$, $\epsilon_0$ and $U$ are mutually independent random variables and $\epsilon_1 \sim \mathcal{N}(0, 0.4^2)$, $\epsilon_0 \sim \mathcal{N}(0, 0.3^2)$. We set $\gamma_1 = 3.5$ and $\gamma_0 = 4.5$. The constants $\gamma \in [0, 1]$, $\rho \in [0, 1]$ and $\mu \in [0, 1]$ are varied in experiments.

It is easy to see that the DGP above satisfies Assumption 2.1(i) in Section 2 no matter what values of $\mu$ taking. The conditional independence assumption holds only when $\mu$ takes value of zero. We use standard normal kernel functions for both $K_1(\cdot)$ and $K(\cdot)$ with the bandwidth chosen by $h_1 = \hat{\sigma}_X n^{-1/5}$, $h_x = a \cdot \hat{\sigma}_X n^{-1/4}$ and $h_z = a \cdot \hat{\sigma}_Z n^{-1/4}$, where $\hat{\sigma}_X$ and $\hat{\sigma}_Z$ are the sample standard deviations of $\{X_i\}_{i=1}^n$ and $\{Z_i\}_{i=1}^n$, respectively. To check the sensitivity of the tests with respect to different values of the bandwidths, the value of $h_1$ is fixed and $h$ is changed via different values of $a$. We set $a = 0.5$, $1.0$ and $1.5$, respectively. Finally, the number of replications are 2,000 for all cases.

The actual sizes of the $J_n$ test based on asymptotic one-sided normal critical values are reported in Table 1 which presents the empirical rejection probabilities of the test $J_n$ when the different values of bandwidth are considered. Generally speaking, the test works reasonably well in finite samples in various situations. First, the empirical sizes are not sensitive to different choices of the bandwidth. Secondly, the choice of $\gamma$ has very little influence on empirical sizes. Finally, when the sample size is 100, the empirical sizes are a little bit undersized, particularly when $\gamma = 0.8$ and $a = 1.5$, the empirical rejection rate is only 0.039. However, when the sample size increases to 400, the test works very well in all cases.
### Table 1: Estimated sizes of $J_n$ (nominal size $\alpha_0 = 5\%$)

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\gamma(= \rho)$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.040</td>
<td>0.055</td>
<td>0.050</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0</td>
<td>0.048</td>
<td>0.060</td>
<td>0.054</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0</td>
<td>0.042</td>
<td>0.056</td>
<td>0.052</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\gamma(= \rho)$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.045</td>
<td>0.058</td>
<td>0.052</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0</td>
<td>0.043</td>
<td>0.055</td>
<td>0.051</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0</td>
<td>0.044</td>
<td>0.061</td>
<td>0.055</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\gamma(= \rho)$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.043</td>
<td>0.057</td>
<td>0.053</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0</td>
<td>0.040</td>
<td>0.065</td>
<td>0.057</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0</td>
<td>0.039</td>
<td>0.058</td>
<td>0.052</td>
</tr>
</tbody>
</table>

Next, we exam the power curves with nominal size $\alpha_0 = 5\%$ under various cases which are depicted in Figures 1-3 displaying the estimated power curves for the $J_n$ test for different values of the bandwidth respectively. In general, the test $J_n$ has reasonably powerful in detecting the deviation from the null in all cases even when $\mu$ is small. It is not surprising that the powers increase quickly when both the sample size and the value of $\mu$ increase. One of interesting facts is that the power performance depends on the the correlation between the auxiliary variable and potential outcomes and covariates. In all cases, when the values of the $\rho$ and $\gamma$ increase, the powers also increase immensely. It is also noticed from these figures that the higher value is the bandwidth, the larger are the powers of the $J_n$ test. This result can be explained by the fact that the test $J_n$ diverge to $\infty$ at the rate of $nh^{p/2}$ under $H_1$. Therefore, a higher $h$ value (in certain range) should lead to a more powerful test against some fixed alternatives. But this does not mean that one should prefer a higher value of $h$ in practice, since there is always a tradeoff between powers and sizes.
Figure 1: Power curves for test statistic $J_n$ with nominal size $\alpha_0 = 5\%$ and $a = 0.5$. 
Figure 2: Power curves for test statistic $J_n$ with nominal size $\alpha_0 = 5\%$ and $a = 1.0$. 

\rho = 0.0, \gamma = 0.0

\rho = 0.3, \gamma = 0.3

\rho = 0.6, \gamma = 0.6

\rho = 0.9, \gamma = 0.9
Example 2. In this example, we investigate the size and power of the test statistic $M_n$ using the following data generating process:

$$Y_t = \alpha Z_t + \beta Y_{t-1} + \gamma D_t + \varepsilon_t, \quad \text{with} \quad Y_t = D_t Y_t(1) + (1-D_t) Y_t(0),$$

$$D_t = I \left\{ \frac{\mu}{2} (Y_t(1) + Y_t(0)) - \sqrt{1 - \mu^2/2Y_{t-1}} > \eta_t \right\},$$

where $\beta = 0.5$, $\gamma = 0.5$, $Z_t = 0.5 Z_{t-1} + \omega_t$, $\varepsilon_t = 0.5 \varepsilon_{t-1} + v_t$, $\omega_t$, $v_t$ and $\eta_t$ are independent processes with $\omega_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.5^2)$, $v_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.3^2)$ and $\eta_t \overset{i.i.d.}{\sim} \text{unif}(0, 1)$. Again, the con-
stants $\alpha \in [0, 1]$ and $\mu \in [0, 1]$ are varied in experiments. In this example, it is easy to check that the auxiliary variable $Z_t$ always satisfies Assumption 3.2 in Section 3 no matter what values of $\mu$ taking. The selection on observables assumption (Assumption 3.1) is satisfied only when $\mu = 0$ which corresponds to the null model. The correlation between the auxiliary variable $Z_t$ and the potential outcomes $Y_t(1)$ and $Y_t(0)$ is dominated by the constant $\alpha$.

Again we use standard normal kernel functions and the smoothing parameters are chosen using the same methods as in Example 1. In particular, we fix the smoothing parameter $h_1$ and change the smoothing parameter $h$ via different choices of $a$ ($a = 0.5, 1.0, 1.5$). The number of replications are 2,000 for all cases.

Table 2: Estimated sizes of $M_n$ (nominal size $\alpha_0 = 5\%$)

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\alpha$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.035</td>
<td>0.042</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.6</td>
<td>0.036</td>
<td>0.041</td>
<td>0.052</td>
</tr>
<tr>
<td>0.9</td>
<td>0.038</td>
<td>0.045</td>
<td>0.054</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\alpha$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.030</td>
<td>0.035</td>
<td>0.042</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.6</td>
<td>0.033</td>
<td>0.040</td>
<td>0.044</td>
</tr>
<tr>
<td>0.9</td>
<td>0.034</td>
<td>0.044</td>
<td>0.055</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\alpha$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.021</td>
<td>0.026</td>
<td>0.041</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.6</td>
<td>0.028</td>
<td>0.033</td>
<td>0.039</td>
</tr>
<tr>
<td>0.9</td>
<td>0.026</td>
<td>0.037</td>
<td>0.043</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 reports the estimated sizes for $M_n$ for the DGP in Example 2. From Table 2 we observe that the estimated sizes for $M_n$ under-estimates the nominal sizes for almost all cases considered, but the estimated sizes does indicate that as the sample size $n$ increases, the estimated sizes convergent to their nominal sizes. In particular, when the sample size increases to 400, $M_n$ performs well in most cases considered. Also one can see from Table
that the estimated sizes are closer to their nominal sizes for smaller values of \( a \) (for the range of \( a \) values considered), but this does not mean that one should use a very small value of \( h \) in practice. Because too small a \( h \) may cause the kernel estimation to be inaccurate.

Similar to Example 1, Figure 4-6 present the estimated power curves for \( M_n \) for different values of bandwidths, respectively. When the sample size and the value of \( \mu \) increase, the estimated powers also increases sharply. Again one can observe that the power performance relies on the the correlation between the auxiliary variable \( Z_t \) and the potential outcomes measured by \( \alpha \).

Figure 4: Power curves for test statistic \( M_n \) with nominal size \( \alpha_0 = 5\% \) and \( a = 0.5 \).
Figure 5: Power curves for test statistic $M_n$ with nominal size $\alpha_0 = 5\%$ and $\alpha = 1.0$. 
Empirical Rejection Rate

\alpha = 0.3

\mu

Empirical Rejection Rate

\alpha = 0.6

\mu

Empirical Rejection Rate

\alpha = 0.9

\mu

Figure 6: Power curves for test statistic \( M_n \) with nominal size \( \alpha_0 = 5\% \) and \( \alpha = 1.5 \).

5 Real Examples

5.1 The effect of 401(k) participation on asset holdings

It is of interest for policy makers to investigate the effect of the participation into the 401(k) programs on financial assets holdings, see, for example, Abadie (2003), Chernozhukov and Hansen (2004), Ogburn et al. (2015), and Belloni et al. (2017)). The 401(k) plans were introduced in the 1980s with the goal of encouraging retirement saving. Though it is clear that the plans have been widely used as vehicles for retirement saving, their effects on assets holdings is less clear. The main challenge comes from the unobservable saver heterogeneity coupled with nonrandom selection into participation states. To be specific, among people
who are eligible to participate into the 401(k) plans, those who choose to participate are likely to save more than those who choose not. In other words, the underlying saving preference is possible to becomes an unmeasured confounder in the treatment-outcome relationship. Any estimation without dealing with this concern carefully will be biased upward, tending to overestimating the actual effect of the plan participation. To overcome this endogeneity issue, many studies adopted the instrumental variable method and the eligibility of the 401(k) plan was used as the instrumental variable. For example, see Abadie (2003), Chernozhukov and Hansen (2004), Ogburn et al. (2015), Belloni et al. (2017), Poterba and Venti (1994), Poterba et al. (1995), Poterba and Venti (2004), Benjamin (2003), and among others.

