

# Specification Testing for Transformation Models with an Application to Generalized Accelerated Failure-time Models\*

Arthur Lewbel,<sup>a</sup> Xun Lu<sup>b</sup> and Liangjun Su<sup>c</sup>

<sup>a</sup> Department of Economics, Boston College

<sup>b</sup> Department of Economics, Hong Kong University of Science and Technology

<sup>c</sup> School of Economics, Singapore Management University

First version: May 2012. Current revision: August 2014

## Abstract

This paper provides a nonparametric test of the specification of a transformation model. Specifically, we test whether an observable outcome  $Y$  is monotonic in the sum of a function of observable covariates  $X$  plus an unobservable error  $U$ . Transformation models of this form are commonly assumed in economics, including, e.g., standard specifications of duration models and hedonic pricing models. Our test statistic is asymptotically normal under local alternatives and consistent against nonparametric alternatives violating the implied restriction. Monte Carlo experiments show that our test performs well in finite samples. We apply our results to test for specifications of generalized accelerated failure-time (GAFT) models of the duration of strikes.

**Keywords:** additivity, control variable, endogenous variable, monotonicity, nonparametric nonseparable model, hazard model, specification test, transformation model, unobserved heterogeneity

**JEL Classification:** C12, C14

## 1 Introduction

Consider a scalar observable outcome  $Y$ , a  $d_x \times 1$  vector of observable covariates of interest  $X$ , and a scalar unobservable cause or error  $U$ . Our goal of this paper is to test the following hypotheses:<sup>1</sup>

$\mathbb{H}_{10}$  : There exist two measurable functions  $G : \mathbb{R} \rightarrow \mathbb{R}$  and  $H_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$   
such that  $Y = G[H_1(X) + U]$  a.s., and  $G$  is strictly monotonic;

$\mathbb{H}_{1A}$  :  $\mathbb{H}_{10}$  is false

---

\*Halbert White inspired this project, brought us together to work on it, and provided substantial advice, discussion, and enthusiasm. We deeply mourn his passing. We gratefully thank the co-editor Jianqing Fan, the associate editor, and three anonymous referees for their many constructive comments. We are also indebted to Songnian Chen, Yanqin Fan, and Shakeeb Khan for helpful comments and suggestions. Lu acknowledges support from Hong Kong University of Science and Technology under grant number FSECS13BM02. Su acknowledges support from the Singapore Ministry of Education for Academic Research Fund under grant number MOE2012-T2-2-021.

<sup>1</sup>The error term  $U$  is a scalar under the null  $\mathbb{H}_{10}$ , however, under the alternative  $\mathbb{H}_{1A}$ ,  $Y$  could be a function of  $X$  and a vector of unobservable errors  $\mathbb{U}$ . For example, the alternative might include models with random coefficients. More generally, we could write  $\mathbb{H}_{1A}$  as  $Y = R(X, \mathbb{U})$ , and then under the null there would exist a scalar valued function  $H_2$  such that  $U = H_2(\mathbb{U})$ . We later define a more specific set of alternatives that our test has power against.

Specifications that are monotonic functions of additive models have been called “transformation models” (e.g., Chiappori et al., 2013), or “transformed additively separable models” (e.g., Jacho-Chávez et al., 2010), or “generalized additive models with unknown link function” (e.g., Horowitz, 2001, and Horowitz and Mammen, 2004).

Broadly speaking, there are two kinds of transformation models that are common in the economics literature. The first type assumes that  $Y$  and  $X$  are observable,  $U$  is unobservable, and the link function  $G(\cdot)$  may be known or unknown. Our paper belongs to this category. Ridder (1990), Horowitz (1996), Ekeland et al. (2004), Ichimura and Lee (2011), and Chiappori et al. (2013) discuss identification and estimation for transformation models of this category. In this class of models, the functions  $G$  and  $H_1$  and the distribution of  $U$  are identified and estimated. In the second type of transformation model, both  $X$  and  $U$  are observable, and  $Y$  is an object that can be estimated such as a conditional mean or quantile function. Horowitz (2001), Horowitz and Mammen (2004, 2007, 2011), Horowitz and Lee (2005), and Jacho-Chávez et al. (2010) provide identification and estimation results for this second kind of transformation model, while Gozalo and Linton (2001) consider specification tests for such models. See also Horowitz (2014) for a recent survey on the latter class of models.

The transformation models under our null are commonly used (and hence assumed to hold) in a wide range of economic applications. For example, they are often used to study duration data (see, e.g., Heckman and Singer, 1984, Keifer, 1988, Mata and Portugal, 1994, Engle, 2000, and Abbring et al., 2008). In particular generalized accelerated failure-time (GAFT) models, which includes accelerated failure-time (AFT) models, proportional hazard (PH) models, and mixed proportional hazard (MPH) models as special cases, are all examples of models that satisfy our null hypothesis. The MPH specification in particular is a widely used class of duration data specifications (for a review, see Van den Berg, 2001).

Despite its popularity, economic theory rarely justifies the MPH or other GAFT specifications. For example, Van den Berg (2001, p. 3400) points out that “the MPH model specification is not derived from economic theory and it remains to be seen whether the MPH specification is actually able to capture important theoretical relations.” He also provides some specific economic examples where the MPH specification is violated. In their microeconometrics textbook, Cameron and Trivedi (2005, p. 613) say that “the multiplicative heterogeneity assumption [in MPH models] is also rather special, but it is mathematically convenient...” Given the popularity (and the limitations) of GAFT models, especially MPH models, it is obvious that a formal specification test of these models would be useful for empirical research. While some specification tests for certain parametric forms of duration models exist (see, e.g., Fernandes and Grammig, 2005), to the best of our knowledge, ours is the first that specifically tests for some testable implications of the general specification of GAFT models.<sup>2</sup>

Another major set of applications of transformation model specifications where  $U$  is unobservable are hedonic models (see, e.g., Ekeland et al., 2004, and Heckman et al., 2005). Here again, we believe that our paper is the first to provide a general specification test for this class of transformation models.

A conditional exogeneity assumption is imposed to test  $\mathbb{H}_{10}$ , i.e., we assume that  $U$  and  $X$  are conditionally independent, conditioning on an observable covariate vector  $Z$ . This is analogous to the conditional unconfoundedness assumption in the treatment effect literature, and to the assumptions required for use of control function type methods of dealing with endogeneity (see, e.g., Heckman and Robb 1986 and Blundell and Powell, 2003. In a control function setting  $Z$  would be the errors obtained after regressing endogenous

---

<sup>2</sup>Recently, Chiappori et al. (2013) provide a nonparametric test, not for the transformation model specification itself, but for a conditional exogeneity assumption within the context of a transformation model. Still, their test might be interpreted as a model specification test. See Remark 2.6 in Section 2.1 for details.

elements of  $X$  on a vector of instruments). Chiappori et al. (2013) provide a nonparametric estimator for the transformation model under similar assumptions. Our test allows for a covariate vector  $Z$ , but unlike some other estimators and tests (see below), we do not require a vector  $Z$ , i.e., our test can also be applied when  $U$  and  $X$  are unconditionally independent and no other relevant covariates are observed.

We first show that if the data are generated by a transformation model, i.e., if  $\mathbb{H}_{10}$  holds, then the ratio of the derivatives with respect to  $Y$  and to  $X$  of the conditional CDF of  $Y$  given  $(X, Z)$  can be written as a product of functions of  $X$  and  $Y$ .<sup>3</sup> We then use local polynomial methods to estimate these derivatives, and construct test statistics based on the  $L_2$  distance between restricted and unrestricted estimators of this ratio of derivatives. We show that our test statistic is asymptotically normal under the null and under a sequence of Pitman local alternatives and is consistent against the alternatives violating the implied restriction. To facilitate the application of our test, we use subsampling to obtain  $p$ -values or critical values. We also evaluate our test both in a Monte Carlo setting, and in an empirical application concerning duration of strikes by manufacturing workers.

Our null  $\mathbb{H}_{10}$  is weaker than additive separability but stronger than monotonicity. Lu and White (2014) and Su et al. (2014) propose tests for additive separability under the same conditional exogeneity assumption that  $U$  is independent of  $X$  given  $Z$ . Specifically, they test whether there exists an unknown measurable function  $G_1$  such that

$$Y = G_1(X) + U \text{ a.s.}$$

Testing  $\mathbb{H}_{10}$  is more general than testing for separability, since our null is equivalent to additive separability in the special case where  $G$  is known to be the identity function. Hence if we reject our  $\mathbb{H}_{10}$ , then we also reject their additive separability.

Hoderlein et al. (2014, HSWY hereafter) test for monotonicity under a conditional exogeneity assumption. HSWY test whether there exists a function  $R$  such that

$$Y = R(X, U)$$

where  $R$  is strictly monotonic in its second argument. Our null is stronger than monotonicity, so if the HSWY test rejects monotonicity, then our null  $\mathbb{H}_{10}$  is also rejected. Our null  $\mathbb{H}_{10}$  combines monotonicity with the additional restriction that the observable  $X$  and unobservable  $U$  are additively separable under a transformation function  $G$ . Our test exploits this additivity restriction, and so should be generally stronger than HSWY for testing our null  $\mathbb{H}_{10}$ . Also, the HSWY test requires that  $Z$  not be empty, while our test of  $\mathbb{H}_{10}$  can be applied even if we have no conditioning covariates  $Z$ .

The rest of the paper is organized as follows. In Section 2, we propose and motivate our test. In Section 3, we show that our test statistics are asymptotically normal under the null, and we analyze their global and local power. In Section 4, we conduct some Monte Carlo simulations to evaluate the finite sample performance of our test statistics. In Section 5, we provide an empirical application to testing for the specification of GAFT models in data on the durations of strikes. In Section 6, we discuss extensions to other closely related hypotheses. Section 7 concludes. The proofs of the main results in the paper are relegated to Appendices A-C and those for the technical lemmas are given in the supplementary material.

---

<sup>3</sup>Horowitz (1996) considers the estimation of the semiparametric model under our null, where the function  $H_1$  takes a parametric form (unlike our nonparametric case) and without covariates  $Z$ . His estimator also relies on the implication that the ratio of the derivatives is a multiplicative function of  $X$  and  $Y$ .

## 2 A Specification Test for Transformation Models

In this section, we describe implications of  $\mathbb{H}_{10}$  that are used to motivate our test construction, and then describe our proposed test statistic.

### 2.1 Motivation

To construct our test, we first impose a conditional exogeneity assumption. Let  $X \perp U \mid Z$  denote that  $X$  and  $U$  are independent given  $Z$ .

**Assumption A.1.** Let  $Z$  be an observable random vector of dimension  $d_z \in \mathbb{N}$ , such that  $X \perp U \mid Z$  and that  $X$  and  $U$  are not measurable with respect to the sigma-field generated by  $Z$ .

Assumption A.1 is equivalent to the unconfoundedness assumption in the treatment effect literature and is widely used to identify causal effects. For detailed discussions, see Altonji and Matzkin (2005), Hoderlein and Mammen (2007), Imbens and Newey (2009), and White and Lu (2011), among others. It is also closely related to the assumptions used to allow for endogeneity in the control function literature, where  $Z$  would equal the residuals from a regression of  $X$  on exogenous instruments. See, e.g., Heckman and Robb (1986) and Blundell and Powell (2003, 2004).

Let  $F(\cdot \mid x, z) \equiv F_{Y \mid X, Z}(\cdot \mid x, z)$  and  $f(\cdot \mid x, z) \equiv f_{Y \mid X, Z}(\cdot \mid x, z)$  denote the conditional cumulative distribution function (CDF) and probability density function (PDF) of  $Y$  given  $(X, Z) = (x, z)$ , respectively. Let  $V = (X', Z)'$ . Let  $\mathcal{X}$ ,  $\mathcal{Y}$ ,  $\mathcal{V}$ , and  $\mathcal{W}$  denote the supports of  $X$ ,  $Y$ ,  $V$ , and  $W$ , respectively. Note that we allow the support of  $F(\cdot \mid x, z)$  to change according to the values  $x$  and  $z$ . Let  $r^0(y; x, z) \equiv \frac{D_x F(y \mid x, z)}{f(y \mid x, z)}$ , so  $r^0(y; x, z)$  is the ratio of two partial derivatives of  $F(y \mid x, z)$ , since  $f(y \mid x, z) = \partial F(y \mid x, z) / \partial y$  and  $D_x F(y \mid x, z) \equiv \partial F(y \mid x, z) / \partial x$ .

The following theorem characterizes some useful properties of the transformation model under  $\mathbb{H}_{10}$ .

**Theorem 2.1** *Suppose that  $f(y \mid x, z) \neq 0$  for all  $(y, x, z) \in \mathcal{W}$ .*

(a) *If  $\mathbb{H}_{10}$  and A.1 hold and the first order (partial) derivatives of  $G$  and  $H_1$  exist, then there exist two measurable functions  $s_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_x}$  and  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_+$  (or  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_-$ ) such that*

$$r^0(Y; X, Z) = s_1(X) s_2(Y) \text{ a.s.}, \quad (2.1)$$

where  $s_1(x) = -\partial S_1(x) / \partial x$  for some measurable function  $S_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ , and  $1/s_2(y) = \partial S_2(y) / \partial y$  for some measurable function  $S_2 : \mathbb{R} \rightarrow \mathbb{R}$ .

(b) *If there exist two measurable functions  $s_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_x}$  and  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_+$  (or  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_-$ ) such that (2.1) holds,  $s_1(x) = -\partial S_1(x) / \partial x$  for some function  $S_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ , and  $1/s_2(y) = \partial S_2(y) / \partial y$  for some measurable function  $S_2 : \mathbb{R} \rightarrow \mathbb{R}$ , then  $\mathbb{H}_{10}$  holds in the sense that there exist two measurable functions  $G : \mathbb{R} \rightarrow \mathbb{R}$  and  $H_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$  such that*

$$Y = G[H_1(X) + U] \text{ a.s.} \quad (2.2)$$

where  $G$  is strictly monotonic and differentiable, all first order partial derivatives of  $H_1$  exist, and  $U$  is a scalar unobservable random variable satisfying  $X \perp U \mid Z$ .

**Remark 2.1.** Theorem 2.1(a) says that under  $\mathbb{H}_{10}$  and the conditional exogeneity condition in A.1, the ratio  $r^0(y; x, z)$  is free of  $z$  and can be factored into the product of a function  $s_1$  of  $x$  and a function

$s_2$  of  $y$ , where the function  $s_1$  can be written as the derivative of a scalar function, and the function  $s_2$  does not alternate in sign on its support. Theorem 2.1(b) says the converse is also true: as long as the factorization in (2.1) holds with  $s_1$  and  $s_2$  satisfying appropriate conditions, the observables  $(Y, X, Z)$  will satisfy the version of transformation model (2.2) under the null.

**Remark 2.2.** Theorem 2.1 gives a characterization of  $\mathbb{H}_{10}$ , but it does not by itself provide a test for  $\mathbb{H}_{10}$ . The proof of Theorem 2.1(a) shows that  $s_1$  and  $s_2$  in the theorem depend on the unknown functions  $H_1$  and  $G$ , respectively, so we cannot directly test equation (2.1). We instead propose a feasible and straightforward test statistic that is based on implications of the factorization in (2.1).