We revisit this example by focusing on testing whether the participation in 401(k) programs is independent of potential financial asset holdings conditional on some observed covariates. Even though the 401(k) eligibility is possible to be a valid instrumental variable, it is still worth of considering a conditional independence test because it is well known that the instrumental variable method can only identify the treatment effect of compliers rather than the whole population. Since different instrumental variables drive different compliers, it is hard to compare estimates from different instrumental variables.

To this end, we use the same data as Chernozhukov and Hansen (2004) and Belloni et al. (2017), which consist of a sample of 9,915 observations at the household level drawn from the Survey of Income and Program Participation (SIPP) of 1991. The sample is limited to the households with reference persons between 25 and 64 years old, at least one person being employed, and without self-employed people. We denote $D$ as the treatment variable which is equal to one if one participates into the 401(k) plans and zero otherwise. We also use the same set of covariates $X$ in Chernozhukov and Hansen (2004) and Belloni et al. (2017), which includes age, annual household income (in thousands of dollars), family size, education, marital status, two-earner status, defined benefit (DB) pension status which indicates whether the household’s employer offers a DB pension plan, IRA participation status, and homeownership status. More details about variable definitions can be found in Chernozhukov and Hansen (2004).

Table 3 provides descriptive statistics of household characteristics for the entire sample and split by the participation status. Table 3 shows that, expect age and family size, the means of most characteristics for participants and non-participants are significantly different. Compared to non-participants, the 401(k) plan participants have more incomes, and they are more likely to be married, to have IRAs and DB pensions, to be homeowners, and to be well educated. However, the means of age and family size are similar between the two groups, which implies that age and family size can not be reasons to influence the participation decision. Therefore, we consider to apply our aforementioned test method using age or
family size as the auxiliary variable $Z$.

Table 3: Descriptive statistics (means and standard deviations)

<table>
<thead>
<tr>
<th></th>
<th>Entire sample</th>
<th>By 401(k) participation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Participants</td>
</tr>
<tr>
<td><strong>Outcome variables:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net financial assets</td>
<td>18,051</td>
<td>38,262</td>
</tr>
<tr>
<td></td>
<td>(63,523)</td>
<td>(79,088)</td>
</tr>
<tr>
<td>Total wealth</td>
<td>63,817</td>
<td>96,920</td>
</tr>
<tr>
<td></td>
<td>(111,530)</td>
<td>(127,790)</td>
</tr>
<tr>
<td><strong>Covariates:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>37,201</td>
<td>49,367</td>
</tr>
<tr>
<td></td>
<td>(24,774)</td>
<td>(27,208)</td>
</tr>
<tr>
<td>Age</td>
<td>41.06</td>
<td>41.51</td>
</tr>
<tr>
<td></td>
<td>(10.34)</td>
<td>(9.66)</td>
</tr>
<tr>
<td>Married</td>
<td>0.60</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>(0.49)</td>
<td>(0.46)</td>
</tr>
<tr>
<td>Family size</td>
<td>2.87</td>
<td>2.92</td>
</tr>
<tr>
<td></td>
<td>(1.54)</td>
<td>(1.47)</td>
</tr>
<tr>
<td>Years education</td>
<td>13.21</td>
<td>13.81</td>
</tr>
<tr>
<td></td>
<td>(2.81)</td>
<td>(2.66)</td>
</tr>
<tr>
<td>Participation in IRA</td>
<td>0.24</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>(0.43)</td>
<td>(0.48)</td>
</tr>
<tr>
<td>Home owner</td>
<td>0.64</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>(0.48)</td>
<td>(0.42)</td>
</tr>
<tr>
<td>Defined benefit pension</td>
<td>0.27</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>(0.44)</td>
<td>(0.49)</td>
</tr>
</tbody>
</table>

Note: The sample is drawn from the 1991 SIPP and consists of 9915 observations. The observational units are household reference persons aged 25-64 and spouse if present. Households are included in the sample if at least one person is employed and no one is self-employed. Standard deviations are in parentheses.

Table 4 reports the testing results using either age or family size as the auxiliary variable conditional on various covariates. The standard normal kernel functions is adopted and the bandwidths are chosen as $h_1 = \tilde{\sigma}_X n^{-1/5}$ and $h = 0.5 \tilde{\sigma}_X$, where $\tilde{\sigma}_X n^{-1/4}$ is the sample standard deviation. For all cases, our test very significantly rejects the null hypothesis for different choice of auxiliary variables and various combination of covariates. Our results strongly conclude that the selection on observables assumption is not true for this example so that the classical treatment effect approaches can not be applied. Therefore, it strongly supports using the instrumental variable method.
### Table 4: Results for the conditional unconfoundedness test

<table>
<thead>
<tr>
<th>Auxiliary variables (Z)</th>
<th>Covariates (X)</th>
<th>Test statistic $J_n$ (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Income, Family size, Education</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>Income, Family size, Education, Marital status</td>
<td>0.000</td>
</tr>
<tr>
<td>Family size</td>
<td>Income, Family size, Education, Two-earner status</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Income, Family size, Education, Defined benefit (DB) pension status</td>
<td>0.008</td>
</tr>
<tr>
<td>Family size</td>
<td>Income, Family size, Education, IRA participation status</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Income, Family size, Education, Homeownership status</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Income, Age, Education</td>
<td>0.000</td>
</tr>
<tr>
<td>Family size</td>
<td>Income, Education, Age, Marital status</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Income, Education, Age, Two-earner status</td>
<td>0.003</td>
</tr>
<tr>
<td>Family size</td>
<td>Income, Age, Education, Defined benefit (DB) pension status</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>Income, Education, Age, IRA participation status</td>
<td>0.002</td>
</tr>
<tr>
<td>Family size</td>
<td>Income, Education, Age, Homeownership status</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### 5.2 Return to College Education

The second real example is about return to college education. The endogeneity issue arises because it is concerned to have some unobservable individual abilities that may influence both education and future income. Chen et al. (2017) considered this example and tested the conditional mean independence using a Kolmogorov-Smirnov test with the assumption of symmetric distributions in error terms. Their test can not reject the null hypothesis when some relevant covariates are controlled. We revisit the same issue using the same data in Chen et al. (2017). The data come from the China Health and Nutrition Survey (CHNS) of the year of 2004, 2006, and 2009. The data set includes various provinces in China and consists of 525 individuals aged between 18 and 65 with individual characteristics information including gender, residence type, income, education level and family background. Variable definitions and further details are referred to Table 4 in Chen et al. (2017).

In this example, the outcome variable of interest, denoted by $Y$, is the logarithm of annual income and the treatment variable $D$ is a binary variable which takes a value of 1 for college graduates and 0 otherwise. The covariates $X$ include experience, gender, residence type (urban or rural) and the family background which is represented by parents’ income.
Table 5 shows that most individual characteristics, except age, are very different for people receiving college education and not receiving. However, the average age between the treated and control group is very similar, 29.39 for the treated group and 29.05 for the control group, which motivate us to consider using age as the proper candidate for the auxiliary variable $Z$. For computing our test statistic, we adopt the normal kernel functions and the bandwidths $h$ and $h_1$ are determined by the same procedure used in the previous example.

Table 5: Descriptive statistics (means and standard deviations)

<table>
<thead>
<tr>
<th></th>
<th>Entire sample</th>
<th>Receiving college education</th>
<th>Not receiving college education</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome variables:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>15226.400</td>
<td>17923.910</td>
<td>14157.440</td>
</tr>
<tr>
<td></td>
<td>(11678.750)</td>
<td>(11173.620)</td>
<td>(11716.48)</td>
</tr>
<tr>
<td><strong>Covariates:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experience (Years of working)</td>
<td>10.714</td>
<td>8.456</td>
<td>11.609</td>
</tr>
<tr>
<td></td>
<td>(7.456)</td>
<td>(7.150)</td>
<td>(7.394)</td>
</tr>
<tr>
<td>Age</td>
<td>29.153</td>
<td>29.396</td>
<td>29.057</td>
</tr>
<tr>
<td></td>
<td>(7.331)</td>
<td>(7.123)</td>
<td>(7.418)</td>
</tr>
<tr>
<td>Gender</td>
<td>0.686</td>
<td>0.611</td>
<td>0.715</td>
</tr>
<tr>
<td></td>
<td>(0.464)</td>
<td>(0.489)</td>
<td>(0.452)</td>
</tr>
<tr>
<td>Residence</td>
<td>0.518</td>
<td>0.718</td>
<td>0.438</td>
</tr>
<tr>
<td></td>
<td>(0.500)</td>
<td>(0.451)</td>
<td>(0.496)</td>
</tr>
<tr>
<td>Mother’s income</td>
<td>10806.760</td>
<td>13561.590</td>
<td>9715.091</td>
</tr>
<tr>
<td></td>
<td>(13682.630)</td>
<td>(13198.820)</td>
<td>(13734.490)</td>
</tr>
<tr>
<td>Father’s income</td>
<td>15963.790</td>
<td>20974.080</td>
<td>13978.330</td>
</tr>
<tr>
<td></td>
<td>(16157.910)</td>
<td>(22606.050)</td>
<td>(12214.68)</td>
</tr>
</tbody>
</table>

Table 6 reports the testing results for the conditional independence using age as the auxiliary variable, conditional on different covariates. Our results are basically similar to those in Chen et al. (2017). The test rejects the null hypothesis of conditional unconfoundedness only in the case that a single covariate experience is controlled, with a $p$-value 0.011. However, when more conditioning covariates, such as gender, residence type and parents’ income, are added, our test can not reject the null hypothesis any more. The results also show that the $p$-values increases as more covariates are included in the model, which is in line with the tuition that the conditional unconfoundedness assumption is more likely to hold when more relevant variables are included into the model.
Table 6: Results for the unconfoundedness test

<table>
<thead>
<tr>
<th>Auxiliary variables (Z)</th>
<th>Covariates (X)</th>
<th>Test statistic $J_n$ (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experience</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>Experience, Gender</td>
<td>0.259</td>
</tr>
<tr>
<td>Age</td>
<td>Experience, Gender, Residence</td>
<td>0.276</td>
</tr>
<tr>
<td></td>
<td>Experience, Gender, Residence, Logarithm of mother’s income</td>
<td>0.321</td>
</tr>
<tr>
<td></td>
<td>Experience, Gender, Residence, Logarithm of mother’s income, Logarithm of father’s income</td>
<td>0.401</td>
</tr>
</tbody>
</table>

6 Conclusion

This paper proposes an alternative method to test the conditional independence assumption which is a key vehicle to provide identification power in the literature of policy evaluation. The existing testing methods either depend on the availability of a binary instrumental variable or assume symmetric distributions of the error terms in both the outcome and selection equations. We provide another choice. Our test relies on an auxiliary variable whose potential influence on the treatment decision is fully captured by potential outcomes and observable covariates. In practice, any covariates which are insignificant in the balance check for the treated and control groups can be considered as possible candidates for the auxiliary variable. As extension, we also extend the proposed procedure to time series setting.

We also establish the asymptotic properties of our proposed test. The simulation experiments show that our test works very well even in a small sample size. We finally apply our testing method to two empirical examples, the 401(k) plan participation and return to college education. For the first example, we find strong evidence to support the use of the instrumental variable method. However, for the second example, we find the conditional independence assumption can basically hold when the conditioning variables are appropriately chosen.

References


Mathematical Appendix

A Appendix

This appendix collects proofs of the results stated in Section 2. In this section, the letter \( C \) denotes a generic positive constant whose value can change from context to context.

**Proof of Theorem 2.1:** Recall that \( m_i = m(X_i) = E(D_i|X_i), \varepsilon_i = D_i - m(X_i) \) and

\[
\hat{D}_i = \frac{1}{(n-1)h_i^d} \sum_{j \neq i, j=1}^n K_1 \left( \frac{X_j - X_i}{h_i} \right) D_j / \hat{f}(X_i),
\]

as well as

\[
\hat{f}(X_i) = \frac{1}{(n-1)h_i^d} \sum_{j \neq i, j=1}^n K_1 \left( \frac{X_j - X_i}{h_i} \right).
\]

Define

\[
\tilde{m}_i = \tilde{m}(X_i) = \frac{1}{(n-1)h_i^d} \sum_{j \neq i, j=1}^n K_1 \left( \frac{X_j - X_i}{h_i} \right) m(X_j) / \hat{f}(X_i),
\]

and

\[
\tilde{\varepsilon}_i = \frac{1}{(n-1)h_i^d} \sum_{j \neq i, j=1}^n K_1 \left( \frac{X_j - X_i}{h_i} \right) \varepsilon_j / \hat{f}(X_i),
\]

then we have \( \hat{D}_i = \hat{m}_i + \tilde{\varepsilon}_i \). Using \( \tilde{\varepsilon}_i = D_i - \hat{D}_i = (m_i - \hat{m}_i) + \varepsilon_i - \tilde{\varepsilon}_i \), the test statistic \( T_n \) defined in Section 2 can be rewritten as

\[
T_n = \frac{1}{n(n-1)h^p} \sum_{i=1}^n \sum_{j \neq i, j=1}^n \left[ (m_i - \tilde{m}_i) \hat{f}(X_i) (m_j - \tilde{m}_j) \hat{f}(X_j) + \varepsilon_i \varepsilon_j \hat{f}(X_i) \hat{f}(X_j) + 2 \varepsilon_i \hat{f}(X_i) (m_j - \tilde{m}_j) \hat{f}(X_j) \\
- 2 \varepsilon_i \hat{f}(X_i) (m_j - \tilde{m}_j) \hat{f}(X_j) - 2 \varepsilon_i \hat{f}(X_i) \varepsilon_j \hat{f}(X_j) \right] K_{ij}
\]

where \( K_{ij} \) is defined in Section 2.

We shall complete the proof of Theorem 2.1 by investigating \( T_{ni} \) for \( i = 1, \ldots, 6 \), respectively in the following Lemmas A.1 to A.6. Since the proof is similar to that of Theorem 3.1 of Fan and Li (1996) and Li (1999), we only provide some key steps.

**Lemma A.1.** \( T_{n1} = o_p((nh^p/2)^{-1}) \).

35
where the last equality results from Assumption 2.3. Thus, we have $T_i$ where the inequality follows from Lemma B.2 and Lemma B.3 presented in Appendix B and is the kernel estimator of $f_{ij}$

Proof. First, note that $K(\cdot)$ is a non-negative kernel function, then we can obtain

$$|T_{n1}| = \left| \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} (m_i - \hat{m}_i)\hat{f}(X_i)(m_j - \hat{m}_j)\hat{f}(X_j)K_{ij} \right|$$

$$\leq \frac{1}{2n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \left\{ (m_i - \hat{m}_i)^2 \hat{f}^2(X_i) + (m_j - \hat{m}_j)^2 \hat{f}^2(X_j) \right\} K_{ij}$$

$$= \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} (m_i - \hat{m}_i)^2 \hat{f}^2(X_i)K_{ij} = \frac{1}{n} \sum_{i=1}^{n} (m_i - \hat{m}_i)^2 \hat{f}^2(X_i)\hat{f}_W(W_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[ (m_i - \hat{m}_i)^2 \hat{f}^2(X_i)f_W(W_i) \right] + \frac{1}{n} \sum_{i=1}^{n} \left[ (m_i - \hat{m}_i)^2 \hat{f}^2(X_i)(\hat{f}_W(W_i) - f_W(W_i)) \right]$$

$$\leq C \sum_{i=1}^{n} \left[ (m_i - \hat{m}_i)^2 \hat{f}^2(X_i) \right] + \frac{1}{n} \sum_{i=1}^{n} \left[ (m_i - \hat{m}_i)^2 \hat{f}^2(X_i)(\hat{f}_W(W_i) - f_W(W_i)) \right]$$

$$:= C \cdot T_{n1}^{(1)} + T_{n1}^{(2)}$$

where

$$\hat{f}_W(W_i) = \frac{1}{(n-1)h^p} \sum_{j \neq i, j=1}^{n} K\left( \frac{X_j - X_i}{h}, \frac{Z_j - Z_i}{h} \right)$$

is the kernel estimator of $f_W(W_i)$. For $T_{n1}^{(1)}$, observing that

$$E(T_{n1}^{(1)}) = \frac{1}{n} \sum_{i=1}^{n} E\left[ (m_i - \hat{m}_i)^2 \hat{f}^2(X_i) \right] = E\left[ (m_1 - \hat{m}_1)^2 \hat{f}^2(X_1) \right]$$

$$= \frac{1}{(n-1)^2h_{1}^{2d}} \sum_{i \neq j, j=1}^{n} E\left[ (m_i - m_1)K_{1,i1}(m_j - m_1)K_{1,j1} \right]$$

$$= \frac{1}{(n-1)^2h_{1}^{2d}} \sum_{i \neq j, j=1}^{n} E\left[ E((m_i - m_1)^2K_{1,i1}^2|X_1) \right]$$

$$+ \frac{1}{(n-1)^2h_{1}^{2d}} \sum_{i \neq j, j=1}^{n} E\left[ E((m_i - m_1)K_{1,i1}|X_1) \cdot E((m_j - m_1)K_{1,j1}|X_1) \right]$$

$$\leq \frac{h_{1}^{2d+2\nu}}{(n-1)^2h_{1}^{2d}} \sum_{i \neq j, j=1}^{n} E(M_m(X_i)) + \frac{h_{1}^{2d+2\nu}}{(n-1)^2h_{1}^{2d}} \sum_{i \neq j, j=1}^{n} E\left[ D^2_m(X_1) \right]$$

$$\leq o((nh_1^{d-2})^{-1}) + o(h_1^{2\nu}) = o((nh^{p/2})^{-1})$$

where the inequality follows from Lemma B.2 and Lemma B.3 presented in Appendix B and the last equality results from Assumption 2.3. Thus, we have $T_{n1}^{(1)} = o_p((nh^{p/2})^{-1})$. As far as $T_{n1}^{(2)}$, one only needs to note that $\sup_{w \in W} |\hat{f}_W(w) - f_W(w)| = o_p(1)$ under some regular conditions, where $W$ is the support of $W$. Therefore, we also have $T_{n1}^{(2)} = o_p((nh^{p/2})^{-1})$. This completes the proof of Lemma A.1. 

\[ \square \]
Lemma A.2. \( nh^{p/2}T_{n2} \stackrel{d}{\rightarrow} \mathcal{N}(0, \sigma_{T}^{2}) \), where \( \sigma_{T}^{2} = 2E \left[ f_{W}(W)\sigma^{4}(W)f^{4}(X) \right] \left[ \int K^{2}(u)du \right] \). Furthermore, the variance \( \sigma_{T}^{2} \) can be consistently estimated by \( \hat{\sigma}_{T}^{2} \) given by

\[
\hat{\sigma}_{T}^{2} = \frac{1}{n(n-1)h^{p}} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} (\hat{\varepsilon}_{i} \hat{f}(X_{i}))^{2}(\hat{\varepsilon}_{j} \hat{f}(X_{j}))^{2}K_{ij} \left[ \int K^{2}(u)du \right].
\]