Let  $\mathcal{Y}_0 \equiv [y, \bar{y}] \subset \mathcal{Y}$  for finite real numbers  $y$  and  $\bar{y}$ . Let  $\mathbf{1}\{\cdot\}$  denote the indicator function that equals one when  $\cdot$  is true and zero otherwise, and let  $E_Y(\cdot)$  and  $E_{XZ}(\cdot)$  denote expectations with respect to  $Y$  and  $(X, Z)$ , respectively. Define

$$\begin{aligned} r(y; x, z) &\equiv r^0(y; x, z) \mathbf{1}\{y \in \mathcal{Y}_0\}, \\ r_0 &\equiv E_Y E_{XZ} [r(Y; X, Z)], \\ r_1(x) &\equiv E[r(Y; x, Z)], \\ r_2(y) &\equiv E[r(y; X, Z)], \end{aligned} \tag{2.3}$$

where  $r(y; x, z)$  denotes a trimmed version of  $r^0(y; x, z)$ . Note that  $r$ ,  $r_0$ ,  $r_1$  and  $r_2$  are all  $d_x \times 1$  vectors. The following corollary summarizes a testable implication of (2.1) under  $\mathbb{H}_{10}$  and A.1.

**Corollary 2.2** *Suppose that  $\mathbb{H}_{10}$  and A.1 hold and  $r_0 \neq 0$ . Then*

$$r(Y; X, Z) \circ r_0 = r_1(X) \circ r_2(Y) \text{ a.s.}, \tag{2.4}$$

where  $\circ$  denotes the Hadamard product.

**Remark 2.3.** This corollary remains valid if we drop the indicator  $\mathbf{1}\{y \in \mathcal{Y}_0\}$  in the definition of  $r$  in (2.3). Equivalently, one can take  $\mathcal{Y}_0 = \mathcal{Y}$  in the definition of  $r$  and still obtain the above result provided that  $r$  is well defined. We incorporate the indicator function in our theorem to permit the trimming of the data in the tails that facilitates the rigorous establishment of the asymptotic properties of our test. Specifically our asymptotic theory below requires consistent estimation of  $r(y; x, z)$  uniformly in  $(y; x, z) \in \mathcal{W}_0 \equiv \mathcal{Y}_0 \times \mathcal{V}$ . If  $f(y | x, z)$  is too close to zero for some values of  $(y, x, z) \in \mathcal{W}$ , then we cannot estimate  $r(y; x, z)$  uniformly in  $(y; x, z) \in \mathcal{W}$  at a sufficiently fast rate. We therefore restrict our attention to a subset  $\mathcal{W}_0$  such that  $f(y | x, z)$  is bounded away from zero on it.

Based on Corollary 2.2, consider the following null hypothesis

$$\mathbb{H}_0 : \Pr[r(Y; X, Z) \circ r_0 - r_1(X) \circ r_2(Y) = 0] = 1. \tag{2.5}$$

The alternative hypothesis  $\mathbb{H}_A$  is the negation of  $\mathbb{H}_0$ , i.e.,

$$\mathbb{H}_A : \Pr[r(Y; X, Z) \circ r_0 - r_1(X) \circ r_2(Y) = 0] < 1. \tag{2.6}$$

According to the characterization result in Theorem 2.1, rejection of (2.5) can only be due to the violation of either  $\mathbb{H}_{10}$ , the original null hypothesis of interest, or the conditional exogeneity condition in A.1. Maintaining the conditional exogeneity assumption, we may therefore use the null hypothesis  $\mathbb{H}_0$  to test the original null of interest,  $\mathbb{H}_{10}$ . Alternatively, if we maintain the transformation model specification in

$\mathbb{H}_{10}$ , our test can be used to test the conditional exogeneity assumption A.1 (see remark 2.6 below for more on this last point).

To test the null hypothesis  $\mathbb{H}_0$  in (2.5), we use a construction analogous to that of Härdle and Mammen (1993) by considering the weighted  $L_2$  distance between  $r \circ r_0$  and  $r_1 \circ r_2$ :

$$\Gamma \equiv E \left[ \|r(Y; X, Z) \circ r_0 - r_1(X) \circ r_2(Y)\|^2 a(Y; X, Z) \right], \quad (2.7)$$

where  $\|\cdot\|$  denotes the Euclidean norm, and  $a(y; x, z)$  is a nonnegative weight function that is continuous on its compact support  $\mathcal{Y}_0 \times \mathcal{V}_0$ , where  $\mathcal{V}_0 \subset \mathcal{V}$ . Then  $\Gamma = 0$  under  $\mathbb{H}_0$  and generally deviates from zero under  $\mathbb{H}_A$ . In the next subsection we consider the sample version of  $\Gamma$  based on local polynomial estimates of  $r$ ,  $r_0$ ,  $r_1$ , and  $r_2$ .

**Remark 2.4.** In a previous version of this paper, we considered another set of testable implications:

$$\Pr [r(Y; X, Z) r_0^\pi - r_1(X) r_2^\pi(Y) = 0] = 1 \text{ under } \mathbb{H}_{10}, \quad (2.8)$$

where  $r_2^\pi(y) \equiv \pi' r_2(y)$ ,  $r_0^\pi \equiv \pi' r_0$ , and  $\pi \equiv (\pi_1, \dots, \pi_{d_x})'$  as a  $d_x \times 1$  weight vector (e.g.,  $\pi = (1/d_x, \dots, 1/d_x)$ ). This exploits the additional implication that  $s_2(\cdot)$  is a scalar function under  $\mathbb{H}_{10}$ , however, constructing a test based on equation (2.8) introduces the problem of choosing a weight vector  $\pi$ . Thus, following the suggestion of a referee, we now construct a test based on the null hypothesis (2.5) above, which is free of  $\pi$ . Note that when  $d_x = 1$ ,  $\pi$  can only equal 1, in which case (2.5) and (2.8) are equivalent.

The null  $\mathbb{H}_0$  only exploits some of the implications of Theorem 2.1(a). This suggests that there might be some additional testable restrictions in  $\mathbb{H}_{10}$  that  $\mathbb{H}_0$  ignores. Corollary 2.3 below shows that when  $d_x = 1$ , the only additional restriction implied by  $\mathbb{H}_{10}$  that  $\mathbb{H}_0$  ignores is a sign restriction on  $s_2(\cdot)$ , which is  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_+$  or  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_-$ . That is, the sign of  $s_2(y)$  does not depend on  $y$ .

**Corollary 2.3** *Suppose that  $d_x = 1$ ,  $f(y | x, z) \neq 0$  for all  $(y, x, z) \in \mathcal{W}$ ,  $\mathcal{Y}_0 = \mathcal{Y}$  and  $r_0 \neq 0$ . Then  $\mathbb{H}_{10}$  holds if and only if*

$$\Pr [r(Y; X, Z) \cdot |r_0| - r_1(X) \cdot |r_2(Y)| = 0] = 1. \quad (2.9)$$

In principle, we could test equation (2.9) directly. However, a test based on (2.9) would involve the non-differentiable absolute value function, which would greatly complicate the asymptotic analysis. When  $d_x > 1$  there are, in addition to the sign restriction of  $s_2(\cdot)$ , two more restrictions in  $\mathbb{H}_{10}$  that  $\mathbb{H}_0$  ignores, as follows. (i) We do not exploit the fact that  $s_2(\cdot)$  is a scalar function under  $\mathbb{H}_{10}$ . This information could be incorporated into our test as discussed in Remark 2.4. (ii) We do not exploit the fact that  $s_1(\cdot)$ , a  $d_x$ -multivariate function, equals a vector of the derivatives of an unknown scalar function, i.e.,  $s_1(\cdot)$  can be written as  $s_1(x) = -\partial S_1(x) / \partial x$  for some measurable function  $S_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ . It might be possible to exploit this restriction in the null hypothesis by imposing the implied constraint that the matrix of derivatives  $\partial s_1(x) / \partial x'$  is symmetric (though this would require an additional smoothness assumption). Overall, these differences between  $\mathbb{H}_{10}$  and  $\mathbb{H}_0$  appear to be relatively minor, so the main substance of the testable implications of  $\mathbb{H}_{10}$  is given by the product form in  $\mathbb{H}_0$  that we test.

**Remark 2.5.**  $\mathbb{H}_0$  is a necessary condition of  $\mathbb{H}_{10}$ . Following a referee's suggestion, we provide a characterization of the intersection of  $\mathbb{H}_0$  and  $\mathbb{H}_{1A}$ . Let  $\mathcal{Y}_0 = \mathcal{Y}$ . When  $d_x = 1$ , the intersection of  $\mathbb{H}_0$  and  $\mathbb{H}_{1A}$  includes all data generating processes (DGPs) that satisfy equation (2.5), but fail to satisfy equation (2.9). Even though we feel that the difference between equations (2.5) and (2.9) is small, the intersection

of  $\mathbb{H}_0$  and  $\mathbb{H}_{1A}$  is not empty. To illustrate, consider the example under  $\mathbb{H}_{1A}$  that  $Y = R(X, U) = X^2U$ , where  $X$  has support on  $\mathbb{R}^+ \setminus \{0\}$  and  $U$  has support  $[a, b]$  where  $a < 0$  and  $b > 0$ .<sup>4</sup> In this example,

$$r(Y; X, Z) = -\frac{2Y}{X}, r_0 = -2E\left(\frac{1}{X}\right)E(Y), r_1(X) = -\frac{2}{X}E(Y), \text{ and } r_2(Y) = -2E\left(\frac{1}{X}\right)Y.$$

It is easy to show that equation (2.5) holds. However,

$$r(Y; X, Z) \cdot |r_0| - r_1(X) \cdot |r_2(Y)| = -4\frac{1}{X} \cdot \left|E\left(\frac{1}{X}\right)\right| \cdot |E(Y)| \cdot \left[Y - \frac{E(Y)}{|E(Y)|}|Y|\right] \neq 0.$$

When  $d_x > 1$ , the intersection of  $\mathbb{H}_0$  and  $\mathbb{H}_{1A}$  includes all DGPs that satisfy equation (2.5), but fail to satisfy one (or both) of the following conditions: (i)  $r(Y; X, Z) \cdot |r_0^\pi| - r_1(X) \cdot |r_2^\pi(Y)| = 0$ , where  $r_2^\pi(y) \equiv \pi' r_2(y)$ ,  $r_0^\pi \equiv \pi' r_0$ , and  $\pi \equiv (\pi_1, \dots, \pi_{d_x})'$  as a  $d_x \times 1$  weight vector (e.g.,  $\pi = (1/d_x, \dots, 1/d_x)$ ). (For the role of the weight  $\pi$ , see Remark 2.4 above); and (ii)  $r_1(\cdot)$  can be written as  $r_1(x) = -\partial S_1(x) / \partial x$  for some measurable function  $S_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ .

**Remark 2.6.** Chiappori et al. (2013, CKK hereafter) consider a model that is similar to our null model. Using our notation, their model can be written as  $Y = G(\mathcal{H}_1(X, Z) + U)$ , where  $X$  and  $Z$  are  $d_x \times 1$  and  $d_z \times 1$  vectors of observable variables, respectively,  $U$  is a scalar unobservable error term, and  $G : \mathbb{R} \rightarrow \mathbb{R}$  is an unknown strictly monotone function and  $\mathcal{H}_1 : \mathbb{R}^{d_x+d_z} \rightarrow \mathbb{R}$  is an unknown measurable function. They assume that  $X \perp U \mid Z$  and mainly focus on the identification and estimation of  $G(\cdot)$  and  $\mathcal{H}_1(\cdot)$ .<sup>5</sup> In an intermediate step in the proof of their identification result, CKK show that (again using our notation)  $r^0(Y; X, Z) = \tilde{s}_1(X, Z) \cdot s_2(Y)$  a.s., where  $\tilde{s}_1 : \mathbb{R}^{d_x+d_z} \rightarrow \mathbb{R}$  and  $s_2 : \mathbb{R} \rightarrow \mathbb{R}$  are two measurable functions. This is similar to our characterization of  $r^0$ , though CKK use their result in a completely different way. In particular, CKK use their result to show that  $\theta(y; x, z)$  is a constant function of  $(x, z)$ , where  $\theta(y; x, z) = S(y; x, z) / E[S(Y; x, z)]$  and  $S(y; x, z) = \int_0^y r^0(u; x, z) du$ . Equivalently,

$$\int \theta(y; x, z) w(x, z) d(x, z) - \theta(y; x, z) = 0 \quad \forall (y, x, z), \quad (2.10)$$

where  $w(\cdot)$  is a weight function such that  $\int w(x, z) d(x, z) = 1$ . CKK then propose a weighted  $L_2$ -distance-based test of the conditional exogeneity condition (i.e., that  $X \perp U \mid Z$ ) by testing (2.10).

We test implications of  $Y = G[H_1(X) + U]$  against the general alternative given either  $X \perp U \mid Z$  or just  $X \perp U$ . In contrast, CKK assume  $G(\mathcal{H}_1(X, Z) + U)$  holds and test if  $X \perp U \mid Z$ . Thus both our null and our alternative differ from CKK's test, and so we view our test as a complement to theirs, not a substitute. However, one point of commonality is that, if one maintains the assumption that  $X \perp U \mid Z$ , then either CKK's test or our test could be applied to test our null  $\mathbb{H}_{10}$ . But even when maintaining this conditional independence assumption, there are still substantial differences between the tests for considering the null  $\mathbb{H}_{10}$ . For example, unlike CKK, we exploit the exclusion restriction that  $Z$  does not enter the structural equation  $Y = G[H_1(X) + U]$  under  $\mathbb{H}_{10}$ , and so we have power in those directions.<sup>6</sup>

<sup>4</sup>If  $U$  has support on  $\mathbb{R}^+ \setminus \{0\}$ , then we can write  $Y = \exp[\ln(X^2) + \ln(U)]$ , which is essentially a DGP under the null  $\mathbb{H}_{10}$  by setting  $\tilde{U} = \ln(U)$ .

<sup>5</sup>To identify  $\mathcal{H}_1(\cdot)$ , and hence for estimation, CKK also assume that they observe an additional instrumental variable  $Q$  such that  $E(U \mid Q) = 0$  and that the conditional distribution of  $Z$  given  $Q$  is complete.

<sup>6</sup>It is not surprising that our test would have advantages over CKK in testing  $\mathbb{H}_{10}$ , because that was not CKK's intended null. Other differences are that our test is based directly on Corollary 2.2 while CKK's is based on equation (2.10), we use local polynomials instead of Nadaraya-Watson kernel estimators, and our test is numerically simpler by not requiring any numerical integration, while CKK requires multiple numerical integration steps.

## 2.2 Estimation and test statistic

The derivation in the previous section allows the covariates  $Z$  to be continuous or discrete. To describe our estimators and associated test statistics, we first consider the (more difficult) case where  $Z$  is continuous. Remark 2.7 below then discusses the case where some or all of the elements of  $Z$  are discrete.

We employ local polynomial regression to estimate various unknown population objects. Let  $v \equiv (x', z')' = (v_1, \dots, v_d)'$  be a  $d \times 1$  vector,  $d \equiv d_x + d_z$ , where  $x$  is  $d_x \times 1$  and  $z$  is  $d_z \times 1$ . Let  $\mathbf{j} \equiv (j_1, \dots, j_d)$  be a  $d$ -vector of non-negative integers. Following Masry (1996), adopt the notation

$$v^{\mathbf{j}} \equiv \prod_{i=1}^d v_i^{j_i}, \quad \mathbf{j}! \equiv \prod_{i=1}^d j_i!, \quad |\mathbf{j}| \equiv \sum_{i=1}^d j_i, \quad \sum_{0 \leq |\mathbf{j}| \leq p} \equiv \sum_{k=0}^p \sum_{j_1=0}^k \cdots \sum_{\substack{j_d=0 \\ j_1+\dots+j_d=k}}^k.$$

From  $v^{\mathbf{j}} \equiv \prod_{i=1}^d v_i^{j_i}$ , the  $j_i$ 's represent powers applied to the elements of  $v$  when constructing polynomials.

Consider the  $p$ -th order local polynomial estimators  $D_x \hat{F}_b(y|x, z)$  of  $D_x F(y|x, z)$ . The subscript  $b = b_n$  is a bandwidth parameter. Let  $V_i \equiv (X_i', Z_i')'$  so  $V_i - v = ((X_i - x)', (Z_i - z)')'$ . Given observations  $\{(Y_i, V_i), i = 1, \dots, n\}$ , we estimate  $D_x F(y|v)$  by solving the weighted least squares problem

$$\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^n \left[ \mathbf{1}\{Y_i \leq y\} - \sum_{0 \leq |\mathbf{j}| \leq p} \beta'_{\mathbf{j}} ((V_i - v)/b)^{\mathbf{j}} \right]^2 K_b(V_i - v). \quad (2.11)$$

Here  $\boldsymbol{\beta}$  stacks the  $\beta_{\mathbf{j}}$ 's ( $0 \leq |\mathbf{j}| \leq p$ ) in lexicographic order (with  $\beta_{\mathbf{0}}$ , indexed by  $\mathbf{0} \equiv (0, \dots, 0)$ , in the first position, the element with index  $(0, 0, \dots, 1)$  next, etc.) and  $K_b(\cdot) \equiv K(\cdot/b)/b^d$ , where  $K(\cdot)$  is a symmetric PDF on  $\mathbb{R}^d$ . Let  $\hat{\boldsymbol{\beta}}(y|v)$  denote the solution to the above minimization problem.

Let  $N_l \equiv (l + d - 1)!/(l!(d - 1)!)$  be the number of distinct  $d$ -tuples  $\mathbf{j}$  having  $|\mathbf{j}| = l$ . In the above estimation problem, this denotes the number of distinct  $l$ th order partial derivatives of  $F(y|v)$  with respect to  $v$ . Let  $N \equiv \sum_{l=0}^p N_l$ . Let  $\mu(\cdot)$  be a stacking function such that  $\mu((V_i - v)/b)$  denotes an  $N \times 1$  vector that stacks  $((V_i - v)/b)^{\mathbf{j}}$ ,  $0 \leq |\mathbf{j}| \leq p$ , in lexicographic order (e.g.,  $\mu(v) = (1, v)'$  when  $p = 1$ ). Let  $\mu_b(v) \equiv \mu(v/b)$ . Then

$$\hat{\boldsymbol{\beta}}(y|v) = [\mathbf{S}_b(v)]^{-1} n^{-1} \sum_{i=1}^n K_b(V_i - v) \mu_b(V_i - v) \mathbf{1}\{Y_i \leq y\}, \quad (2.12)$$

where  $\mathbf{S}_b(v) \equiv n^{-1} \sum_{i=1}^n K_b(V_i - v) \mu_b(V_i - v) \mu_b(V_i - v)'$ . The  $p$ -th order local polynomial estimator  $D_x \hat{F}_b(y|x, z)$  of  $D_x F(y|x, z)$  is given by

$$D_x \hat{F}_b(y|x, z) = e_1 \hat{\boldsymbol{\beta}}(y|x, z) / b \quad (2.13)$$

where  $e_1 \equiv [0_{d_x \times 1}, I_{d_x}, 0_{d_x \times (N - d_x - 1)}]$  selects the estimator of the coefficient of  $(X_i - x)/b$  in the above regression.

To estimate  $f(y|v)$ , the conditional PDF of  $Y_i$  given  $V_i = v$ , we again employ local polynomial regression. Like Fan et al. (1996), we estimate  $f(y|v)$  as  $\hat{f}_c(y|v)$ , the minimizing constant in the weighted least squares problem

$$\min_{\gamma} n^{-1} \sum_{i=1}^n \left[ L_c(Y_i - y) - \sum_{0 \leq |\mathbf{j}| \leq p} \gamma'_{\mathbf{j}} ((V_i - v)/c)^{\mathbf{j}} \right]^2 K_c(V_i - v),$$

where  $\boldsymbol{\gamma}$  stacks the  $\gamma_{\mathbf{j}}$ 's ( $0 \leq |\mathbf{j}| \leq p$ ) in lexicographic order and  $L_c(\cdot) \equiv L(\cdot/c)/c$ , with  $L(\cdot)$  a symmetric kernel function defined on  $\mathbb{R}$  and  $c \equiv c_n$  a bandwidth parameter. Here, we use the same bandwidth



sequence for  $Y_i$  and  $V_i$ , although different choices of bandwidths are also possible. To reduce the bias of the estimator  $\hat{f}_c$ , we permit use of a higher-order kernel for  $L$ . It is straightforward to verify that

$$\hat{f}_c(y|v) = e_2'[\mathbf{S}_c(v)]^{-1} n^{-1} \sum_{i=1}^n K_c(V_i - v) \mu_c(V_i - v) L_c(Y_i - y), \quad (2.14)$$

where  $e_2 \equiv (1, 0, \dots, 0)'$  is an  $N \times 1$  vector.<sup>7</sup>

Define

$$\begin{aligned} \hat{r}(y; x, z) &\equiv \frac{D_x \hat{F}_b(y|x, z)}{\hat{f}_c(y|x, z)} \mathbf{1}\{y \in \mathcal{Y}_0\}, & \hat{r}_0 &\equiv \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \hat{r}(Y_i; X_j, Z_j), \\ \hat{r}_1(x) &\equiv \frac{1}{n} \sum_{i=1}^n \hat{r}(Y_i; x, Z_i), & \text{and } \hat{r}_2(y) &\equiv \frac{1}{n} \sum_{i=1}^n \hat{r}(y; X_i, Z_i). \end{aligned}$$

Our proposed test statistic is

$$\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^n \|\hat{r}(Y_i; X_i, Z_i) \circ \hat{r}_0 - \hat{r}_1(X_i) \circ \hat{r}_2(Y_i)\|^2 a(Y_i; X_i, Z_i), \quad (2.15)$$

which is a sample analogue of  $\Gamma$  in (2.7). We next study the asymptotic properties of  $\hat{\Gamma}$  under  $\mathbb{H}_0$ ,  $\mathbb{H}_A$ , and a sequence of Pitman local alternatives.

**Remark 2.7.** The above estimators and associated tests are easily extended to allow some or all elements of  $Z$  to be discrete. To estimate  $r(y; x, z)$  in this case, just stratify the sample by each distinct discrete outcome. Specifically, suppose  $Z = (Z_c, Z_d)$ , where  $Z_c$  is continuous and  $Z_d$  discrete. Then estimate  $r(y; x, z) = r(y; x, z_c, z_d)$  as above (replacing  $Z$  with  $Z_c$  everywhere), just using the data having  $Z_{di} = z_d$ , and repeat for each value  $z_d$  in the support of  $Z_d$ . The functions  $r_0$ ,  $r_1$ , and  $r_2$  can be estimated exactly the same way, by averaging out  $(X_i, Y_i, Z_i)$ ,  $(Y_i, Z_i)$ , and  $(X_i, Z_i)$ , respectively, and then our test statistic  $\hat{\Gamma}$  is still given by (2.15). More sophisticated estimators (e.g., smoothing across the discrete  $Z_d$  cells as proposed in Li and Racine, 2003) could also be used to estimate  $r$  these functions. We omit the details for brevity.

## 3 Asymptotic Properties of the Test Statistic

### 3.1 Basic assumptions

To study asymptotic properties of  $\hat{\Gamma}$ , make the following assumptions.

**Assumption C.1.** Let  $W_i \equiv (Y_i, X_i', Z_i)'$ ,  $i = 1, 2, \dots, n$ , be IID random variables on  $(\Omega, \mathcal{F}, P)$ , with  $(Y_i, X_i, Z_i)$  distributed identically to  $(Y, X, Z)$ .

**Assumption C.2.** (i) The PDF  $f(v)$  of  $V_i$  is continuous in  $v \in \mathcal{V}$ , and  $f(y|v)$  is continuous in  $(y, v) \in \mathcal{Y}_0 \times \mathcal{V}$ .

(ii) There exist  $C_1, C_2 \in (0, \infty)$  such that  $C_1 \leq \inf_{v \in \mathcal{V}} f(v) \leq \sup_{v \in \mathcal{V}} f(v) \leq C_2$ , and  $C_1 \leq \inf_{(y,v) \in \mathcal{Y}_0 \times \mathcal{V}} f(y|v) \leq \sup_{(y,v) \in \mathcal{Y}_0 \times \mathcal{V}} f(y|v) \leq C_2$ .

---

<sup>7</sup>To ensure that the estimator of  $f(y|v)$  is positive, we can replace  $\hat{f}_c(y|v)$  here with a trimmed version defined as  $\max\{\hat{f}_c(y|v), \epsilon/\sqrt{n}\}$ , where  $\epsilon$  is a small positive number, e.g.  $\epsilon = 0.01$ . This would not change our resulting asymptotic theory.

**Assumption C.3.** (i)  $F(\cdot|v)$  is equicontinuous on  $\mathcal{Y}_0$ :  $\forall \epsilon > 0, \exists \delta > 0 : |y - \tilde{y}| < \delta \Rightarrow \sup_{y \in \mathcal{Y}_0} |F(y|v) - F(\tilde{y}|v)| < \epsilon$ . For each  $y \in \mathcal{Y}_0$ ,  $F(y|\cdot)$  is Lipschitz continuous on  $\mathcal{V}$  and has all partial derivatives up to order  $p+1$ ,  $p \in \mathbb{N}$ .

(ii) Let  $D^{\mathbf{j}}F(y|v) \equiv \partial^{|\mathbf{j}|}F(y|v)/\partial^{j_1}v_1 \dots \partial^{j_d}v_d$ . For each  $y \in \mathcal{Y}_0$ ,  $D^{\mathbf{j}}F(y|\cdot)$  with  $|\mathbf{j}| = p+1$  is uniformly bounded and Lipschitz continuous on  $\mathcal{V}$ : for all  $v, \tilde{v} \in \mathcal{V}$ ,  $|D^{\mathbf{j}}F(y|v) - D^{\mathbf{j}}F(y|\tilde{v})| \leq C_3 \|v - \tilde{v}\|$  for some  $C_3 \in (0, \infty)$  where  $\|\cdot\|$  is the Euclidean norm.

(iii) For each  $v \in \mathcal{V}$  and for all  $y, \tilde{y} \in \mathcal{Y}_0$ ,  $|D^{\mathbf{j}}F(y|v) - D^{\mathbf{j}}F(\tilde{y}|v)| \leq C_4 |y - \tilde{y}|$  for some  $C_4 \in (0, \infty)$  where  $|\mathbf{j}| = p+1$ .

**Assumption C.4.** Let  $r \geq 2$ . The  $r$ th derivative  $f^{(r)}(y|v)$  of  $f(y|v)$  with respect to  $y$  and all the  $(p+1)$ th partial derivatives of  $f(y|v)$  with respect to  $v$  are uniformly continuous on  $\mathcal{Y}_0 \times \mathcal{V}$ .

**Assumption C.5.** (i) The kernel  $K : \mathbb{R}^d \rightarrow \mathbb{R}^+$  is a continuous, bounded, and symmetric PDF.

(ii)  $v \rightarrow \|v\|^{2p+1} K(v)$  is integrable on  $\mathbb{R}^d$  with respect to the Lebesgue measure.

(iii) Let  $\mathbf{K}_{\mathbf{j}}(u) \equiv v^{\mathbf{j}}K(v)$  for all  $\mathbf{j}$  with  $0 \leq |\mathbf{j}| \leq 2p+1$ . For some finite constants  $\sigma_K, \bar{\sigma}_1$ , and  $\bar{\sigma}_2$ , either  $K(\cdot)$  is compactly supported such that  $K(v) = 0$  for  $\|u\| > \sigma_K$ , and  $|\mathbf{K}_{\mathbf{j}}(v) - \mathbf{K}_{\mathbf{j}}(\tilde{v})| \leq \bar{\sigma}_2 \|v - \tilde{v}\|$  for any  $v, \tilde{v} \in \mathbb{R}^d$  and for all  $\mathbf{j}$  with  $0 \leq |\mathbf{j}| \leq 2p+1$ ; or  $K(\cdot)$  is differentiable,  $\|\partial \mathbf{K}_{\mathbf{j}}(v)/\partial v\| \leq \bar{\sigma}_1$ , and for some  $\iota_0 > 1$ ,  $|\partial \mathbf{K}_{\mathbf{j}}(v)/\partial v| \leq \bar{\sigma}_1 \|v\|^{-\iota_0}$  for all  $\|v\| > \sigma_K$  and for all  $\mathbf{j}$  with  $0 \leq |\mathbf{j}| \leq 2p+1$ .

**Assumption C.6.** The univariate kernel function  $L$  satisfies  $\int L(y)^2 dy < \infty$  and is a symmetric  $r$ th order kernel, i.e.,  $\int L(y) dy = 1$ ,  $\int y^s L(y) dy = 0$  for all  $s = 1, \dots, r-1$ , and  $\int y^r L(y) dy < \infty$ . The  $r$ th derivative of  $L$  exists and is continuous.

**Assumption C.7.** (i)  $p > d/2$ .

(ii) As  $n \rightarrow \infty$ ,  $(c^{p+1} + c^r)/b^{d/2} \rightarrow 0$ ,  $(b^p + c^{p+1} + c^r)b^{d/2+2}/c^{d+1} \rightarrow 0$ ,  $b^{d+4}/c^{d+1} \rightarrow 0$ ,  $nb^{2(p+1)+d} \rightarrow 0$ , and  $nb^{d+2}(c^{2(p+1)} + c^{2r}) \rightarrow 0$ .

(iii) As  $n \rightarrow \infty$ ,  $\min\{nb^{2d}, nb^{3d/2+1}/\ln n, nb^{d+2}, nb^{d_x+2}/\ln n, nb^{d+1}c^{(d+1)/2}/\ln n, nb^{d/2}c^{d+1}/\ln n, nb^{-(d/2+2)}c^{2(d+1)}/\ln n, nb^{-1}c^{3(d+1)/2}/\ln n, nb^{-(d+4)}c^{3(d+1)}\} \rightarrow \infty$ .

We assume IID observations in Assumption C.1, which is standard in cross-section studies. Assumptions C.2-C.4 impose smoothness conditions on the conditional CDF  $F(y|v)$  and PDF  $f(y|v)$  that are used to ensure uniform consistency of our local polynomial estimators, based on results of Masry (1996) and Hansen (2008). Assumption C.2(ii) is relatively strong, but it is needed given the definitions of  $r(y; x, z)$ ,  $r_0$ ,  $r_1(x)$  and  $r_2(y)$  and the uniform consistency of  $\hat{r}(y; x, z)$  (uniformly in  $(y; (x, z)) \in \mathcal{Y}_0 \times \mathcal{V}$ ) to establish the asymptotic properties of the estimators  $\hat{r}_0$ ,  $\hat{r}_1(x)$  and  $\hat{r}_2(y)$ ; see the proofs of Lemmas B.3(ii)-(iii) in the supplemental material. However, if we modify the definition of  $r(y; x, z)$  to be

$$r(y; x, z) = r^0(y; x, z) \mathbf{1}\{(y; (x, z)) \in \mathcal{Y}_0 \times \mathcal{V}_0\}$$

and define  $r_0$ ,  $r_1(x)$  and  $r_2(y)$  accordingly, then we can relax Assumption C.2(ii) to:

**Assumption C.2.** (ii\*) There exist  $C_1, C_2 \in (0, \infty)$  such that  $C_1 \leq \inf_{v \in \mathcal{V}_0} f(v) \leq \sup_{v \in \mathcal{V}_0} f(v) \leq C_2$ , and  $C_1 \leq \inf_{(y,v) \in \mathcal{Y}_0 \times \mathcal{V}_0} f(y|v) \leq \sup_{(y,v) \in \mathcal{Y}_0 \times \mathcal{V}_0} f(y|v) \leq C_2$ .