**Proof.** First, \( T_{n2} \) can be decomposed into three parts:

\[
T_{n2} = \frac{1}{n(n-1)h^{p}} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \varepsilon_{i} \varepsilon_{j} \hat{f}(X_{i}) \hat{f}(X_{j})K_{ij}
\]

\[
= \frac{1}{n(n-1)h^{p}} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \varepsilon_{i} \varepsilon_{j} f(X_{i}) f(X_{j})K_{ij}
\]

\[
+ \frac{2}{n(n-1)h^{p}} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \varepsilon_{i} \varepsilon_{j} (\hat{f}(X_{i}) - f(X_{i})) f(X_{j})K_{ij}
\]

\[
+ \frac{1}{n(n-1)h^{p}} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \varepsilon_{i} \varepsilon_{j} (\hat{f}(X_{i}) - f(X_{i})) (\hat{f}(X_{j}) - f(X_{j}))K_{ij}
\]

\[
:= T_{n2}^{(1)} + 2T_{n2}^{(2)} + T_{n2}^{(3)}.
\]

Under the null hypothesis, we show that \( nh^{p/2}T_{n2}^{(1)} \) is normally distributed and \( nh^{p/2}T_{n2}^{(i)} = o_{p}(1) \) for \( i = 2, 3 \). It is easy to see that \( T_{n2}^{(1)} \) can be written in a \( U \)-statistic form

\[
T_{n2}^{(1)} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} H_{n}(R_{i}, R_{j})
\]

with

\[
H_{n}(R_{i}, R_{j}) = \frac{1}{h^{p}} K \left( \frac{X_{j} - X_{i}}{h}, \frac{Z_{j} - Z_{i}}{h} \right) \eta_{i} \eta_{j},
\]

where \( \eta_{i} = \varepsilon_{i} f(X_{i}), \eta_{j} = \varepsilon_{j} f(X_{j}) \) and \( R_{i} = (D_{i}, X_{i}, Z_{i}) \). Under the null hypothesis, we have

\[
E \left( H_{n}(R_{1}, R_{2}) \mid R_{1} \right) = E \left( \frac{1}{h^{p}} K \left( \frac{W_{2} - W_{1}}{h} \right) \eta_{1} \eta_{2} \mid R_{1} \right)
\]

\[
= \frac{1}{h^{p}} \varepsilon_{1} f(X_{1}) E \left( K \left( \frac{W_{2} - W_{1}}{h} \right) \varepsilon_{2} f(X_{2}) \mid D_{1}, W_{1} \right)
\]

\[
= \frac{1}{h^{p}} \varepsilon_{1} f(X_{1}) E \left\{ E \left[ K \left( \frac{W_{2} - W_{1}}{h} \right) \varepsilon_{2} f(X_{2}) \mid D_{1}, W_{1}, W_{2} \right] \mid D_{1}, W_{1} \right\}
\]

\[
= \frac{1}{h^{p}} \varepsilon_{1} f(X_{1}) \left\{ K \left( \frac{W_{2} - W_{1}}{h} \right) f(X_{2}) \right\}
\]

\[
\times E \left( D_{2} - m(X_{2}) \mid D_{1}, W_{1}, W_{2} \right) \mid D_{1}, W_{1}
\]

\[
= 0.
\]
Thus, $T_{n2}^{(1)}$ is a degenerate $U$ statistic. To apply Theorem 1 of Hall (1984) to derive the asymptotic distribution of $T_{n2}^{(1)}$, we need to verify its conditions as in Hall (1984). To this end, define

$$G_n(R_1, R_2) = E \left[ H_n(R_3, R_1) H_n(R_3, R_2) | R_1, R_2 \right].$$

Note that

$$G_n(R_1, R_2) = \frac{1}{h^{2p}} \epsilon_1 f(X_1) \epsilon_2 f(X_2) \left[ K \left( \frac{W_3 - W_1}{h} \right) K \left( \frac{W_3 - W_2}{h} \right) \epsilon_3^2 f^2(X_3) | R_1, R_2 \right].$$

$$= \frac{1}{h^{2p}} \epsilon_1 f(X_1) \epsilon_2 f(X_2) \left[ E \left[ K \left( \frac{W_3 - W_1}{h} \right) K \left( \frac{W_3 - W_2}{h} \right) \epsilon_3^2 f^2(X_3) | R_1, R_2, W_3 \right] | R_1, R_2 \right].$$

$$= \frac{1}{h^{2p}} \epsilon_1 f(X_1) \epsilon_2 f(X_2) \left[ K \left( \frac{W_3 - W_1}{h} \right) K \left( \frac{W_3 - W_2}{h} \right) f^2(X_3) \sigma^2(W_3) | W_1, W_2 \right].$$

$$= \frac{1}{h^{2p}} \epsilon_1 f(X_1) \epsilon_2 f(X_2) \int K \left( \frac{w_3 - W_1}{h} \right) K \left( \frac{w_3 - W_2}{h} \right) f^2(x_3) \sigma^2(w_3) f_W(w_3) dw_3.$$

$$= \frac{1}{h^p} \epsilon_1 f(X_1) \epsilon_2 f(X_2) \int K(u) K \left( u + \frac{W_1 - W_2}{h} \right) f^2(X_1 + hu_x)$$

$$\times \sigma^2(W_1 + hu) f_W(W_1 + uh) du,$$

where $u = (u_x, u_z) \in \mathbb{R}^p$ and $\sigma^2(W_3) = E(\epsilon_3^2 | W_3)$, then, we obtain

$$E \left( G_n^2(R_1, R_2) \right) = \frac{1}{h^{2p}} E \left\{ \epsilon_1^2 f^2(X_1) \epsilon_2^2 f^2(X_2) \left[ \int K(u) K \left( u + \frac{W_1 - W_2}{h} \right) f^2(X_1 + hu_x) \right. \right.$$

$$\left. \times \sigma^2(W_1 + hu) f_W(W_1 + uh) du \right\}^2 \right\}.$$

$$= \frac{1}{h^{2p}} E \left\{ \epsilon_1^2 f^2(X_1) \epsilon_2^2 f^2(X_2) \left( \int K(u) K \left( u + \frac{W_1 - W_2}{h} \right) \right. \right.$$

$$\left. \times f^2(X_1 + hu_x) \sigma^2(W_1 + hu) f_W(W_1 + hu) du \right) \sigma^2(W_1, W_2) \right\}.$$

$$= \frac{1}{h^{2p}} \left\{ \int K(u) K \left( u + \frac{W_1 - W_2}{h} \right) f^2(X_1 + hu_x) \right.$$

$$\left. \times \sigma^2(W_1 + hu) f_W(W_1 + hu) du \right\}^2 f^2(X_1) f^2(X_2) \sigma^2(W_1) \sigma^2(W_2) E(\epsilon_3^2 | W_1, W_2) \right\}.$$

$$= \frac{1}{h^{2p}} \left\{ f^2(X_1) f^2(X_2) \sigma^2(W_1) \sigma^2(W_2) \left( \int K(u) K \left( u + \frac{W_1 - W_2}{h} \right) \right. \right.$$

$$\left. \times f^2(X_1 + hu_x) \sigma^2(W_1 + hu) f_W(W_1 + hu) du \right\}$$

$$= \frac{1}{h^{2p}} \left\{ \int f^2(x_1) f^2(x_2) \sigma^2(w_1) \sigma^2(w_2) \left( \int K(u) K \left( u + \frac{w_1 - w_2}{h} \right) \right. \right.$$

$$\left. \times f^2(x_1 + hu_x) \sigma^2(w_1 + hu) f_W(w_1 + hu) du \right)^2 f_W(w_1) f_W(w_2) dw_1 dw_2 \right\}.$$
that is, by Assumption 2.3. Thus, by Theorem 1 of Hall (1984), it follows that

\[
\frac{1}{h^p} \int f^2(x_1)f^2(x_1 - hv_x)\sigma^2(w_1)\sigma^2(w_1 - hv) \left( \int K(u)K(u + v) \right.
\]

\[
\times f^2(x_1 + hu_x)\sigma^2(w_1 + hu)f_W(w_1 + hu)du \right) f_W(w_1)f_W(w_1 - hv)dw_1dv
\]

\[= O(1/h^p),\]

where \(v = (v_x, v_z), \sigma^2(W_i) = E(\varepsilon_i^2 | W_i), i = 1, 2. \) Also, we can get

\[
E \left[ H_n^2(R_1, R_2) \right] = E \left\{ E \left[ H_n^2(R_1, R_2) | W_1, W_2 \right] \right\}
\]

\[
= E \left\{ E \left[ \frac{1}{h^{2p}} K^2 \left( \frac{W_1 - W_2}{h} \right) \varepsilon_1 f^2(X_1)\varepsilon_2 f^2(X_2) | W_1, W_2 \right] \right\}
\]

\[
= E \left\{ \frac{1}{h^{2p}} K^2 \left( \frac{W_1 - W_2}{h} \right) f^2(X_1)f^2(X_2)\sigma^2(W_1)\sigma^2(W_2) \right\}
\]

\[
= \frac{1}{h^p} \int K^2(u)f^2(x_2 + hu_x)f^2(x_2)\sigma^2(w_2 + hu)\sigma^2(w_2)
\]

\[
\times f_W(w_2 + hu)f_W(w_2)du dw_2
\]

\[
= \frac{1}{h^p} \left\{ \left[ \int f^4(x_2)\sigma^4(w_2)f_W^2(w_2)dw_2 \right] \cdot \left[ \int K^2(u)du \right] + o(1) \right\}
\]

\[
= \frac{1}{h^p} \left\{ E \left( f^4(X)\sigma^4(W)f_W(W) \right) \cdot \left[ \int K^2(u)du \right] + o(1) \right\}
\]

\[
= \frac{1}{h^p} (\sigma_T^4 + o(1)) = O(h^{-p}).
\]