In this case, we would only need to restrict  $f(v)$  and  $f(y|v)$  to be well behaved on  $\mathcal{V}_0$  and  $\mathcal{Y}_0 \times \mathcal{V}_0$ , respectively.<sup>8</sup> The asymptotic theory developed in the next two subsections would remain essentially the same.

<sup>8</sup>We would like to thank an associate editor for pointing out that Assumption C.2 (ii) could be relaxed in this way.

Assumptions C.5 and C.6 impose conditions on the kernels  $K$  and  $L$ , which are standard in the literature for local polynomial regression or conditional density estimation. Assumption C.7 restricts the choice of bandwidth sequences  $b$  and  $c$ , the order  $p$  of local polynomial regressions, and the order  $r$  of the kernel  $L$ . This assumption allows  $c$  to differ from  $b$ , but in the case where  $b = c$  Assumption C.7 simplifies to the following assumption:

**Assumption C.7\***. (i)  $p > d/2$  and  $r > d/2$ .

(ii) As  $n \rightarrow \infty$ ,  $nb^{2(p+1)+d} \rightarrow 0$  and  $nb^{2r+d+2} \rightarrow 0$ .

(iii) As  $n \rightarrow \infty$ ,  $\min\{nb^{2d}, nb^{3(d+1)/2}/\ln n, nb^{d+2}, nb^{d_x+2}/\ln n\} \rightarrow \infty$ .

Note that we allow  $d_z = 0$ , otherwise the condition  $nb^{d_x+2}/\ln n \rightarrow \infty$  as  $n \rightarrow \infty$  becomes redundant.

### 3.2 Asymptotic null distribution

In this section, we study the asymptotic behavior of the test statistic in (2.15). To state the next result, let  $w_i \equiv (y_i, v_i)'$  and introduce the following notation:

$$\begin{aligned}\zeta_{1k}(y; v) &\equiv b^{-1}f(y|v)^{-1}e_1\bar{\mathbf{S}}_b(v)^{-1}\mu_b(V_k - v)K_b(V_k - v)\bar{\mathbf{I}}_y(W_k)\mathbf{1}\{y \in \mathcal{Y}_0\}, \\ \zeta_{2k}(y; v) &\equiv f(y|v)^{-2}D_x F(y|v)e_2'\bar{\mathbf{S}}_c(v)^{-1}\mu_c(V_k - v)K_c(V_k - v)\bar{\mathbf{L}}_y(W_k)\mathbf{1}\{y \in \mathcal{Y}_0\}, \\ \zeta_k(y; v) &\equiv \zeta_{1k}(y; v) - \zeta_{2k}(y; v), \\ \zeta(W_i, W_j, W_k) &\equiv (\zeta_j(Y_i; V_i) \circ r_0)'(\zeta_k(Y_i; V_i) \circ r_0)a_i, \\ \varphi(w_i, w_j) &\equiv E[\zeta(W_1, w_i, w_j)],\end{aligned}$$

where  $\bar{\mathbf{S}}_b(v) \equiv E[\mathbf{S}_b(v)]$ ,  $\bar{\mathbf{I}}_y(W_k) \equiv \mathbf{1}\{Y_k \leq y\} - F(y|V_k)$ ,  $\bar{\mathbf{L}}_y(W_k) \equiv L_c(Y_k - y) - \alpha(y|V_k)$ ,  $\alpha(y|v) \equiv E[L_c(Y_k - y)|V_k = v]$  and  $a_i \equiv a(Y_i; X_i, Z_i)$ . Define the asymptotic bias term

$$\begin{aligned}\mathbb{B}_n &\equiv n^{-1}b^{\frac{d}{2}+2}\sum_{i=1}^n\varphi(W_i, W_i) + n^{-4}b^{\frac{d}{2}+2}\sum_{i=1}^n\left\|\sum_{j=1}^n\sum_{k=1}^n\zeta_k(Y_j; X_i, Z_j) \circ r_2(Y_i)\right\|^2 a_i \\ &\quad - 2n^{-3}b^{\frac{d}{2}+2}\sum_{i=1}^n\left(\sum_{l=1}^n\zeta_l(Y_i; V_i) \circ r_0\right)'\left(\sum_{j=1}^n\sum_{k=1}^n\zeta_k(Y_j; X_i, Z_j) \circ r_2(Y_i)\right) a_i \\ &\equiv \mathbb{B}_{1n} + \mathbb{B}_{2n} - 2\mathbb{B}_{3n}, \text{ say.}\end{aligned}$$

We establish the asymptotic null distribution of the  $\hat{\Gamma}$  test statistic as follows:

**Theorem 3.1** *Suppose Assumptions A.1 and C.1-C.7 hold. Then*

$$nb^{\frac{d}{2}+2}\hat{\Gamma} - \mathbb{B}_n \xrightarrow{d} N(0, \sigma_0^2),$$

where  $\sigma_0^2 \equiv \lim_{n \rightarrow \infty} \sigma_n^2$  and  $\sigma_n^2 = 2b^{d+4}E[\varphi(W_1, W_2)^2]$ .

**Remark 3.1.** The asymptotic bias  $\mathbb{B}_n$  of  $nb^{\frac{d}{2}+2}\hat{\Gamma}$  contains three terms  $\mathbb{B}_{1n}$ ,  $\mathbb{B}_{2n}$ , and  $-2\mathbb{B}_{3n}$ . The first two terms reflect the contributions of  $\hat{r}(Y_i; V_i) \circ \hat{r}_0$  and  $\hat{r}_1(X_i) \circ \hat{r}_2(Y_i)$ , respectively, and the last term reflects the interaction between these latter two terms. We show that  $\mathbb{B}_{1n} = O_P(b^{\frac{d}{2}+2}(b^{-d-2} + c^{-d-1}))$  in Lemma B.4,  $\mathbb{B}_{2n} = O_P(b^{\frac{d}{2}+2}(b^{-d_x-2} + c^{-d_x}))$  in Lemma B.5(b), and  $\mathbb{B}_{3n} = O_P(b^{\frac{d}{2}+2}(b^{-d_x-2} + c^{-d_x}))$  in Lemma B.6(b). Here  $\mathbb{B}_{1n}$  never vanishes asymptotically whereas  $\mathbb{B}_{2n}$  and  $\mathbb{B}_{3n}$  are asymptotically negligible under appropriate conditions, such as when  $b = c$  and  $d_z > d_x$ . The asymptotic variance  $\sigma_n^2$  of  $nb^{\frac{d}{2}+2}\hat{\Gamma}$

only reflects the contribution of  $\hat{r}(Y_i; V_i) \circ \hat{r}_0$ , due to the faster convergence rate of  $\hat{r}_1(X_i) \circ \hat{r}_2(Y_i)$  to  $r_1(X_i) \circ r_2(Y_i)$  than that of  $\hat{r}(Y_i; V_i) \circ \hat{r}_0$  to  $r(Y_i; V_i) \circ r_0$ .

To implement the test, we need consistent estimates of the asymptotic bias and variance. Let

$$\begin{aligned}\hat{\zeta}_{1k}(y; v) &\equiv b^{-1} \hat{f}_c(y|v)^{-1} e_1 \mathbf{S}_b(v)^{-1} \mu_b(V_k - v) K_b(V_k - v) \hat{\mathbf{1}}_y(W_k) \mathbf{1}\{y \in \mathcal{Y}_0\}, \\ \hat{\zeta}_{2k}(y; v) &\equiv \hat{f}_c(y|v)^{-2} D_x \hat{F}_b(y|v) e_2 \mathbf{S}_c(v)^{-1} \mu_c(V_k - v) K_c(V_k - v) \hat{\mathbf{L}}_y(W_k) \mathbf{1}\{y \in \mathcal{Y}_0\}, \\ \hat{\zeta}_k(y; v) &\equiv \hat{\zeta}_{1k}(y; v) - \hat{\zeta}_{2k}(y; v), \\ \hat{\varphi}(W_j, W_k) &\equiv n^{-1} \sum_{i=1}^n \left( \hat{\zeta}_j(Y_i; V_i) \circ \hat{r}_0 \right)' \left( \hat{\zeta}_k(Y_i; V_i) \circ \hat{r}_0 \right) a_i,\end{aligned}$$

where  $\hat{\mathbf{1}}_y(W_k) \equiv \mathbf{1}\{Y_k \leq y\} - \hat{F}_b(y|V_k)$ ,  $\hat{\mathbf{L}}_y(W_k) \equiv L_c(Y_k - y) - \hat{f}_c(y|V_k)$ , and  $\hat{F}_b(y|V_k)$  is the  $p$ th order local polynomial estimator of  $F(y|V_k)$  by using the kernel  $K$  and bandwidth  $b$ . We propose to estimate the asymptotic bias  $\mathbb{B}_n$  and variance  $\sigma_n^2$  respectively by

$$\begin{aligned}\hat{\mathbb{B}}_n &\equiv n^{-1} b^{\frac{d}{2}+2} \sum_{i=1}^n \hat{\varphi}(W_i, W_i) + n^{-4} b^{\frac{d}{2}+2} \sum_{i=1}^n \left\| \sum_{j=1}^n \sum_{k=1}^n \hat{\zeta}_k(Y_j; X_i, Z_j) \circ \hat{r}_2(Y_i) \right\|^2 a_i \\ &\quad - 2n^{-3} b^{\frac{d}{2}+2} \sum_{i=1}^n \left( \sum_{l=1}^n \hat{\zeta}_l(Y_i; V_i) \circ \hat{r}_0 \right)' \left( \sum_{j=1}^n \sum_{k=1}^n \hat{\zeta}_k(Y_j; X_i, Z_j) \circ \hat{r}_2(Y_i) \right) a_i, \\ \hat{\sigma}_n^2 &\equiv 2n^{-2} b^{d+4} \sum_{i=1}^n \sum_{j=1}^n \hat{\varphi}(W_i, W_j)^2.\end{aligned}$$

It is straightforward to show  $\hat{\mathbb{B}}_n - \mathbb{B}_n = o_P(1)$  and  $\hat{\sigma}_n^2 - \sigma_n^2 = o_P(1)$ . We can now compare

$$T_n \equiv \left( nb^{\frac{d}{2}+2} \hat{\Gamma} - \hat{\mathbb{B}}_n \right) / \sqrt{\hat{\sigma}_n^2} \quad (3.1)$$

to the critical value  $z_\alpha$  defined as the upper  $\alpha$  percentile from the  $N(0, 1)$  distribution (since the test is one-sided) and reject the null when  $T_n > z_\alpha$ .

### 3.3 Consistency and asymptotic local power

The following theorem shows that the test  $T_n$  is consistent for the class of global alternatives

$$\mathbb{H}_A : \mu_A \equiv E \left\{ \|r(Y; X, Z) \circ r_0 - r_1(X) \circ r_2(Y)\|^2 a(Y; X, Z) \right\} > 0.$$

**Theorem 3.2** *Suppose Assumptions C.1-C.7 hold. Then  $T_n$  diverges to infinity at the rate of  $nb^{\frac{d}{2}+2}$  under  $\mathbb{H}_A$ , i.e.,  $P(T_n > e_n) \rightarrow 1$  as  $n \rightarrow \infty$  under  $\mathbb{H}_A$  for any nonstochastic sequence  $e_n = o(nb^{\frac{d}{2}+2})$ .*

To study the local power of our test, we consider the model

$$Y_{ni} \equiv R_n(X_{ni}, U_{ni}) = G[H_1(X_{ni}) + U_{ni} + \gamma_n \delta(X_{ni}, U_{ni})], \quad (3.2)$$

where  $\gamma_n \rightarrow 0$  as  $n \rightarrow \infty$ ,  $\delta(X_{ni}, U_{ni})$  is not additively separable in  $X_{ni}$  and  $U_{ni}$ , and  $G$  and  $H_1$  are as defined under  $\mathbb{H}_{10}$  in Section 1. Note that we allow both  $Y_{ni}$  and  $(X_{ni}, U_{ni})$  to be double-array and the structural function  $R_n$  is now  $n$ -dependent. As before, we assume that  $X_{ni}$  and  $U_{ni}$  are independent given the  $d_z$ -vector  $Z_{ni} : X_{ni} \perp U_{ni} \mid Z_{ni}$ . Following the literature on nonseparable models, we also assume that

$R_n(x, \cdot)$  is strictly increasing for each  $x$  on the support of  $X_{ni}$ . Formally, we put these requirements into the following assumption.

**Assumption A.1\***  $X_{ni} \perp U_{ni} \mid Z_{ni}$  and equation (3.2) holds such that  $R_n(x, \cdot)$  is strictly increasing for each  $x$  on the support of  $X_{ni}$ .

Let  $\delta_n(x, u) \equiv u + \gamma_n \delta(x, u)$ . Given the strict monotonicity of  $G$ , without loss of generality, we assume  $G$  is also strictly increasing. This, in conjunction with Assumption A.1\* implies that  $\delta_n(x, \cdot)$  is strictly increasing for each  $x$  on the support of  $X_{ni}$ .

Let  $F_n(\cdot \mid x, z)$  and  $f_n(\cdot \mid x, z)$  denote the conditional CDF and PDF of  $Y_{ni}$  given  $(X_{ni}, Z_{ni}) = (x, z)$ , respectively. We now use  $\mathcal{X}_n$ ,  $\mathcal{Z}_n$ , and  $\mathcal{W}_n$  to denote the supports of  $X_{ni}$ ,  $Z_{ni}$ , and  $(Y_{ni}, X_{ni}, Z_{ni})$ , respectively. Let  $r_n^0(y; x, z) \equiv \frac{D_x F_n(y \mid x, z)}{f_n(y \mid x, z)}$ . The following theorem parallels Theorem 2.1(a) and lays down the foundation for the asymptotic local analysis of our test statistic.

**Theorem 3.3** *Suppose that A.1\* holds. Suppose that  $f_n(y \mid x, z)$  is continuously differentiable with respect to both  $y$  and  $x$  for each  $z \in \mathcal{Z}_n$  and  $f_n(y \mid x, z) \neq 0$  for all  $(y, x, z) \in \mathcal{W}_n$ . Suppose that  $G : \mathbb{R} \rightarrow \mathbb{R}$  is strictly increasing and continuously differentiable,  $H_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$  is continuously differentiable, and  $\delta(x, \cdot)$  is continuously differentiable for each  $x \in \mathcal{X}_n$ . Then*

$$r_n^0(y; x, z) = s_{1n}(x) s_2(y) + \gamma_n \Delta_n(y; x, z) + o(\gamma_n) \text{ for all } (y, x, z) \in \mathcal{W}_n \quad (3.3)$$

where  $s_{1n} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_x}$  and  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_+$  (or  $\mathbb{R}_-$ ) are defined in (B.6) and  $\Delta_n(y; x, z)$  is defined in (B.5).

As in Section 2.1, let  $r_n(y; x, z) \equiv r_n^0(y; x, z) \mathbf{1}\{y \in \mathcal{Y}_0\}$ ,  $r_{0n} \equiv E_{Y_n} E_{X_n Z_n} [r_n(Y_{ni}; X_{ni}, Z_{ni})]$ ,  $r_{1n}(x) \equiv E[r_n(Y_{ni}; x, Z_{ni})]$ , and  $r_{2n}(y) \equiv E[r_n(y; X_{ni}, Z_{ni})]$ . The following corollary indicates that  $r_{1n}(x) \circ r_{2n}(y)$  only deviates from  $r_n(y; x, z) \circ r_{0n}$  locally.

**Corollary 3.4** *Suppose that the conditions in Theorem 3.3 hold. Then*

$$r_n(y; x, z) \circ r_{0n} - r_{1n}(x) \circ r_{2n}(y) = \gamma_n \bar{\Delta}_n(y; x, z) + o(\gamma_n) \text{ for all } (y, x, z) \in \mathcal{W}_{0n},$$

where  $\bar{\Delta}_n(y; x, z)$  is defined in (B.7) in the appendix and  $\mathcal{W}_{0n}$  is defined as  $\mathcal{W}_n$  but with  $y$  restricted on  $\mathcal{Y}_0$ .

**Remark 3.2.** To see  $\bar{\Delta}_n(y; x, z)$  more clearly, we consider a special case of equation (3.2):

$$Y_{ni} \equiv R_n(X_{ni}, U_{ni}) = G[H_1(X_{ni}) + U_{ni} + \gamma_n \cdot \delta_1(X_{ni}) \cdot U_{ni}],$$

where  $\delta_1(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$  is a measurable function. Further, we assume that  $\delta_1(X_{ni}) > 0$  a.s. and  $\mathcal{Y}_0 = \mathcal{Y}$ . Then we can show that in this case,

$$r(y; x, z) = -\frac{H_1'(x)}{\partial[G^{-1}(y)]/\partial y} - \gamma_n \cdot \left[ \delta_1'(x) \frac{G^{-1}(y)}{\partial[G^{-1}(y)]/\partial y} - H_1(x) \delta_1'(x) \frac{1}{\partial[G^{-1}(y)]/\partial y} \right] + o(\gamma_n).$$

Hence,  $\Delta_n(y; x, z)$  in Theorem 3.3 is:

$$\Delta_n(y; x, z) = -\delta_1'(x) \frac{G^{-1}(y)}{\partial[G^{-1}(y)]/\partial y} + H_1(x) \delta_1'(x) \frac{1}{\partial[G^{-1}(y)]/\partial y},$$

and  $\bar{\Delta}_n(y; x, z)$  in Corollary 3.4 becomes:

$$\begin{aligned} \bar{\Delta}_n(y; x, z) &= [E(H_1'(X_{ni})) \circ \delta_1'(x) - H_1'(x) \circ E(\delta_1'(X_{ni}))] \\ &\times \left[ \frac{G^{-1}(y)}{\partial[G^{-1}(y)]/\partial y} \cdot E \left[ \frac{1}{\partial[G^{-1}(Y_{ni})]/\partial y} \right] - E \left[ \frac{G^{-1}(Y_{ni})}{\partial[G^{-1}(Y_{ni})]/\partial y} \right] \cdot \frac{1}{\partial[G^{-1}(y)]/\partial y} \right]. \end{aligned}$$

To see  $\bar{\Delta}_n(y; x, z) \neq 0$  in this special case, let  $G(\cdot)$  be the identity function:  $G(y) = y$  and  $d_x = 1$ . Then

$$\bar{\Delta}_n(y; x, z) = [E(H'_1(X_{ni})) \cdot \delta'_1(x) - H'_1(x) \cdot E(\delta'_1(X_{ni}))] \cdot [y - E(Y_{ni})].$$

$y - E(Y_{ni})$  is not zero when  $Y_{ni}$  is random.  $E(H'_1(X_{ni})) \cdot \delta'_1(x) - H'_1(x) \cdot E(\delta'_1(X_{ni}))$  is also in general not zero unless  $H_1(x) = b + c \cdot \delta_1(x)$  for all  $x$ , where  $b$  and  $c$  are some constants.<sup>9</sup>

With the above corollary, we can study the local power of our test. We consider the following sequence of Pitman local alternatives:

$$\mathbb{H}_A(\gamma_n) : r_n(y; x, z) \circ r_{0n} - r_{1n}(x) \circ r_{2n}(y) = \gamma_n \bar{\Delta}_n(y; x, z) + o(\gamma_n), \quad (3.4)$$

where  $\gamma_n \rightarrow 0$  as  $n \rightarrow \infty$ , and  $\bar{\Delta}_n$  is a nonzero measurable function with  $\mu_0 \equiv \lim_{n \rightarrow \infty} E[\bar{\Delta}_n(Y_{ni}; X_{ni}, Z_{ni})^2 a(Y_{ni}; X_{ni}, Z_{ni})] < \infty$ . For technical reasons, we assume that the term  $o(\gamma_n)$  in (3.4) holds uniformly in  $(y, x, z)$  on the support of the weight function  $a(y; x, z)$ .

We continue to use  $\hat{r}$ ,  $\hat{r}_0$ ,  $\hat{r}_1$ , and  $\hat{r}_2$  to denote the local polynomial estimates of  $r_n$ ,  $r_{0n}$ ,  $r_{1n}$ , and  $r_{2n}$ , respectively. The asymptotic bias and variance terms are estimated as before with slight notational changes to account for double-array processes. The final test statistic  $T_n$  is constructed as before. The following theorem reports the asymptotic property of  $T_n$  under  $\mathbb{H}_A(\gamma_n)$ .

**Theorem 3.5** *Suppose Assumptions C.1-C.7 hold with the obvious notational changes that allows double-array IID processes. Then under  $\mathbb{H}_A(\gamma_n)$  with  $\gamma_n = n^{-1/2}b^{-d/4-1}$ ,  $T_n \xrightarrow{d} N(\mu_0/\sigma_0, 1)$ .*

**Remark 3.3.** Theorem 3.5 implies that the  $T_n$  test has non-trivial power against Pitman local alternatives that converge to zero at rate  $n^{-1/2}b^{-d/4-1}$ , provided  $0 < \mu_0 < \infty$ . The asymptotic local power function of the test is given by  $1 - \Phi(z_\alpha - \mu_0/\sigma_0)$ , where  $\Phi$  is the standard normal CDF. It is worth mentioning that the local alternative in (3.4) may be motivated from models other than that considered in (3.2). Generally speaking, the local alternative in (3.4) represents a class of local deviations from the implied null hypothesis  $\mathbb{H}_0$  which our test has power to detect. Such a local deviation may be caused by the violation of any conditions specified under  $\mathbb{H}_{01}$ . This includes the model in (3.2) where  $\delta_n(x, u) \equiv u + \gamma_n \delta(x, u)$  may or may not be monotone in its second argument (even though we assume monotonicity to facilitate the derivation), and the case where the conditional exogeneity condition in Assumption A.1\* is locally violated.<sup>10</sup>

**Remark 3.4.** Alternatively, following Remark 2.4, one can exploit the testable implication in (2.8) and construct the test statistic:

$$\hat{\Gamma}(\pi) = \frac{1}{n} \sum_{i=1}^n \|\hat{r}(Y_i; X_i, Z_i) \hat{r}_0^\pi - \hat{r}_1(X_i) \hat{r}_2^\pi(Y_i)\|^2 a(Y_i; X_i, Z_i),$$

---

<sup>9</sup>If  $H_1(X) = b + c \cdot \delta_1(X)$  holds and  $U_{ni} + \frac{c}{\gamma_n} > 0$  a.s., then

$$\begin{aligned} Y_{ni} &\equiv R_n(X_{ni}, U_{ni}) = G[H_1(X_{ni}) + U_{ni} + \gamma_n \cdot \delta_1(X_{ni}) \cdot U_{ni}] \\ &= G[b + c\delta_1(X_{ni}) + U_{ni} + \gamma_n \cdot \delta_1(X_{ni}) \cdot U_{ni}] \\ &= G\left[\exp\left[\ln\left[\gamma_n\left(\delta_1(X_{ni}) + \frac{1}{\gamma_n}\right)\right] + \ln\left[\left(U_{ni} + \frac{c}{\gamma_n}\right)\right]\right] + b - \frac{c}{\gamma_n}\right]. \end{aligned}$$

This is essentially a DGP under the null  $\mathbb{H}_{10}$  by defining  $\tilde{G}_n(\cdot) \equiv G\left(\exp(\cdot) + b - \frac{c}{\gamma_n}\right)$ ,  $\tilde{H}_{1n}(x) \equiv \ln\left[\gamma_n\left(\delta_1(x) + \frac{1}{\gamma_n}\right)\right]$ , and  $\tilde{U}_{ni} \equiv \ln\left[\left(U_{ni} + \frac{c}{\gamma_n}\right)\right]$ . Hence, it is not surprising that  $\bar{\Delta}_n(y; x, z) = 0$  in this case.

<sup>10</sup>This means the conditional PDF  $f_n(\cdot|x)$  of  $Y_{ni}$  given  $X_{ni} = x$  deviates from the conditional PDF  $f_n(\cdot|x, z)$  of  $Y_{ni}$  given  $X_{ni} = x$  and  $Z_{ni} = z$  locally:  $f_n(y|x, z) - f_n(y|x) = \gamma_n \eta(y; x, z) + o(\gamma_n)$ , where  $\gamma_n \rightarrow 0$  as  $n \rightarrow \infty$  and  $\eta(y; x, z)$  is a measurable function of  $(y, x, z)$ .

where  $\hat{r}_0^\pi \equiv \pi' \hat{r}_0$  and  $\hat{r}_2^\pi(y) \equiv \pi' \hat{r}_2(y)$ . In an earlier version of this paper, we showed that a suitable normalization of  $\hat{\Gamma}(\pi)$ , say,  $T_n(\pi)$ , has the standard normal limiting null distribution. This test would require selection of a weigh vector  $\pi$ . For convenience, one could choose equal weights  $\pi = (1/d_x, \dots, 1/d_x)$ . A referee suggested that one might improve power by maximizing over possible weight vectors, e.g., basing a test on the statistic  $\sup_{\pi \in \mathcal{S}} \hat{\Gamma}(\pi)$  where  $\mathcal{S}$  is a unit sphere in  $\mathbb{R}^{d_x}$ . Such a test would likely be considerably more demanding computationally than our proposed test.

**Remark 3.5.** Like many nonparametric specification tests in the literature (e.g., Härdle and Mammen (1993), CKK, and HSWY), our test suffers from a typical curse of dimensionality. Our test can detect local alternatives that converge to the null at the rate of  $\gamma_n = n^{-1/2} b^{-d/4-1}$ , so as  $d$  increases, the local power of our test deteriorates, and does so at rates that are common for nonparametric tests. In our simulations and empirical applications,  $d$  is small, equalling one or two. For estimating the model under the null, to alleviate the curse of dimensionality we could consider semiparametric models of  $H_1(x)$ . Examples of such specifications for  $H_1(x)$  could include partially linear models, single-index models, or additive models. See, e.g., Fan et al. (2001) or Fan and Jiang (2005), among others. However, for the  $L_2$ -distance type test like ours, these semiparametric restrictions do not help to alleviate the curse of dimensionality, as we need to estimate the model under the alternative. Alternatively, to alleviate the curse of dimensionality, one may consider the integrated conditional moment type test (see, e.g., Bierens and Ploberger (1997)).

### 3.4 Simulating the null distribution

It is well known that the asymptotic normal null distributions with estimated variance matrices often do not provide good approximations for kernel-based tests, in part because tests based on normal critical values can be very sensitive to the choice of bandwidth and suffer from substantial finite sample size distortions. We found that to be the case in some experiments with our test (not reported to save space). An alternative, we consider resampling methods to obtain the simulated  $p$ -values or critical values for our test.

As discussed earlier in remark 2.6, CKK propose a test related to ours. They employ a nonparametric bootstrap method to obtain  $p$ -values for their test, but they do not formally demonstrate its asymptotic validity for technical reasons.<sup>1112</sup> In contrast, the asymptotic validity of subsampling can be justified by standard arguments, so we use subsampling instead of a standard bootstrap. Let  $m = m_n$  be a sequence of positive integers such that  $m \rightarrow \infty$  and  $m/n \rightarrow 0$  as  $n \rightarrow \infty$ . Let  $B$  be a large integer. The subsampling (or equivalently, the  $m$ -out-of- $n$  bootstrap) procedure is as follows:

1. Randomly draw  $B$  subsamples  $\left\{ \left( X_i^{*(k)}, Y_i^{*(k)}, Z_i^{*(k)} \right), i = 1, \dots, m \right\}_{k=1}^B$  of size  $m$  from the original sample  $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ .
2. For  $k = 1, \dots, B$ , compute  $T_n$  using the subsample  $\left\{ \left( X_i^{*(k)}, Y_i^{*(k)}, Z_i^{*(k)} \right) \right\}_{i=1}^m$  and denote this as  $\hat{T}_{n,m}^{*(k)}$ .

<sup>11</sup>This bootstrap depends on rates of uniform convergence of kernel estimated objects, and such rates can fail due to issues associated with unbounded support or to boundary biases.

<sup>12</sup>As an associate editor kindly points out, bootstrap tests can have size distortion when the null hypothesis is nonparametric or semiparametric as in our case. For the recent literatures on this, see, e.g., Barrientos-Marin and Sperlich (2010) and Rodriguez-Poo et al. (2014).

3. Calculate the subsampling  $p$ -value as

$$p = B^{-1} \sum_{k=1}^B \mathbf{1} \left\{ T_n < \hat{T}_{n,m}^{*(k)} \right\}.$$

The asymptotic validity of the above subsampling method can be readily established as in Politis et al. (1999). Intuitively, under the null hypothesis both  $T_n$  and  $\hat{T}_{n,m}^{*(k)}$  are asymptotically distributed as  $N(0, 1)$  and thus the test based on the subsampling-based  $p$ -value has the correct asymptotic size, and under the fixed alternative  $T_n$  diverges to infinity at a speed faster than  $\hat{T}_{n,m}^{*(k)}$ , giving the test its power.

## 4 Monte Carlo Simulations

In this section, we use simulations to examine the finite sample performance of our test. We consider four data generating processes (DGPs):

DGPs 1 and 3:  $Y = X + U + \lambda X \sqrt{1 + U^2}$ ;

DGPs 2 and 4:  $Y = \Phi(X + U + \lambda X \sqrt{1 + U^2})$ ;

For each DGP,  $\lambda$  takes four values: 0, 0.25, 0.5, and 1.  $\lambda = 0$  corresponds to the null DGP and  $\lambda = 0.25, 0.5$  and 1 represent gradual departure from the null. In DGPs 1 and 2,  $X \sim \text{Uniform}(-1, 1)$ ,  $U \sim \text{Uniform}(-1, 1)$ , and  $X$  and  $U$  are independent. In DGPs 3 and 4,  $X$  and  $U$  are no longer independent:  $X = 0.5Z + 0.5\varepsilon_1$  and  $U = 0.5Z + 0.5\varepsilon_2$ , where  $\varepsilon_1 \sim \text{Uniform}(-1, 1)$ ,  $\varepsilon_2 \sim \text{Uniform}(-1, 1)$ ,  $Z$  follows a standard normal distribution truncated by  $-2$  and  $2$  in the tails, and  $\varepsilon_1, \varepsilon_2$ , and  $Z$  are mutually independent. By construction,  $X \perp U \mid Z$ .

We use second order (quadratic) local polynomial estimators, i.e.,  $p = 2$ , with a Gaussian PDF for the kernel function. For the bandwidth sequence  $b$  and  $c$ , we use the rule  $\kappa \cdot \text{std}(V) \cdot n^{-\frac{1}{2(p+1)+1}}$  and  $\kappa \cdot \text{std}(Y) \cdot n^{-\frac{1}{2(p+1)+1}}$  associated with  $V$  and  $Y$ , respectively, where  $\kappa$  is a constant and  $\text{std}(V)$  and  $\text{std}(Y)$  are sample standard deviations of  $V$  and  $Y$ , respectively. In general, the optimal  $\kappa$  depends on the underlying specific DGPs. For simplicity we let  $\kappa = 1$  for DGPs 1 and 2 and  $\kappa = 2$  for DGPs 3 and 4. For DGPs 1 and 2, we specify the weight function  $a = 1$ , corresponding to no trimming, whereas for DGPs 3 and 4,  $a$  trims out 2.5% data on each tail of each dimension of  $(Y, X, Z)$ , so

$$a(Y; X, Z) = \mathbf{1}[y_{0.025} \leq Y \leq y_{0.975}] \cdot \mathbf{1}[x_{0.025} \leq X \leq x_{0.975}] \cdot \mathbf{1}[z_{0.025} \leq Z \leq z_{0.975}],$$

where  $y_{0.025}$  and  $y_{0.975}$  are the 0.025 and 0.975 quantiles of  $Y$  respectively, and similarly for  $x_{0.025}, x_{0.975}, z_{0.025}$ , and  $z_{0.975}$ .