Similarly, by some straightforward calculation, it is easy to obtain that

\[
E \left[ H_n^4(R_1, R_2) \right] = O(1/h^{3p}).
\]

Therefore, we have, as \( n \to \infty, \)

\[
\frac{E \left[ G_n^2(R_1, R_2) \right] + n^{-1} E \left[ H_n^4(R_1, R_2) \right]}{\left( E \left[ H_n^2(R_1, R_2) \right] \right)^2} = \frac{O(1/h^p) + n^{-1}O(1/h^{3p})}{O(1/h^{2p})} \to 0,
\]

by Assumption 2.3. Thus, by Theorem 1 of Hall (1984), it follows that

\[
\frac{nT_{n2}^{(1)}}{\sqrt{2E[H_n^2(R_1, R_2)]}} \overset{d}{\to} \mathcal{N}(0, 1),
\]

that is,

\[
n h^{p/2} T_{n2}^{(1)} \overset{d}{\to} \mathcal{N}(0, \sigma_T^2).
\]
Next, we consider $T^{(2)}_{n2}$. Noting that

$$T^{(2)}_{n2} = \frac{1}{n(n-1)^3 h_1^4 h^p} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i} \varepsilon_i \varepsilon_j (K_{1,ki} - h_1^d f(X_i)) f(X_j) K_{ij},$$

where $K_{1,ki} = K_1\left(\frac{X_k - X_i}{h_i}\right)$, its second moment is

$$E\left(T^{(2)}_{n2}\right)^2 = \frac{1}{n^2(n-1)^6 h_1^2 h^2 p} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i} \sum_{l \neq i} \sum_{k \neq j} \varepsilon_i \varepsilon_j \varepsilon_l \varepsilon_j' (K_{1,ki} - h_1^d f(X_i)) \times f(X_j) K_{ij} (K_{1,k'i'} - h_1^d f(X_{i'})) f(X_{j'}) K_{i'j'}.$$
\[ := C \cdot T_{n2,1}^{(3)} + T_{n2,2}^{(3)}. \]

For \( T_{n2,1}^{(3)} \), we have
\[
E(T_{n2,1}^{(3)}) = \frac{1}{n(n-1)^2 h_1^{2d}} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \sum_{k \neq j, k=1}^{n} E \left\{ \varepsilon_i^2 \left( K_{1,ji} - h_1^d f(X_i) \right) \left( K_{1,ki} - h_1^d f(X_i) \right) \right\}
\]
\[
= \frac{1}{n(n-1)^2 h_1^{2d}} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \sum_{k \neq j, k=1}^{n} E \left\{ \varepsilon_i^2 \left( K_{1,ji} - h_1^d f(X_i) \right)^2 \right\}
\]
\[
+ \frac{1}{n(n-1)^2 h_1^{2d}} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \sum_{k \neq j, k=1}^{n} \sum_{l \neq k, l=1}^{n} E \left\{ \varepsilon_i^2 \left( K_{1,ji} - h_1^d f(X_i) \right) \left( K_{1,ki} - h_1^d f(X_i) \right) \right\}
\]
\[ := A_1 + A_2. \]

Observing that
\[
E \left\{ \varepsilon_i^2 \left( K_{1,21} - h_1^d f(X_1) \right)^2 \right\} = E \left( (K_{1,21} - h_1^d f(X_1))^2 \sigma^2(W_1) \right) = O(h_1^d),
\]

it follows that
\[
A_1 = \frac{1}{n(n-1)^2 h_1^{2d}} \cdot n(n-1) \cdot O(h_1^d) = O((1/nh_1^d)^{-1}) = o((1/nh_1^{p/2})^{-1}),
\]

by assumption \( h_1^d/nh_1^{p/2} = o(1) \). Also observing that
\[
E \left[ \varepsilon_1^2 \left( K_{1,21} - h_1^d f(X_1) \right) \left( K_{1,31} - h_1^d f(X_1) \right) \right]
\]
\[
= h_1^{2d} E \left\{ \left[ \frac{1}{h_1^d} K_1 \left( \frac{X_2 - X_1}{h_1} \right) - f(X_1) \right] \left[ \frac{1}{h_1^d} K_1 \left( \frac{X_3 - X_1}{h_1} \right) - f(X_1) \right] \sigma^2(W_1) \right\}
\]
\[
= h_1^{2d} E \left\{ \left[ E \left( \frac{1}{h_1^d} K_1 \left( \frac{X_2 - X_1}{h_1} \right) \right) \left| X_1 \right. \right] - f(X_1) \right\} \cdot \left[ E \left( \frac{1}{h_1^d} K_1 \left( \frac{X_3 - X_1}{h_1} \right) \right) \left| X_1 \right. \right] - f(X_1) \right\} \sigma^2(W_1)
\]
\[
\leq \frac{Ch_1^{2d}}{2} \sigma^2(W_1)
\]
\[
\leq C h_1^{2d} \cdot h_1^{2\nu} E(D_2^2(X_1))
\]
\[
= O(h_1^{2(d+\nu)}),
\]

where the second inequality follows from Lemma B.1 in Appendix B. Hence, we have
\[
A_2 = \left( n(n-1)^2 h_1^{2d} \right)^{-1} \cdot n(n-1)(n-2) \cdot O(h_1^{2(d+\nu)}) = O(h_1^{2\nu}) = o((nh_1^{p/2})^{-1}),
\]
by assumption \( nh_1^{p/2} h_1^{2\nu} = o(1) \). Therefore, we have \( T_{n2,1}^{(3)} = o_p((nh_1^{p/2})^{-1}) \). For \( T_{n2,2}^{(3)} \), by the same argument of Lemma A.1, we can obtain \( T_{n2,2}^{(3)} = o((nh_1^{p/2})^{-1}) \), thus, \( T_{n2}^{(3)} = o((nh_1^{p/2})^{-1}) \). This completes the proof of Lemma A.2. \( \square \)
Lemma A.3. \( T_{n3} = o_p\left((nh^p/2)^{-1}\right) \).

Proof. First, we have

\[
|T_{n3}| = \left| \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \varepsilon_i \varepsilon_j f(X_i) \hat{f}(X_j) K_{ij} \right|
\]

\[
\leq \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \left[\varepsilon_i^2 f^2(X_i) K_{ij} \right] = \frac{1}{n} \sum_{i=1}^{n} \left[\varepsilon_i^2 f^2(X_i) \hat{f}_W(W_i) \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left[\varepsilon_i^2 \hat{f}^2(X_i) f_W(W_i) \right] + \frac{1}{n} \sum_{i=1}^{n} \left[\varepsilon_i^2 \hat{f}^2(X_i) (\hat{f}_W(W_i) - f_W(W_i)) \right]
\]

\[
\leq \frac{C}{n} \sum_{i=1}^{n} \left[\varepsilon_i^2 \hat{f}^2(X_i) \right] + \frac{1}{n} \sum_{i=1}^{n} \left[\varepsilon_i^2 \hat{f}^2(X_i) (\hat{f}_W(W_i) - f_W(W_i)) \right]
\]

\[
:= C \cdot T_{n3}^{(1)} + T_{n3}^{(2)}.
\]

Note that

\[
E(T_{n3}^{(1)}) = E\left(\varepsilon_i^2 \hat{f}^2(X_1) \right) = \frac{1}{(n-1)^2 h_1^{2d}} \sum_{i \neq 1}^{n} \sum_{j \neq 1}^{n} E\left(\varepsilon_i \varepsilon_j K_{1,i} K_{1,j} \right)
\]

\[
= \frac{1}{(n-1)^2 h_1^{2d}} \sum_{i \neq 1}^{n} \sum_{j \neq 1}^{n} E\left(\varepsilon_i^2 K_{1,i}^2 \right) = \frac{1}{(n-1)^2 h_1^{2d}} E\left(\sigma^2(W_i) K_{1,i}^2 \right)
\]

\[
\leq C \cdot \frac{1}{(n-1)^2 h_1^{2d}} O(h_1^d) = O((nh_1^d)^{-1}) = o((nh^{p/2})^{-1}),
\]

by the assumption \( h^p/h_1^{2d} = o(1) \), which implies \( T_{n3}^{(1)} = o_p\left((nh^p/2)^{-1}\right) \). Similar to the argument of Lemma A.1, we also obtain \( T_{n3}^{(2)} = o_p\left((nh^p/2)^{-1}\right) \). Thus, we have \( T_{n3} = o_p\left((nh^p/2)^{-1}\right) \), which completes the proof of Lemma A.3.

\[\square\]

Lemma A.4. \( T_{n4} = o_p\left((nh^p/2)^{-1}\right) \).