We consider the subsampling test with the sample sizes  $n = 200$  and 300. We try three different subsample sizes  $m = \lfloor n^{0.80} \rfloor, \lfloor n^{0.85} \rfloor$ , and  $\lfloor n^{0.90} \rfloor$ , where  $\lfloor \cdot \rfloor$  denotes the integer part of  $\cdot$ . The number of subsamples is  $B = 200$  and the number of replications is 500.

We consider two conventional nominal levels: 0.05 and 0.10. Tables 1-4 present the rejection frequencies for DGPs 1-4, respectively. In each Table,  $\lambda = 0$  corresponds to the null DGP. When the sample size is 200, the subsampling tests are undersized. However, when the sample size increases to 300, the performance improves and the rejection frequencies are closer to their nominal levels. This suggests that a moderate to large sample is required for the test to have good level behavior. This is not surprising, as the estimation of derivatives is much harder and has a slower convergence rate than the estimation of the conditional expectation itself. In general, the tests are less under-sized when the subsample size  $m$  is relatively small.



Table 1: Empirical rejection frequency: DGP 1

$\lambda$	$n$	Subsample size					
		$\lfloor n^{0.80} \rfloor$		$\lfloor n^{0.85} \rfloor$		$\lfloor n^{0.90} \rfloor$	
		0.05	0.10	0.05	0.10	0.05	0.10
0	200	0.006	0.032	0.006	0.038	0.004	0.026
	300	0.034	0.134	0.032	0.134	0.016	0.090
0.25	200	0.124	0.286	0.110	0.258	0.056	0.196
	300	0.428	0.642	0.358	0.653	0.230	0.504
0.5	200	0.460	0.686	0.388	0.622	0.258	0.498
	300	0.860	0.954	0.820	0.926	0.638	0.860
1	200	0.910	0.990	0.864	0.968	0.748	0.908
	300	0.998	1.000	0.996	1.000	0.982	0.996

Table 2: Empirical rejection frequency: DGP 2

$\lambda$	$n$	Subsample size					
		$\lfloor n^{0.80} \rfloor$		$\lfloor n^{0.85} \rfloor$		$\lfloor n^{0.90} \rfloor$	
		0.05	0.10	0.05	0.10	0.05	0.10
0	200	0.010	0.040	0.008	0.040	0.004	0.022
	300	0.036	0.098	0.028	0.092	0.020	0.066
0.25	200	0.186	0.380	0.134	0.344	0.046	0.216
	300	0.412	0.686	0.326	0.592	0.178	0.444
0.5	200	0.700	0.872	0.596	0.836	0.390	0.674
	300	0.938	0.986	0.892	0.970	0.694	0.924
1	200	0.958	0.996	0.878	0.988	0.594	0.908
	300	0.994	1.000	0.980	0.998	0.734	0.972

In each Table,  $\lambda = 0.25, 0.5$  and  $1$  correspond to alternative DGPs and thus they are used to examine the power of the tests. The test has substantial power. For example, at the 5% level, the rejection frequency is 0.910 for DGP 1 when  $\lambda = 1$ , sample size  $n = 200$ , and subsample size  $m = \lfloor n^{0.80} \rfloor$ . For all the values of  $\lambda = 0.25, 0.5$  and  $1$ , the power increases rapidly as the sample size increases. For example, when  $n$  increases from 200 to 300, the rejection frequency increases from 0.460 to 0.860 for DGP 1 with  $\lambda = 0.5$  and  $m = \lfloor n^{0.80} \rfloor$ . As expected, the rejection frequencies increase with the value of  $\lambda$  for all DGPs. The power of the tests increases when the subsample size  $m$  decreases. This is likely because, under the alternative, the test statistics diverge, so the difference between the original test statistics and subsampled test statistics is large when the difference between the original sample size and subsample size is large.

In general, for the same sample size, the rejection frequencies for DGPs 1 and 2 under the alternative are higher than those for DGPs 3 and 4. This suggests that when  $d$  is large, we need to have a relatively large sample to achieve reasonable powers. This reflects the ‘‘curse of dimensionality’’ of our test.

Table 3: Empirical rejection frequency: DGP 3

$\lambda$	$n$	Subsample size					
		$\lfloor n^{0.80} \rfloor$		$\lfloor n^{0.85} \rfloor$		$\lfloor n^{0.90} \rfloor$	
		0.05	0.10	0.05	0.10	0.05	0.10
0	200	0.010	0.040	0.000	0.016	0.002	0.014
	300	0.034	0.106	0.018	0.074	0.004	0.034
0.25	200	0.056	0.162	0.034	0.092	0.016	0.074
	300	0.174	0.370	0.122	0.334	0.034	0.174
0.5	200	0.216	0.482	0.122	0.368	0.072	0.242
	300	0.546	0.754	0.446	0.720	0.220	0.514
1	200	0.806	0.928	0.642	0.864	0.484	0.782
	300	0.972	0.990	0.940	0.986	0.808	0.952

Table 4: Empirical rejection frequency: DGP 4

$\lambda$	$n$	Subsample size					
		$\lfloor n^{0.80} \rfloor$		$\lfloor n^{0.85} \rfloor$		$\lfloor n^{0.90} \rfloor$	
		0.05	0.10	0.05	0.10	0.05	0.10
0	200	0.018	0.054	0.012	0.030	0.002	0.022
	300	0.040	0.096	0.022	0.074	0.004	0.044
0.25	200	0.030	0.096	0.020	0.062	0.000	0.038
	300	0.084	0.158	0.048	0.134	0.012	0.074
0.5	200	0.074	0.182	0.038	0.102	0.010	0.068
	300	0.134	0.292	0.098	0.252	0.042	0.142
1	200	0.250	0.482	0.138	0.344	0.060	0.226
	300	0.448	0.678	0.346	0.612	0.166	0.404

## 5 Empirical Applications

In this section, we consider testing whether duration data obey the class of nonlinear generalized accelerated failure-time (GAFT) models. We then apply our test empirically on a data set of duration of strikes among manufacturing workers in the US.

### 5.1 Testing for GAFT models

Let  $Y$  be the duration of a certain state (a nonnegative random variable) such as duration of a strike. Our test is directly applicable to nonlinear GAFT models, since such models can be written in the form  $Y = G[H_1(X) + U]$ , where  $X$  is a vector of covariates, and  $U$  an unobservable random variable (see, e.g., equation (2.5) in Ridder, 1990).

MPH models are a particularly popular class of GAFT models. Below we provide a direct link between our null hypothesis and MPH models. Let  $h(Y, X, \xi)$  denote the hazard function for  $Y$ . An MPH model of survival time  $Y$  is one where

$$h(Y, X, \xi) = \lambda(Y) \cdot \theta(X) \cdot \xi \tag{5.1}$$

holds for some baseline hazard function  $\lambda(Y)$  and some nonnegative function of covariates  $\theta(X)$ . The MPH model is widely applied in empirical research (for a detailed review, see Van den Berg (2001)). For example, when  $\xi = 1$ , this is the standard proportion hazard (PH) model developed by Cox (1972). A particularly popular parametric specification of the MPH model due to Lancaster (1979) assumes that  $\lambda(Y) = \alpha Y^{\alpha-1}$ ,  $\theta(X) = \exp(X'\beta)$  and  $\xi$  is a gamma distributed random variable. The following Proposition provides a general characterization of MPH models.

**Proposition 5.1** *Suppose that the hazard function of the survival time  $Y$  is  $h(Y, X, \xi)$ , where  $Y \in \mathbb{R}_+$ ,  $X \in \mathbb{R}^{d_x}$ ,  $\xi \in \mathbb{R}_+$  and  $\xi \neq 0$  with probability 1. Let  $\lambda : \mathbb{R} \rightarrow \mathbb{R}_+$  and  $\theta : \mathbb{R}^{d_x} \rightarrow \mathbb{R}_+$  be two measurable functions such that  $\lambda(Y) = 0$  with probability 0 and  $\theta(X) = 0$  with probability 0. Then  $h(Y, X, \xi)$  is a MPH model:*

$$h(Y, X, \xi) = \lambda(Y) \cdot \theta(X) \cdot \xi,$$

if and only if

$$Y = G[H_1(X) + U],$$

where  $G : \mathbb{R} \rightarrow \mathbb{R}_+$  is a strictly increasing function that is differentiable a.e. on its support,  $H_1 : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ , and  $U = \ln\left[\frac{-\ln(1-\varepsilon)}{\xi}\right]$ , where  $\varepsilon$  is a uniform random variable on  $[0, 1]$  and  $\varepsilon \perp (X, \xi)$ .

Proposition 5.1 shows that the MPH model has two important implications: (i) it equals a transformation model of the type given by our null, and (ii)  $U$  allows a distribution determined by  $\ln(-\ln(1-\varepsilon)/\xi)$ . In principle, both restrictions might be testable, though we focus on implication (i), corresponding to our null hypothesis.<sup>13</sup> If our null is rejected, then the specification of MPH models is rejected, so our test can be used as a falsification test for MPH models.

## 5.2 Duration of strikes

In this subsection, we test the specification of GAFT models using data on the duration of strikes. Here  $Y$  is the duration of strikes in U.S. manufacturing firms, defined as the number of days since the start of a strike. Our  $X$  is a scalar variable indicator of the business cycle position of the economy, measured by the deviation of output from its trend. Positive values of  $X$  mean that the economy is above its growth trend. We assume that A.1 holds with  $X \perp U$ , i.e.,  $Z$  is empty.

Our dataset was used in Kennan (1985) and is employed in several econometrics textbooks including as Cameron and Trivedi (2005) and Greene (2011). The sample size is 566. Table 5 presents data summary statistics.

Table 5: Summary statistics for the strike data (sample size  $n = 566$ )

Variable	Name	Mean	Median	Standard Deviation	Min	Max
$Y$	Duration of strikes (days)	43.62	28.00	44.67	1	235
$X$	Business cycle position	0.006	0.008	0.050	-0.140	0.086

<sup>13</sup>More broadly, this proposition shows that nonparametrically the only difference between GAFT and MPH models is some regularity conditions, since if one is given a GAFT model which by Ridder (1990) satisfies  $Y = G[H_1(X) + U]$ , then given the regularity assumed in Proposition 5.1, one can construct an equivalent MPH model by letting  $\xi = [-\ln(1-\varepsilon)]e^{-U}$  where  $\varepsilon$  is uniform.

We apply our subsampling based test. The details of implementation is the same as in the simulations. For the bandwidth sequence  $b$  and  $c$ , we try various values of the constant  $\kappa$ , letting  $\kappa = 0.5, 0.75, 1, 1.25, 1.5$  and  $2$ . Results based on 1000 subsamples are reported in Table 6. Our results are robust, yielding similar  $p$ -values across different bandwidths and subsample sizes. The  $p$ -values are high for all subsample sizes under investigation. This suggests that our test supports the specification of GAFT models.

Table 6:  $p$ -values for the strike data (sample size  $n = 566$ )

Subsample size	$\lfloor n^{0.80} \rfloor = 159$	$\lfloor n^{0.85} \rfloor = 219$	$\lfloor n^{0.90} \rfloor = 300$
$\kappa = 0.5$	0.605	0.497	0.472
$\kappa = 0.75$	0.617	0.522	0.496
$\kappa = 1$	0.651	0.578	0.572
$\kappa = 1.25$	0.599	0.558	0.572
$\kappa = 1.5$	0.540	0.534	0.530
$\kappa = 2$	0.374	0.437	0.445

## 6 Extensions

Our methodology can be extended to test other related hypotheses for specifications in nonseparable models. For example, consider a simple bivariate  $X$  such that  $X \equiv (X_1, X_2)'$ , where  $X_1$  and  $X_2$  are two scalar random variables. Then our results can be used to test the hypotheses:

$\mathbb{H}_{20}$  : There exist two measurable functions  $R_2$  and  $H_3$  such that

$$Y = R_2 [H_3 (X_1, X_2), U] \text{ a.s.}$$

$\mathbb{H}_{2A}$  :  $\mathbb{H}_{20}$  is false;

and

$\mathbb{H}_{30}$  : There exist three measurable functions  $R_3$ ,  $H_4$  and  $H_5$  such that

$$Y = R_3 [H_4 (X_1) + H_5 (X_2), U] \text{ a.s.}$$

$\mathbb{H}_{3A}$  :  $\mathbb{H}_{30}$  is false.

Given the key conditional exogeneity assumption A.1, a testable implication of  $\mathbb{H}_{20}$  is

$$\frac{\partial F_{Y|X_1, X_2, Z}(y | x_1, x_2, z) / \partial x_1}{\partial F_{Y|X_1, X_2, Z}(y | x_1, x_2, z) / \partial x_2} = r_3(x_1, x_2), \quad (6.1)$$

where  $F_{Y|X_1, X_2, Z}(y | x_1, x_2, z)$  is the conditional CDF of  $Y$  given  $(X_1, X_2, Z)$  and  $r_3$  some unknown measurable function. Similarly,  $\mathbb{H}_{30}$  implies that

$$\frac{\partial F_{Y|X_1, X_2, Z}(y | x_1, x_2, z) / \partial x_1}{\partial F_{Y|X_1, X_2, Z}(y | x_1, x_2, z) / \partial x_2} = r_4(x_1) \cdot r_5(x_2) \quad (6.2)$$

for some unknown measurable functions  $r_4$  and  $r_5$ .

Our test can also be extended to test for semiparametric specifications. For example, consider a multi-dimensional  $X$  and one may be interested in testing

$\mathbb{H}_{40}$  : There exist  $\beta \in \mathbb{R}^{d_x}$  and one measurable functions  $R_4$

$$Y = R_4 [X'\beta + U] \text{ a.s., and } R_4 \text{ is strictly monotonic.}$$

$\mathbb{H}_{4A}$  :  $\mathbb{H}_{40}$  is false;

Then  $\mathbb{H}_{40}$  implies that

$$\frac{D_x F(y | x, z)}{f(y | x, z)} = r_6(y) \quad (6.3)$$

for some unknown measurable function  $r_6$ .

To test equations (6.1), (6.2), and (6.3), one can readily construct test statistics similar to ours, using marginal integration as proposed in testing  $\mathbb{H}_0$ .

## 7 Concluding Remarks

In this paper, we proposed a specification test for a transformation model containing a vector of covariates and an unobservable error. This test is related to tests for separability and monotonicity in nonseparable structural equations. We exploit the testable implication of the transformation model that the ratio of the derivatives of a conditional CDF takes a product form. Our test statistics are based on the  $L_2$  distance between restricted and unrestricted estimators of this ratio of derivatives. We show that the test statistics are asymptotically normal and consistent against the alternative of this testable implication. We provide limit normal distribution theory as well as subsampling methods for obtaining  $p$ -values under the null. Our simulations suggest that the test statistics perform well in moderate size samples. We apply our statistic to test the specification of GAFT models for data on the durations of strikes among manufacturers in the US and fail to reject the specification of GAFT models. We find this result to be stable and robust over a wide range of tuning parameter values.

## Appendix

### A Proof of the main results in Section 2

**Proof of Theorem 2.1.** We first prove (a).

$$\begin{aligned} F(y | x, z) &= \Pr[Y \leq y | X = x, Z = z] \\ &= \Pr[G[H_1(X) + U] \leq y | X = x, Z = z] \\ &= \Pr[U \leq G^{-1}(y) - H_1(x) | Z = z] \\ &= F_{U|Z}[G^{-1}(y) - H_1(x), z], \end{aligned}$$

where  $F_{U|Z}(\cdot, z)$  denotes the conditional CDF of  $U$  given  $Z = z$ . Let  $F_{1,U|Z}$  be the derivative of  $F_{U|Z}$  with respect to its first argument. Then,

$$\frac{\partial F(y | x, z) / \partial x}{\partial F(y | x, z) / \partial y} = \frac{F_{1,U|Z}[G^{-1}(y) - H_1(x), z] \cdot [-\partial H_1(x) / \partial x]}{F_{1,U|Z}[G^{-1}(y) - H_1(x), z] \cdot [\partial G^{-1}(y) / \partial y]} = \frac{-\partial H_1(x) / \partial x}{\partial G^{-1}(y) / \partial y}.$$

So the functions  $s_1$  and  $s_2$  exist and are given by  $s_1(x) = -C \partial H_1(x) / \partial x$  and  $s_2(y) = \frac{1}{C \partial G^{-1}(y) / \partial y}$ , where  $C \neq 0$  is an arbitrary constant. Clearly,  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_+$  if  $C > 0$  and  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_-$  if  $C < 0$ . The measurable functions  $S_1$  and  $S_2$  are given by  $CH_1$  and  $CG^{-1}$ , respectively.

We now prove (b). Without loss of generality, assume that  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_+$ . We can always find two scalar functions  $S_1$  and  $S_2$  such that  $\partial S_1(x) / \partial x = -s_1(x)$  and  $\partial S_2(y) / \partial y = 1/s_2(y)$ , where  $S_2(\cdot)$  is strictly

increasing. Combining this with the definition of  $r(y; x, z)$  gives

$$\frac{D_x F(y | x, z)}{D_y F(y | x, z)} = s_1(x) s_2(y) = \frac{-\partial S_1(x) / \partial x}{\partial S_2(y) / \partial y} \text{ for all } (x, y, z) \in \mathcal{W}. \quad (\text{A.1})$$

Let  $\tilde{U} \equiv S_2(Y) - S_1(X)$  and  $\tilde{u} \equiv S_2(y) - S_1(x)$ . By the monotonicity of  $S_2$ , we have  $Y = S_2^{-1}[S_1(X) + \tilde{U}]$  and  $y = S_2^{-1}[S_1(x) + \tilde{u}]$ . It follows that

$$\begin{aligned} F_{\tilde{U}|X,Z}(\tilde{u}, x, z) &\equiv P\left(\tilde{U} \leq \tilde{u} | X = x, Z = z\right) = P(S_2(Y) - S_1(X) \leq \tilde{u} | X = x, Z = z) \\ &= P(Y \leq S_2^{-1}(S_1(x) + \tilde{u}) | X = x, Z = z) \\ &= P(Y \leq y | X = x, Z = z) = F(y | x, z). \end{aligned}$$

Then

$$\begin{aligned} \frac{D_x F(y | x, z)}{D_y F(y | x, z)} &= \frac{D_x F_{\tilde{U}|X,Z}(\tilde{u}, x, z)}{D_y F_{\tilde{U}|X,Z}(\tilde{u}, x, z)} = \frac{\partial F_{\tilde{U}|X,Z}(\tilde{u}, x, z) / \partial \tilde{u} \cdot s_1(x) + \partial F_{\tilde{U}|X,Z}(\tilde{u}, x, z) / \partial x}{\partial F_{\tilde{U}|X,Z}(\tilde{u}, x, z) / \partial \tilde{u} \cdot (1/s_2(y))} \\ &= s_1(x) s_2(y) + \frac{\partial F_{\tilde{U}|X,Z}(\tilde{u}, x, z) / \partial x \cdot (1/s_2(y))}{\partial F_{\tilde{U}|X,Z}(\tilde{u}, x, z) / \partial \tilde{u}} \text{ for all } (x, y, z) \in \mathcal{W}. \quad (\text{A.2}) \end{aligned}$$

Comparing (A.1) with (A.2) yields  $\partial F_{\tilde{U}|X,Z}(\tilde{u}, x, z) / \partial x = 0$  for all  $(\tilde{u}, x, z) \in \mathcal{U} \times \mathcal{V}$  where  $\mathcal{U}$  denotes the support of  $\tilde{U}$ . Therefore,  $\tilde{U} \perp X | Z$ . So far, we have shown that

$$Y = S_2^{-1}[S_1(X) + \tilde{U}]$$

where  $S_2^{-1}$  is strictly monotonic and  $X \perp \tilde{U} | Z$ . The conclusion in part (b) follows by setting  $G = S_2^{-1}$ ,  $H_1 = S_1$ , and  $U = \tilde{U}$ . ■

**Proof of Corollary 2.2.** Under  $\mathbb{H}_{10}$  and Assumption A.1, (2.1) in Theorem 2.1(a) holds, implying that

$$\begin{aligned} r_0 &\equiv E_Y E_{XZ} [r(Y; X, Z)] = E[s_1(X)] E[s_2(Y) \mathbf{1}\{Y \in \mathcal{Y}_0\}], \\ r_1(x) &\equiv E[r(Y; x, Z)] = s_1(x) E[s_2(Y) \mathbf{1}\{Y \in \mathcal{Y}_0\}], \\ r_2(y) &\equiv E[r(y; X, Z)] = E[s_1(X)] s_2(y) \mathbf{1}\{y \in \mathcal{Y}_0\}. \end{aligned}$$

It follows that  $r(Y; X, Z) \circ r_0 - r_1(X) \circ r_2(Y) = [s_1(X) s_2(Y) \mathbf{1}\{Y \in \mathcal{Y}_0\}] \circ \{E[s_1(X)] E[s_2(Y) \mathbf{1}\{Y \in \mathcal{Y}_0\}]\} - \{s_1(X) E[s_2(Y) \mathbf{1}\{Y \in \mathcal{Y}_0\}]\} \circ \{E[s_1(X)] s_2(Y) \mathbf{1}\{y \in \mathcal{Y}_0\}\} = 0$ . ■

**Proof of Corollary 2.3** The ‘‘if’’ part follows from Theorem 2.1(b) by letting  $s_1(X) = r_1(X)$  and  $s_2(Y) = |r_2(Y)| / |r_0|$ . The ‘‘only if’’ part is implied by Theorem 2.1(a). Note that when  $s_2(Y) > 0$ ,

$$\begin{aligned} &r(Y; X, Z) \cdot |r_0| - r_1(X) \cdot |r_2(Y)| \\ &= s_1(X) \cdot s_2(Y) \cdot |E[s_1(X)] \cdot E[s_2(Y)]| - s_1(X) \cdot E[s_2(Y)] \cdot |E[s_1(X)] \cdot s_2(Y)| \\ &= s_1(X) \cdot s_2(Y) \cdot |E[s_1(X)]| \cdot |E[s_2(Y)]| - s_1(X) \cdot E[s_2(Y)] \cdot |E[s_1(X)]| \cdot s_2(Y) = 0. \end{aligned}$$

Similarly, when  $s_2(Y) < 0$ ,  $r(Y; X, Z) \cdot |r_0| - r_1(X) \cdot |r_2(Y)| = 0$ . ■

## B Proof of the main results in Section 3

To prove Theorem 3.1, we first establish some technical lemmas. All the proofs of the lemmas can be found in the online supplementary material. Recall that  $V_i \equiv (X'_i, Z'_i)'$ ,  $v \equiv (x', z')'$ ,  $K_b(v) \equiv b^{-d} K(v/b)$ , and

$\mu_b(v) \equiv \mu(v/b)$ . Let  $W_i \equiv (Y_i, V_i)'$  and  $w \equiv (y, v)'$ . Define

$$\mathbf{B}_b(y; v) \equiv \frac{1}{n} \sum_{i=1}^n K_b(V_i - v) \mu_b(V_i - v) \Delta_{i,y}(v) \quad \text{and} \quad \mathbf{V}_b(y; v) \equiv \frac{1}{n} \sum_{i=1}^n K_b(V_i - v) \mu_b(V_i - v) \bar{\mathbf{I}}_y(W_i),$$

where  $\Delta_{i,y}(v) \equiv F(y|V_i) - F(y|v) - \sum_{1 \leq |\mathbf{j}| \leq p} \frac{1}{\mathbf{j}!} D^{\mathbf{j}} F(y|v) (V_i - v)^{\mathbf{j}}$ , and  $\bar{\mathbf{I}}_y(W_i) \equiv \mathbf{1}\{Y_i \leq y\} - F(y|V_i)$ . Let  $\bar{\mathbf{S}}_b(v) \equiv E[\mathbf{S}_b(v)]$  and  $\bar{\mathbf{B}}_b(y; v) \equiv E[\mathbf{B}_b(y; v)]$ , where  $\mathbf{S}_b(v)$  is defined after (2.12). Define

$$\begin{aligned} \mathbf{V}_c^{(L)}(y; v) &\equiv \frac{1}{n} \sum_{i=1}^n K_c(V_i - v) \mu_c(V_i - v) \bar{\mathbf{L}}_y(W_i), \\ \mathbf{B}_c^{(L)}(y; v) &\equiv \frac{1}{n} \sum_{i=1}^n K_c(V_i - v) \mu_c(V_i - v) \left[ \alpha(y|V_i) - f(y|v) - \sum_{1 \leq |\mathbf{j}| \leq p} \frac{1}{\mathbf{j}!} \alpha^{(\mathbf{j})}(y|v) (V_i - v)^{\mathbf{j}} \right], \end{aligned}$$

where  $\bar{\mathbf{L}}_y(W_i) \equiv L_c(Y_i - y) - \alpha(y|V_i)$  and  $\alpha(y|v) \equiv E[L_c(Y_i - y)|V_i = v]$ . Let  $\bar{\mathbf{B}}_c^{(L)}(y; v) \equiv E[\mathbf{B}_c^{(L)}(y; v)]$ .

**Lemma B.1** *Suppose that Assumptions C.1-C.3, C.5, and C.7 hold. Then uniformly in  $(y, v) \in \mathcal{Y}_0 \times \mathcal{V}$ ,*

- (a)  $D_x \hat{F}_b(y|v) - D_x F(y|v) = b^{-1} e_1 \bar{\mathbf{S}}_b(v)^{-1} [\mathbf{V}_b(y; v) + \bar{\mathbf{B}}_b(y; v)] + O_P(\nu_{1b}^2 b^{-1} + \nu_{1b} b^p)$ ,
- (b)  $D_x \hat{F}_b(y|v) - D_x F(y|v) = O_P(\nu_{1b} b^{-1} + b^p)$ ,

where  $\nu_{1b} \equiv n^{-1/2} b^{-d/2} \sqrt{\ln n}$ .

**Lemma B.2** *Suppose that Assumptions C.1-C.7 hold. Then uniformly in  $(y, v) \in \mathcal{Y}_0 \times \mathcal{V}$ ,*

- (a)  $\hat{f}_c(y|v) - f(y|v) = e_2' \bar{\mathbf{S}}_c(v)^{-1} [\mathbf{V}_c^{(L)}(y; v) + \bar{\mathbf{B}}_c^{(L)}(y; v)] + O_P(\nu_{2c}^2 + \nu_{2c} c^{p+1})$ ,
- (b)  $\hat{f}_c(y|v) - f(y|v) = O_P(\nu_{2c} + c^{p+1} + c^r)$ ,

where  $\nu_{2c} \equiv n^{-1/2} c^{-(d+1)/2} \sqrt{\ln n}$ .

**Lemma B.3** *Suppose that Assumptions C.1-C.7 hold. Then*

- (a)  $\hat{r}(y; v) - r(y; v) = b^{-1} e_1 \bar{\mathbf{S}}_b(v)^{-1} \mathbf{V}_b(y; v) f(y|v)^{-1} - D_x F(y|v) e_2' \bar{\mathbf{S}}_c(v)^{-1} \mathbf{V}_c^{(L)}(y; v) f(y|v)^{-2} + O_P(\nu_{bc})$  uniformly in  $(y, v) \in \mathcal{Y}_0 \times \mathcal{V}$ ,
- (b)  $\hat{r}_0 - r_0 = O_P(\nu_{bc} + n^{-1/2} b^{-1})$ ,
- (c)  $\sup_{y \in \mathcal{Y}_0} |\hat{r}_2(y) - r_2(y)| = O_P(\nu_{bc} + n^{-1/2} b^{-1} \sqrt{\ln n})$ ,

where  $\nu_{bc} \equiv \nu_{1b}^2 b^{-1} + b^p + \nu_{2c}^2 + c^{p+1} + c^r + \nu_{1b} \nu_{2c} b^{-1}$ .

**Proof of Theorem 3.1.** Let  $a_i \equiv a(Y_i; X_i, Z_i)$ ,  $r_i \equiv r(Y_i; X_i, Z_i)$ ,  $r_{1i} \equiv r_1(X_i)$ ,  $r_{2i} \equiv r_2(Y_i)$ ,  $\hat{r}_i \equiv \hat{r}(Y_i; X_i, Z_i)$ ,  $\hat{r}_{1i} \equiv \hat{r}_1(X_i)$ , and  $\hat{r}_{2i} \equiv \hat{r}_2(Y_i)$ . Let  $\xi_{1i} = r_i \circ r_0 - r_{1i} \circ r_{2i}$ ,  $\xi_{2i} = (\hat{r}_i - r_i) \circ r_0 + r_i \circ (\hat{r}_0 - r_0) - (\hat{r}_{1i} - r_{1i}) \circ r_{2i} - r_{1i} \circ (\hat{r}_{2i} - r_{2i})$ , and  $\xi_{3i} = (\hat{r}_i - r_i) \circ (\hat{r}_0 - r_0) - (\hat{r}_{1i} - r_{1i}) \circ (\hat{r}_{2i} - r_{2i})$ . Then

$$\begin{aligned} nb^{\frac{d}{2}+2} \hat{\Gamma} &= b^{\frac{d}{2}+2} \sum_{i=1}^n \left\| [(\hat{r}_i - r_i) + r_i] \circ [(\hat{r}_0 - r_0) + r_0] - [(\hat{r}_{1i} - r_{1i}) + r_{1i}] \circ [(\hat{r}_{2i} - r_{2i}) + r_{2i}] \right\|^2 a_i \\ &= b^{\frac{d}{2}+2} \sum_{i=1}^n \|\xi_{1i} + \xi_{2i} + \xi_{3i}\|^2 a_i \\ &= \Gamma_{1n} + \Gamma_{2n} + \Gamma_{3n} + 2\Gamma_{4n} + 2\Gamma_{5n} + 2\Gamma_{6n}, \end{aligned} \tag{B.1}$$

where  $\Gamma_{ln} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \|\xi_{li}\|^2 a_i$  for  $l = 1, 2$ , and  $3$ ,  $\Gamma_{4n} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \xi'_{1i} \xi_{2i} a_i$ ,  $\Gamma_{5n} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \xi'_{1i} \xi_{3i} a_i$ , and  $\Gamma_{6n} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \xi'_{2i} \xi_{3i} a_i$ . Under  $\mathbb{H}_0$ ,  $\Gamma_{jn} = 0$  for  $j = 1, 4, 5$ . It suffices to prove the theorem by showing that (i)  $\Gamma_{2n} - \mathbb{B}_n \xrightarrow{d} N(0, \sigma_0^2)$ , (ii)  $\Gamma_{3n} = o_P(1)$ , and (iii)  $\Gamma_{6n} = o_P(1)$ .