Proof. Note that

\[
T_{n4} = \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \varepsilon_i f(X_i) (m_j - \hat{m}_j) \hat{f}(X_j) K_{ij}
\]

\[
= \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \varepsilon_i f(X_i) (m_j - \hat{m}_j) \hat{f}(X_j) K_{ij}
\]

\[
+ \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \varepsilon_i (\hat{f}(X_i) - f(X_i)) (m_j - \hat{m}_j) \hat{f}(X_j) K_{ij}
\]

\[
:= T_{n4}^{(1)} + T_{n4}^{(2)}.
\]
$T_{n4}^{(1)}$ can be rewritten as

$$T_{n4}^{(1)} = \frac{1}{n(n-1)^2 h^p h^d} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq j} \varepsilon_i f(X_i)(m_j - m_k)K_{1,kj}K_{ij},$$

and its second moment is

$$E(T_{n4}^{(1)})^2 = \frac{1}{n^2(n-1)^4 h^{2p} h^{2d}} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq j} \sum_{l \neq j} \sum_{m \neq l} \varepsilon_i \varepsilon_l f(X_i)f(X_l)(m_j - m_k) \times (m_j - m_k) K_{1,kj}K_{1,lj}K_{ij}K_{lj}.$$ 

It is easy to see that, if $i \neq i'$,

$$E(\varepsilon_i \varepsilon_{i'} f(X_i)f(X_{i'})(m_j - m_k)(m_j - m_k') K_{1,kj}K_{1,k'j}K_{ij}K_{ij'}) = 0.$$ 

If $i = i'$ and all five subscripts $i, j, k, j', k'$ are different, we have

$$E(\varepsilon_i^2 f^2(X_i)(m_j - m_k)(m_j - m_k') K_{1,kj}K_{1,k'j}K_{ij}K_{ij'}) = \sigma^2 W_i f^2(X_i)(m_j - m_k)(m_j - m_k') K_{1,kj}K_{1,k'j}K_{ij}K_{ij'}.$$

$$= E(\sigma^2 W_i f^2(X_i)(m_j - m_k)(m_j - m_k') K_{1,kj}K_{1,k'j}K_{ij}K_{ij'})$$

$$= E(\sigma^2 W_i f^2(X_i)K_{ij}K_{ij'}E[\sum_{m}(m_j - m_k)K_{1,jm}|X_j]E[\sum_{m}(m_j - m_k')K_{1,j'm}|X_{j'}]$$

$$\leq h_1^{2d+2p} E[\sigma^2 W_i f^2(X_i)K_{ij}K_{ij'}D_m(X_j)D_m(X_{j'})$$

$$E[\sigma^2 W_i f^2(X_i)K_{ij}K_{ij'}D_m(X_j)D_m(X_{j'})]$$

where the inequality results from Lemma B.2 in Appendix B. Hence,

$$E(T_{n4}^{(1)})^2 = (n^2(n-1)^4 h^{2p} h^{2d})^{-1} \cdot n^5 \cdot h_1^{2d+2p} O(h^{2p}) = o(n^{-2} h^{-p}).$$

Similarly, one can show that for all other cases, $E(T_{n4}^{(1)})^2 = o(n^{-2} h^{-p})$. Therefore, we have $T_{n4}^{(1)} = o_p((nh^{p/2})^{-1})$.

Also note that

$$E|T_{n4}^{(2)}| \leq \frac{1}{2n(n-1) h^p} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} E[\varepsilon_i^2(\hat{f}(X_i) - f(X_i))^2 K_{ij} + (m_j - \hat{m}_j)^2 \hat{f}^2(X_j)K_{ij}]$$

$$= \frac{1}{2n} \sum_{i=1}^{n} E[\varepsilon_i^2(\hat{f}(X_i) - f(X_i))^2 \hat{f}_W(W_i)] + \frac{1}{2n} \sum_{i=1}^{n} E[(m_j - \hat{m}_j)^2 \hat{f}^2(X_j)\hat{f}_W(W_i)]$$

$$= o((nh^{p/2})^{-1}),$$

by the proofs of Lemma A.1 and A.2. Thus, we have $T_{n4} = o_p((nh^{p/2})^{-1})$.  \qed

43
Lemma A.5. \( T_{n5} = o_p((nh^{p/2})^{-1}) \).

Proof. It is easy to obtain

\[
E|T_{n5}| \leq \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} E \left| \varepsilon_i \hat{f}(X_i)(m_j - \hat{m}_j)\hat{f}(X_j) \right| K_{ij}
\]

\[
\leq \frac{1}{2n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} E \left[ \varepsilon_i^2 \hat{f}^2(X_i)K_{ij} + (m_j - \hat{m}_j)^2 \hat{f}^2(X_j)K_{ij} \right]
\]

\[
= o((nh^{p/2})^{-1}),
\]

by the proofs of Lemmas A.1 and A.3. This completes the proof of \( T_{n5} = o_p((nh^{p/2})^{-1}) \).

Lemma A.6. \( T_{n6} = o_p((nh^{p/2})^{-1}) \).

Proof. First, observing that

\[
T_{n6} = \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i} \varepsilon_i \hat{f}(X_i)\varepsilon_j \hat{f}(X_j)K_{ij}
\]

\[
= \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i} \varepsilon_i f(X_i)\varepsilon_j \hat{f}(X_j)K_{ij}
\]

\[
+ \frac{1}{n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i} \varepsilon_i (\hat{f}(X_i) - f(X_i))\varepsilon_j \hat{f}(X_j)K_{ij}
\]

\[
:= T_{n6}^{(1)} + T_{n6}^{(2)}.
\]

For \( T_{n6}^{(1)} \), its second moment is

\[
E \left( T_{n6}^{(1)} \right)^2 = (n^2(n-1)^4h^{2p}h^{2d})^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq j} \sum_{\nu=1}^{n} \sum_{j' \neq \nu} \sum_{k' \neq j'} E \left[ \varepsilon_i \varepsilon_k \varepsilon_{\nu} \varepsilon_{k'} f(X_i) f(X_{\nu}) \right]
\]

\[
\times K_{1,k_j} K_{1,k_j'} K_{i,j} K_{i,j'}
\]

which is \( o(n^{-2}h^{-p}) \) by the similar discussion of showing \( T_{n2}^{(2)} = o_p((nh^{p/2})^{-1}) \). Thus, \( T_{n6}^{(1)} = o_p((nh^{p/2})^{-1}) \).

For \( T_{n6}^{(2)} \), we have

\[
E|T_{n6}^{(2)}| \leq \frac{1}{2n(n-1)h^p} \sum_{i=1}^{n} \sum_{j \neq i} E \left( \varepsilon_i^2 (\hat{f}(X_i) - f(X_i))^2 K_{ij} + \varepsilon_j^2 \hat{f}^2(X_j)K_{ij} \right)
\]

\[
= \frac{1}{2n} \sum_{i=1}^{n} E \left( \varepsilon_i^2 (\hat{f}(X_i) - f(X_i))^2 \hat{f}_W(W_i) \right) + \frac{1}{2n} \sum_{i=1}^{n} E \left( \varepsilon_j^2 \hat{f}^2(X_j) \hat{f}_W(W_i) \right)
\]

\[
= o((nh^{p/2})^{-1}),
\]

44
by the proofs of Lemmas A.2 and A.3. Therefore, we complete the proof of \( T_{n0} = o_p((Nh^{p/2})^{-1}) \).

Finally, \( \hat{\sigma}_T^2 \) is a consistent estimator for \( \sigma_T^2 \) by the Proposition A.7 in Fan and Li (1996).

**Proof of Theorem 2.2:** Based on the result of Theorem 2.1, it suffices to show that

\[
\frac{n(n-1)(n-2)}{n(n-1)(n-2)(n-3)} A_n = o_p((Nh^{p/2})^{-1}),
\]

and

\[
\frac{n(n-1)(n-2)}{n(n-1)(n-2)(n-3)} B_n = o_p((Nh^{p/2})^{-1}).
\]

From the expression of \( A_n \), we have

\[
\frac{n(n-1)(n-2)}{n(n-1)(n-2)(n-3)} A_n = \frac{1}{n(n-1)(n-2)(n-3)h^p h_1^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i,j} (D_i - D_k) \\
\times (D_j - D_k) K_{1,ik} K_{1,jk} K_{ij}
\]

\[
= \frac{1}{n(n-1)(n-2)(n-3)h^p h_1^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i,j} (m_i - m_k) + (\varepsilon_i - \varepsilon_k)
\]

\[
\times [(m_j - m_k) + (\varepsilon_j - \varepsilon_k)] K_{1,ik} K_{1,jk} K_{ij}
\]

\[
:= M_{n1} + M_{n2},
\]

where \( M_{n1} \) denotes the case which does not have an error term \( \varepsilon \) and \( M_{n2} \) denotes the sum of the remaining terms. It follows that

\[
E[M_{n1}] \leq \frac{1}{n(n-1)(n-2)(n-3)h^p h_1^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i,j} E \left[ (m_i - m_k) K_{1,ik} | (m_j - m_k) K_{ij} \right]
\]

\[
\leq \frac{1}{2n(n-1)(n-2)(n-3)h^p h_1^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i,j} E \left[ (m_i - m_k)^2 K_{1,ik}^2 \right]
\]

\[
+ \frac{1}{2n(n-1)(n-2)(n-3)h^p h_1^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i,j} E \left[ (m_j - m_k)^2 K_{1,jk}^2 \right].
\]

Note that

\[
E \left[ (m_i - m_k)^2 K_{1,ik}^2 K_{ij} \right] = E \left[ K_{ij} \cdot E \left[ (m_i - m_k)^2 K_{1,ik}^2 | X_i \right] \right]
\]

\[
\leq h^{2+d} E \left[ M_{m}(X_i) K_{ij} \right] = h^{2+d} \cdot O(h^p),
\]

45
where the inequality follows from the Lemma B.3 in Appendix B. Hence, we obtain

\[ E |M_{n1}| = O\left( (nh_1^{d-2})^{-1} \right) = o\left( (nh^{p/2})^{-1} \right). \]

For \( M_{n2} \), we first consider the terms with one error term \( \varepsilon \). One such term is

\[ M_{n2}^{(1)} = \frac{1}{n(n-1)(n-2)(n-3)h^p h_1^{2d}} \sum_{i=1}^{n} \sum_{j 
eq i} \sum_{k 
eq i,j} (m_i - m_k) \varepsilon_j K_{1,ik} K_{1,jk} K_{ij}, \]

which is \( o_p\left( (nh^{p/2})^{-1} \right) \) by the Proposition A.5 of Fan and Li (1996). Similar arguments show that all the terms with one error term \( \varepsilon \) is of the order of \( o_p\left( (nh^{p/2})^{-1} \right) \).

Finally, we consider the terms with two error terms. One such term is

\[ M_{n2}^{(2)} = \frac{1}{n(n-1)(n-2)(n-3)h^p h_1^{2d}} \sum_{i=1}^{n} \sum_{j 
eq i} \sum_{k 
eq i,j} (m_i - m_k) \varepsilon_i \varepsilon_j K_{1,ik} K_{1,jk} K_{ij}, \]

which is \( o_p\left( (nh^{p/2})^{-1} \right) \) by the Proposition A.1 of Fan and Li (1996). By the same arguments, one can show that all the other terms which have two error terms are of the order of \( o_p\left( (nh^{p/2})^{-1} \right) \). Thus, we have \( \frac{n(n-1)(n-2)}{n(n-1)(n-2)(n-3)} A_n = o_p\left( (nh^{p/2})^{-1} \right) \). Similarly, one can also show that \( \frac{n(n-1)(n-2)}{n(n-1)(n-2)(n-3)} B_n = o_p\left( (nh^{p/2})^{-1} \right) \). Therefore, we have \( \tilde{T}_n = \frac{nh^{p/2} J_n}{\sqrt{2} \sigma_T} = \tilde{T}_n + o_p(1) \xrightarrow{d} \mathcal{N}(0,1). \)

\section{Appendix}

\textbf{Proof of Theorem 3.1:} As in Appendix A, the test statistic \( S_n \) defined in Section 3 is rewritten as

\[ S_n = \frac{1}{n(n-1)h^p} \sum_{t=1}^{n} \sum_{s 
eq t,s=1}^{n} \left[ (\gamma_t - \gamma_t) \hat{f}(X_t)(\gamma_s - \gamma_s) \hat{f}(X_s) + \omega_t \omega_s \hat{f}(X_t) \hat{f}(X_s) + \omega_t \omega_s \hat{f}(X_t) \hat{f}(X_s) + 2 \omega_t \hat{f}(X_t)(\gamma_s - \gamma_s) \hat{f}(X_s) - 2 \omega_t \hat{f}(X_t) \omega_s \hat{f}(X_s) \right] K_{ts} \]

\[ := S_{n1} + S_{n2} + S_{n3} + 2S_{n4} - 2S_{n5} - 2S_{n6}, \]

where \( \gamma_t = \gamma(X_t) = E(D_t|X_t) \), \( \omega_t = D_t - \gamma_t \),

\[ \hat{f}(X_t) = \frac{1}{(n-1)h_1^d} \sum_{s 
eq t} K_1 \left( \frac{X_s - X_t}{h_1} \right), \]

\[ 46 \]
Lemma B.1. Under Assumptions 3.3-3.5, we have

\[ S_{n1} = o_p\left((nh^{p/2})^{-1}\right), \quad S_{n3} = o_p\left((nh^{p/2})^{-1}\right), \]
and

\[ S_{n4} = o_p\left((nh^{p/2})^{-1}\right), \quad S_{n5} = o_p\left((nh^{p/2})^{-1}\right), \quad S_{n6} = o_p\left((nh^{p/2})^{-1}\right). \]


Lemma B.2. (i) \( nh^{p/2}S_{n2} \xrightarrow{d} \mathcal{N}(0, 2\sigma_S^2) \), where \( \sigma_S^2 = E\left[f^4(X_t)f_W(W_t)\sigma^4(W_t)\right] \left( \int K^2(v)dv \right) \).

(ii) \( \tilde{\sigma}_S^2 = \sigma_S^2 + o_p(1) \).

Proof. (i) First note that \( S_{n2} \) is rewritten as

\[
S_{n2} = \frac{1}{n(n-1)h^p} \sum_{t=1}^{n} \sum_{s \neq t} \omega_s \omega_s \tilde{f}(X_t) \tilde{f}(X_s) K_{ts} \\
= \frac{1}{n(n-1)h^p} \sum_{t=1}^{n} \sum_{s \neq t} \omega_s \omega_s f(X_t) f(X_s) K_{ts} \\
+ \frac{2}{n(n-1)h^p} \sum_{t=1}^{n} \sum_{s \neq t} \omega_s \omega_s (\tilde{f}(X_t) - f(X_t)) f(X_s) K_{ts} \\
+ \frac{1}{n(n-1)h^p} \sum_{t=1}^{n} \sum_{s \neq t} \omega_s \omega_s (\tilde{f}(X_t) - f(X_s)) (\tilde{f}(X_s) - f(X_s)) K_{ts} \\
:= S_{n2}^{(1)} + 2S_{n2}^{(2)} + S_{n2}^{(3)}.
\]

By Lemma A.2 in Li (1999), we know that \( S_{n2}^{(k)} = o_p\left((nh^{p/2})^{-1}\right), k = 2, 3 \). Therefore, it remains to establish the asymptotic normality of \( nh^{p/2}S_{n2}^{(1)} \). Here we use Lemma 3.2 in Hjellvik et al. (1998) instead of the central limit theorem in Theorem 2.1 in Fan and Li (1999) for the degenerate U-statistic to show that \( nh^{p/2}S_{n2}^{(1)} \) is normally distributed. To this end, let \( \xi_t = (W_t, \omega_t) \), \( P(\xi_s), P(\xi_s, \xi_t), P(\xi_s, \xi_t, \xi_k) \), and \( P(\xi_s, \xi_t, \xi_l, \xi_k) \) be the probability measures of \( \xi_s, (\xi_s, \xi_t), (\xi_s, \xi_t, \xi_l) \) and \( (\xi_s, \xi_t, \xi_l, \xi_k) \) for different \( s, t, l, k \in \{1, \cdots, n\} \), respectively. Let \( \varphi_{st} = \varphi_{st}(\xi_t, \xi_s) = \omega_s \omega_t f(X_t) f(X_s) K_{ts}/(n(n-1)h^p) \). It is easy to see that \( \varphi_{st} \) is a symmetric function on its arguments. Then, \( S_{n2}^{(1)} = \sum_{1 \leq t \neq s \leq n} \varphi_{st} = 2 \sum_{1 \leq s < t \leq n} \varphi_{st} \) is a degenerate U-
statistics. Let \( \sigma^2_{st} = \text{Var}(\varphi_{st}) \) and \( \sigma^2_n = \sum_{1 \leq s < t \leq n} \sigma^2_{st} \). Define for some constant \( \delta > 0 \),

\[
M_{n1} = \max_{1 < s < t \leq n} \max \left\{ E |\varphi_{st}|^{1+\delta}, \int |\varphi_{st}|^{1+\delta} dP(\xi_1)dP(\xi_s, \xi_t) \right\},
\]

\[
M_{n2} = \max_{1 < s < t \leq n} \max \left\{ E |\varphi_{st}|^{2(1+\delta)}, \int |\varphi_{st}|^{2(1+\delta)} dP(\xi_1)dP(\xi_s, \xi_t), \right. \nonumber \\
\left. \int |\varphi_{st}|^{2(1+\delta)} dP(\xi_1)dP(\xi_s) dP(\xi_t), \int |\varphi_{st}|^{2(1+\delta)} dP(\xi_1)dP(\xi_s) dP(\xi_t) \right\},
\]

\[
M_{n3} = \max_{1 < s < t \leq n} E |\varphi_{st}|^2, \quad M_{n4} = \max_{1 < s, t, l \leq n} \left\{ \max_{s, t, l \text{ different}} P \int |\varphi_{st}|^{2(1+\delta)} dP \right\},
\]

\[
M_{n5} = \max_{1 < s < t \leq n} \max \left\{ E \left| \int \varphi_{st} dP(\xi_1) \right|^{2(1+\delta)}, \right. \nonumber \\
\left. \int \left| \int \varphi_{st} dP(\xi_1) \right|^{2(1+\delta)} dP(\xi_s) dP(\xi_t) \right\},
\]

\[
M_{n6} = \max_{1 < s < t \leq n} E \left| \int \varphi_{st} dP(\xi_1) \right|^2,
\]

where the maximization over \( P \) in the equation for \( M_{n4} \) is taken over the four probability measures \( P(\xi_1, \xi_s, \xi_t, \xi_t), P(\xi_1)P(\xi_s, \xi_t, \xi_t), P(\xi_1)P(\xi_s)P(\xi_t, \xi_t) \) and \( P(\xi_1)P(\xi_s)P(\xi_t, \xi_t) \) for mutually different \( s, t, l \). We assume that all of the above constants are finite.

According to Lemma 3.2 of Hjellvik et al. (1998), \( \sigma^{-1}_n \sum_{1 \leq s < t \leq n} \varphi_{st} = \frac{1}{2} \sigma^{-1}_n \sigma^{(1)}_{n2} \) is asymptotically normal with mean value 0 and variance 1 if for some \( \delta > 0 \), as \( n \to \infty \),

\[
\max \sigma^{-2}_n \left\{ n^2 \left( M_{n1}^{1/(1+\delta)} + M_{n5}^{1/(1+\delta)} + M_{n6}^{1/2} \right), n^{3/2} \left( M_{n2}^{1/(1+\delta)} + M_{n3}^{1/2} + M_{n4}^{1/2(1+\delta)} \right) \right\} \to 0.
\]

To do so, we evaluate only the order of magnitude of \( M_{n1}^{1/(1+\delta)} \), as the other terms can be investigated in a similar fashion. We first work on \( M_{n1} \). Denote \( u_t = \omega_t f(X_t) \) and \( p_{st} = K_{st} / (n(n-1)h^p) \), then \( \varphi_{st} = u_t u_s p_{st} \). Applying Hölder’s inequality, it follows that for some \( 0 < \delta < 1 \) and \( 1 \leq s < t < l \leq n \)

\[
E \left[ |\varphi_{st}|^{1+\delta} \right] = E \left[ |u_t u_s u_t p_{st}|^{1+\delta} \right] \nonumber \\
\leq \left\{ E \left[ |u_t u_s u_t|^2 \right] \right\}^{\frac{1}{2(1+\delta_2)}} \left\{ E \left[ |p_{st}|^{1+\delta_1} \right] \right\}^{\frac{1}{1+\delta_1}} \nonumber \\
\leq \left\{ E \left[ |u_t u_s u_t|^2 \right] \right\}^{\frac{1}{2(1+\delta_2)}} \left\{ E \left[ p_{st}^{1+\delta_1} \right] \right\}^{\frac{1}{1+\delta_1}},
\]

where \( 0 < \delta_1 < 1 \) and \( 0 < \delta_2 < 1 \) satisfy \( \frac{1}{1+\delta_1} + \frac{1}{2(1+\delta_2)} = 1 \) and \( \frac{1+\delta}{3-\delta} < \delta_1 < \frac{1-\delta}{1+\delta} \). It is important to note that

\[
1 < \zeta = (1+\delta)(1+\delta_2) < 2 \quad \text{and} \quad 1 < \eta = (1+\delta)(1+\delta_1) < 2,
\]
so we have that
\[ \left\{ E\left[u_1 u_s u_t^2\right]^{2(1+\delta)(1+\delta_2)} \right\}^{\frac{1}{2(1+\delta_2)}} = \left\{ E\left[u_1^2 u_s^2 u_t^4 \right]^{(1+\delta)(1+\delta_2)} \right\}^{\frac{1}{2(1+\delta_2)}} < \infty \]

by Assumptions 3.3(iii) and 3.5(i). By straightforward calculations, we have
\[
E\left(\left|p_{1t}p_{st}\right|^\eta \right) = \frac{1}{(n(n-1)h^p)^{2\eta}} \int \int \int K\left(\frac{u-r}{h}\right)^{\eta} K\left(\frac{v-r}{h}\right)^{\eta} f(u,v,r) dudvdr
= \frac{1}{(n(n-1)h^p)^{2\eta}} \int \int \int K(x)K(y)^{\eta} f(z+hx,z+hy,z) dxdydz
\leq C \frac{h^{2p}}{(n(n-1)h^p)^{2\eta}}
\]

under Assumption 3.4(ii), where \( f(u,v,r) \) is the joint density function of \((W_1, W_s, W_t)\).

Also, note that
\[
\sigma_n^2 = \frac{1}{2} \sum_{t=1}^{n} \sum_{s \neq t} \text{Var}(\varphi_{st}) = \frac{1}{2} \sum_{t=1}^{n} \sum_{s \neq t} E(\varphi_{st}^2)
= \frac{1}{2(n(n-1)h^p)^2} \sum_{t=1}^{n} \sum_{s \neq t} E\left(\omega_t^2 \omega_s^2 f(X_t) f^2(X_s) K_{ts}^2\right)
= \frac{1}{2n(n-1)h^p} \left(\sigma_S^2 + o(1)\right),
\]

where the last equality follows from Lemma A.2(ii) in Li (1999). Therefore, for any \(1 < s < t \leq n\), we have
\[
n^2 \left[ E\left|\varphi_{1t}\varphi_{st}\right|^{1+\delta} \right] \overset{\text{in prob}}{\longrightarrow} \sigma^2_n \leq \frac{C^{1/\eta} n^{2h^{2p}/\eta} n(n-1)h^p}{(n(n-1)h^p)^2(\sigma^2_S + o(1))} = \frac{C^{1/\eta} h^{2p}/\eta - p}{\sigma^2_S + o(1)} \to 0,
\]

because of \(\eta < 2\).

Now let us consider the second term in \(M_{n1}\). Let \(E_i\) and \(E_{ij}\) be expectations with respect to \(\xi_i\) and \((\xi_i, \xi_j)\), respectively. Then, we have
\[
E_1 E_{st} \left[ \left|\varphi_{1t}\varphi_{st}\right|^{1+\delta} \right] = \int \left|\varphi_{1t}\varphi_{st}\right|^{1+\delta} dP(\xi_1) dP(\xi_s, \xi_t)
\leq C E_1 \left\{ \left[\omega_1\right]^{1+\delta} E_{st} \left[ \left|\omega_s^2 \omega_t^2 p_{1tp_{st}}\right|^{1+\delta} \right] \right\}
\leq C E_1 \left[ \left[\omega_1\right]^{1+\delta} \left\{ E_{st} \left[ \left|\omega_s^2 \omega_t^2\right|^{2(1+\delta)(1+\delta_2)} \right] \right\}^{\frac{1}{1+\delta_2}} \left\{ E_{st} \left[ \left|p_{1tp_{st}}\right|^{(1+\delta)(1+\delta_1)} \right] \right\}^{\frac{1}{1+\delta_1}} \right]
\]

49
\[
\leq \left\{ \frac{Ch^{2p}}{(n(n-1)h^p)^{2\eta}} \right\}^{\frac{1}{1+\delta_1}}.
\]

Therefore,
\[
\sigma_n^{-2}n^2\left\{ \int |\varphi_{1t}\varphi_{st}|^{1+\delta}dP(\xi_1)dP(\xi_s, \xi_t) \right\}^{\frac{1}{1+\delta}} \leq \sigma_n^{-2}n^2\left\{ \frac{Ch^{2p}}{(n(n-1)h^p)^{2\eta}} \right\}^{\frac{1}{2}} \to 0,
\]
as \(n \to \infty\). Thus, we complete the proof of \(\sigma_n^{-2}n^2M_{a1}^{1/(1+\delta)} \to 0\), as \(n \to \infty\).

For the \(M_{a2}\) part, we only consider \(E|\varphi_{1t}\varphi_{st}|^{2(1+\delta)}\) and the other terms can be considered similarly. Note that
\[
\left[ E|\varphi_{1t}\varphi_{st}|^{2(1+\delta)} \right]^{\frac{1}{1+\delta}} \leq \frac{Ch^{2p/\zeta}}{(n(n-1)h^p)^2},
\]
where \(2 < \zeta = 2(1+\delta)(1+\delta_1) < 4\), thus, we have, as \(n \to \infty\)
\[
\sigma_n^{-2}n^{3/2}\left[ E|\varphi_{1t}\varphi_{st}|^{2(1+\delta)} \right] \leq \frac{1}{n^{1/2}h^{p-2p/\zeta}} \frac{1}{\sigma_S^2 + o(1)} \to 0,
\]
by the assumption \(nh^p \to \infty\).

Similarly, one can verify the other terms are true and the details thus are omitted. Therefore, combined Lemma B.1 and Lemma B.2, we can obtain Theorem 3.1. \(\square\)

**Proof of Theorem 3.2:** See the proof of Corollary 3.2 in Li (1999).

### C Appendix

The following Lemmas given below are taken from Robinson (1988), Fan and Li (1996) and Li (1999), which are used repeatedly in deriving the asymptotic properties for the above Lemmas. The proofs can be found in Robinson (1988) and Fan and Li (1996).

**Lemma C.1.** Let \(f(\cdot) \in \mathcal{G}_\nu\) and \(K_1(\cdot) \in \mathcal{S}_\nu\), where \(\nu \geq 2\) is an integer, \(X \in \mathbb{R}^d\), \(h_1 \to 0\) as \(n \to \infty\). Then
\[
\left| E\left[ K_1\left( \frac{X-x}{h_1} \right) - h_1^{d+\nu}Df(x) \right] \right| \leq h_1^{d+\nu}Df(x), \text{ uniformly in } x.
\]
where \(Df(\cdot)\) has finite \(\alpha\)th moments.

**Lemma C.2.** Suppose \(m(\cdot) \in \mathcal{G}_\nu\) and \(K_1(\cdot) \in \mathcal{S}_\nu\), where \(\nu \geq 2\) is an integer, \(X \in \mathbb{R}^d\),
$h_1 \to 0$ as $n \to \infty$. Then

$$\left| E \left[ K_1 \left( \frac{X - x}{h_1} \right) (m(X) - m(x)) \right] \right| \leq h_1^{d+\nu} D_m(x), \text{ uniformly in } x.$$ 

where $D_m(\cdot)$ has finite $\alpha$th moments.

**Lemma C.3.** For $\delta \geq 1$, let $m(\cdot) \in \varphi_\nu^{2\delta}$ and suppose $\sup_u [u^{\delta+d} K_1^\delta(u)] < \infty$. Then

$$\left| E \left( m(X_2) - m(X_1) \right)^\delta K_1^\delta |X_1| \right| \leq M_m(X_1) h^{\delta+d},$$

where $M_m(\cdot)$ has finite second moment and $d$ is the dimension of $X_1$. 