To show (i), we write  $\Gamma_{2n} = \sum_{j=1}^{10} \Gamma_{2n,j}$  where  $\Gamma_{2n,1} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \|(\hat{r}_i - r_i) \circ r_0\|^2 a_i$ ,  $\Gamma_{2n,2} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \|r_i \circ (\hat{r}_0 - r_0)\|^2 a_i$ ,  $\Gamma_{2n,3} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \|(\hat{r}_{1i} - r_{1i}) \circ r_{2i}\|^2 a_i$ ,  $\Gamma_{2n,4} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \|r_{1i} \circ (\hat{r}_{2i} - r_{2i})\|^2 a_i$ ,  $\Gamma_{2n,5}$

$\equiv 2b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_i - r_i) \circ r_0)' (r_i \circ (\hat{r}_0 - r_0)) a_i$ ,  $\Gamma_{2n,6} \equiv -2b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_i - r_i) \circ r_0)' ((\hat{r}_{1i} - r_{1i}) \circ r_{2i}) a_i$ ,  $\Gamma_{2n,7} \equiv -2b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_i - r_i) \circ r_0)' (r_{1i} \circ (\hat{r}_{2i} - r_{2i})) a_i$ ,  $\Gamma_{2n,8} \equiv -2b^{\frac{d}{2}+2} \sum_{i=1}^n (r_i \circ (\hat{r}_0 - r_0))' ((\hat{r}_{1i} - r_{1i}) \circ r_{2i}) a_i$ ,  $\Gamma_{2n,9} \equiv -2b^{\frac{d}{2}+2} \sum_{i=1}^n (r_i \circ (\hat{r}_0 - r_0))' (r_{1i} \circ (\hat{r}_{2i} - r_{2i})) a_i$ , and  $\Gamma_{2n,10} \equiv 2b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_{1i} - r_{1i}) \circ r_{2i})' (r_{1i} \circ (\hat{r}_{2i} - r_{2i})) a_i$ . We show in Lemma B.4 below that  $\Gamma_{2n,1}$  contributes to both the asymptotic bias and variance of our test statistic:  $\Gamma_{2n,1} - \mathbb{B}_{1n} \xrightarrow{d} N(0, \sigma_0^2)$ , and  $\mathbb{B}_{1n} = O_P(b^{\frac{d}{2}+2} (b^{-d-2} + c^{-d-1}))$  is the dominant bias term that never vanishes asymptotically no matter whether  $Z$  is absent or not. The normalization constant  $nb^{\frac{d}{2}+2}$  in the front of  $\hat{\Gamma}$  in (B.1) indicates that the rate of  $\hat{\Gamma}$  converging to zero under the null is uniquely determined by the convergence rate of  $\hat{r}_i$  to  $r_i$  as the asymptotic variance of  $\hat{r}(y; x, z) - r(y; x, z)$  is of the order  $O((n^{-1}b^{-(d+2)}))$ . We show in Lemmas B.5(b) and B.6(c) below that  $\Gamma_{2n,3}$  and  $\Gamma_{2n,6}$  also contribute to the asymptotic bias of our test statistic:  $\Gamma_{2n,3} = \mathbb{B}_{2n} + o_P(1)$  and  $\Gamma_{2n,6} = -2\mathbb{B}_{3n} + o_P(1)$ ,  $\mathbb{B}_{2n} = O_P(b^{(d_z - d_x)/2} + b^{\frac{d}{2}+2} c^{-d_x})$  and  $\mathbb{B}_{3n} = O_P(b^{(d_z - d_x)/2} + b^{\frac{d}{2}+2} c^{-d_x})$ .  $\mathbb{B}_{2n}$  signifies the estimation effect of  $\hat{r}_{1i}$  and  $\mathbb{B}_{3n}$  signifies the estimation effect of both  $\hat{r}_i$  and  $\hat{r}_{1i}$ .  $\mathbb{B}_{2n}$  and  $\mathbb{B}_{3n}$  become  $o_P(1)$  if  $d_z > d_x$  and  $b^{\frac{d}{2}+2} c^{-d_x} = o(1)$ . We also show in the other parts of Lemmas B.5(b) and B.6(c) below that  $\Gamma_{2n,s} = o_P(1)$  for  $s = 2, 4, 5, 7, \dots, 10$ .

In short, by Lemmas B.4, B.5(b) and B.6(c) below,  $\Gamma_{2n,1} + \Gamma_{2n,3} + \Gamma_{2n,6} - \mathbb{B}_n \xrightarrow{d} N(0, \sigma_0^2)$  where  $\mathbb{B}_n \equiv \mathbb{B}_{1n} + \mathbb{B}_{2n} - 2\mathbb{B}_{3n}$ , and by Lemmas B.5(a) and (c) and Lemmas B.6 (a) and (c)-(f),  $\Gamma_{2n,s} = o_P(1)$  for  $s = 2, 4, 5, 7, \dots, 10$ . It follows that  $\Gamma_{2n} - \mathbb{B}_n \xrightarrow{d} N(0, \sigma_0^2)$ .

Next, we show (ii). By Cauchy-Schwarz inequality and the fact that  $\|A \circ B\| \leq \|A\| \|B\|$  for any two conformable vectors  $A$  and  $B$ ,  $\Gamma_{3n} \leq 2\Gamma_{3n,1} + 2\Gamma_{3n,2}$ , where  $\Gamma_{3n,1} \equiv b^{\frac{d}{2}+2} \|\hat{r}_0 - r_0\|^2 \sum_{i=1}^n \|\hat{r}_i - r_i\|^2 a_i$  and  $\Gamma_{3n,2} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \|(\hat{r}_{1i} - r_{1i}) \circ (\hat{r}_{2i} - r_{2i})\|^2 a_i$ . Following the arguments used in the proofs of Lemma B.4 and Lemma B.5(b) respectively, we can readily show that

$$\bar{\Gamma}_{3n,1} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \|\hat{r}_i - r_i\|^2 a_i = O_P\left(b^{\frac{d}{2}+2} (b^{-d-2} + c^{-d-1})\right) \quad \text{and} \quad (\text{B.2})$$

$$\bar{\Gamma}_{3n,2} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n \|\hat{r}_{1i} - r_{1i}\|^2 a_i = O_P(b^{\frac{d}{2}+2} (b^{-d_x-2} + c^{-d_x})). \quad (\text{B.3})$$

Then by Lemma B.3(b),  $\Gamma_{3n,1} = \|\hat{r}_0 - r_0\|^2 \bar{\Gamma}_{3n,1} = O_P(\nu_{bc}^2 + n^{-1}b^{-2}) O_P(b^{\frac{d}{2}+2} (b^{-d-2} + c^{-d-1})) = o_P(1)$  and  $\Gamma_{3n,2} \leq \sup_{y \in \mathcal{Y}_0} \|\hat{r}_2(y) - r_2(y)\|^2 \bar{\Gamma}_{3n,2} = O_P(\nu_{bc}^2 + n^{-1}b^{-2} \ln n) O_P(b^{\frac{d}{2}+2} (b^{-d_x-2} + c^{-d_x})) = o_P(1)$ . Consequently,  $\Gamma_{3n} = o_P(1)$ .

To show (iii), note that  $\Gamma_{6n} = \Gamma_{6n,1} - \Gamma_{6n,2} + \Gamma_{6n,3} - \Gamma_{6n,4} - \Gamma_{6n,5} + \Gamma_{6n,6} - \Gamma_{6n,7} + \Gamma_{6n,8}$ , where  $\Gamma_{6n,1} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [(\hat{r}_i - r_i) \circ r_0]' [(\hat{r}_i - r_i) \circ (\hat{r}_0 - r_0)] a_i$ ,  $\Gamma_{6n,2} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [(\hat{r}_i - r_i) \circ r_0]' [(\hat{r}_{1i} - r_{1i}) \circ (\hat{r}_{2i} - r_{2i})] a_i$ ,  $\Gamma_{6n,3} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [r_i \circ (\hat{r}_0 - r_0)]' [(\hat{r}_i - r_i) \circ (\hat{r}_0 - r_0)] a_i$ ,  $\Gamma_{6n,4} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [r_i \circ (\hat{r}_0 - r_0)]' [(\hat{r}_{1i} - r_{1i}) \circ (\hat{r}_{2i} - r_{2i})] a_i$ ,  $\Gamma_{6n,5} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [(\hat{r}_{1i} - r_{1i}) \circ r_{2i}]' [(\hat{r}_i - r_i) \circ (\hat{r}_0 - r_0)] a_i$ ,  $\Gamma_{6n,6} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [(\hat{r}_{1i} - r_{1i}) \circ r_{2i}]' [(\hat{r}_{1i} - r_{1i}) \circ (\hat{r}_{2i} - r_{2i})] a_i$ ,  $\Gamma_{6n,7} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [r_{1i} \circ (\hat{r}_{2i} - r_{2i})]' [(\hat{r}_i - r_i) \circ (\hat{r}_0 - r_0)] a_i$ , and  $\Gamma_{6n,8} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [r_{1i} \circ (\hat{r}_{2i} - r_{2i})]' [(\hat{r}_{1i} - r_{1i}) \circ (\hat{r}_{2i} - r_{2i})] a_i$ . By Lemma B.3(b) and (B.2),

$$|\Gamma_{6n,1}| \leq \|r_0\| \|\hat{r}_0 - r_0\| \bar{\Gamma}_{3n,1} = O_P\left(\nu_{bc} + n^{-1/2}b^{-1}\right) O_P(b^{\frac{d}{2}+2} (b^{-d-2} + c^{-d-1})) = o_P(1).$$



By Cauchy-Schwarz inequality, Lemma B.3(c), and equations (B.2) and (B.3),

$$\begin{aligned}
|\Gamma_{6n,2}| &\leq \|r_0\| \sup_{y \in \mathcal{Y}_0} \|\hat{r}_2(y) - r_2(y)\| b^{\frac{d}{2}+2} \sum_{i=1}^n \|\hat{r}_i - r_i\| \|\hat{r}_{1i} - r_{1i}\| a_i \\
&\leq \|r_0\| \sup_{y \in \mathcal{Y}_0} \|\hat{r}_2(y) - r_2(y)\| (\bar{\Gamma}_{3n,1} \bar{\Gamma}_{3n,2})^{1/2} \\
&= O_P\left(\nu_{bc} + n^{-1/2} b^{-1} \sqrt{\ln n}\right) \left\{ O_P\left(b^{\frac{d}{2}+2} (b^{-d-2} + c^{-d-1})\right) O_P\left(b^{\frac{d}{2}+2} (b^{-d_x-2} + c^{-d_x})\right) \right\}^{1/2} \\
&= o_P(1).
\end{aligned}$$

Note that  $\Gamma_{6n,3} = b^{\frac{d}{2}+2} ((\hat{r}_0 - r_0) \circ (\hat{r}_0 - r_0))' \bar{\Gamma}_{2n,5}$  where  $\bar{\Gamma}_{2n,5} \equiv \sum_{i=1}^n ((\hat{r}_i - r_i) \circ r_i) a_i$ . By Lemma B.3(b) and the proof of Lemma B.6(a),

$$\Gamma_{6n,3} = b^{\frac{d}{2}+2} O_P\left(\nu_{bc} + n^{-1/2} b^{-1}\right) O_P(b^{-d-1} + n^{1/2} b^{-1} + c^{-d-1}) = o_P(1).$$

Note that  $\Gamma_{6n,4} = (\hat{r}_0 - r_0)' \bar{\Gamma}_{6n,4}$  where  $\bar{\Gamma}_{6n,4} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n (r_i \circ (\hat{r}_{1i} - r_{1i}) \circ (\hat{r}_{2i} - r_{2i})) a_i$ . Following the proof of Lemma B.6(f), we can show that  $\bar{\Gamma}_{6n,4} = o_P(1)$ . This, in conjunction with Lemma B.3(b), implies that  $\Gamma_{6n,4} = o_P(1)$ . Note that  $\Gamma_{6n,5} = (\hat{r}_0 - r_0)' \bar{\Gamma}_{2n,6}$  where  $\bar{\Gamma}_{2n,6} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_i - r_i) \circ (\hat{r}_{1i} - r_{1i}) \circ r_{2i}) a_i$ . By Lemma B.3 and the proof of Lemma B.6(b),  $\Gamma_{6n,5} = O_P(\nu_{bc} + n^{-1/2} b^{-1}) (O_P(b^{(d_z - d_x)/2}) + o_P(1)) = o_P(1)$ . Note that  $|\Gamma_{6n,6}| \leq \sup_{y \in \mathcal{Y}_0} \|\hat{r}_2(y) - r_2(y)\| \bar{\Gamma}_{6n,6}$  where  $\bar{\Gamma}_{6n,6} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_{1i} - r_{1i}) \circ r_{2i} \circ (\hat{r}_{1i} - r_{1i})) a_i$ . Analogously to the proof of Lemma B.4, we can show that  $\bar{\Gamma}_{6n,6} = O_P(b^{\frac{d}{2}+2} (b^{-d_x-2} + c^{-d_x}))$ . Combining this with Lemma B.3(c) yields

$$|\Gamma_{6n,6}| = O_P\left(\nu_{bc} + n^{-1/2} b^{-1} \sqrt{\ln n}\right) O_P\left(b^{\frac{d}{2}+2} (b^{-d_x-2} + c^{-d_x})\right) = o_P(1).$$

Observe that  $\Gamma_{6n,7} = (\hat{r}_0 - r_0)' \bar{\Gamma}_{2n,7}$  where  $\bar{\Gamma}_{2n,7} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [(\hat{r}_i - r_i) \circ r_{1i} \circ (\hat{r}_{2i} - r_{2i})] a_i$ . By Lemma B.3(b) and the proof of B.6(c),  $\Gamma_{6n,7} = O_P(\nu_{bc} + n^{-1/2} b^{-1} \sqrt{\ln n}) o_P(1) = o_P(1)$ . Lastly, by Cauchy-Schwarz inequality and the study of  $\Gamma_{2n,4}$  and  $\Gamma_{3n,2}$  above  $|\Gamma_{6n,8}| \leq \{\Gamma_{2n,4} \Gamma_{3n,2}\}^{1/2} = o_P(1)$ . Consequently we have proved  $\Gamma_{6n} = o_P(1)$ . ■

**Remark.** Admittedly, the formulae for the asymptotic bias and variance are quite complicated because we consider the general local polynomial regressions to estimate both  $f(y|x, z)$  and  $D_x F(y|x, z)$  and each of the four estimates  $\hat{r}_i$ ,  $r_0$ ,  $\hat{r}_{1i}$ , and  $\hat{r}_{2i}$  contribute to the asymptotic bias and variance of our test statistic in different manners.

**Lemma B.4** *Suppose Assumptions C.1-C.7 hold. Then  $\Gamma_{2n,1} - \mathbb{B}_{1n} \xrightarrow{d} N(0, \sigma_0^2)$  where  $\mathbb{B}_{1n} = n^{-1} b^{\frac{d}{2}+2} \sum_{i=1}^n \varphi(W_i, W_i) = O_P(b^{\frac{d}{2}+2} (b^{-d-2} + c^{-d-1}))$ .*

**Lemma B.5** *Suppose Assumptions C.1-C.7 hold. Then*

- (a)  $\Gamma_{2n,2} = b^{\frac{d}{2}+2} \sum_{i=1}^n \|r_i \circ (\hat{r}_0 - r_0)\|^2 a_i = o_P(1)$ ,
- (b)  $\Gamma_{2n,3} = b^{\frac{d}{2}+2} \sum_{i=1}^n \|(\hat{r}_{1i} - r_{1i}) \circ r_{2i}\|^2 a_i = \mathbb{B}_{2n} + o_P(1)$ ,
- (c)  $\Gamma_{2n,4} = b^{\frac{d}{2}+2} \sum_{i=1}^n \|r_{1i} \circ (\hat{r}_{2i} - r_{2i})\|^2 a_i = o_P(1)$ ,

where  $\mathbb{B}_{2n} \equiv n^{-4} b^{\frac{d}{2}+2} \sum_{i=1}^n \left\| \sum_{j=1}^n \sum_{k=1}^n \zeta_k(Y_j; X_i, Z_j) \circ r_{2i} \right\|^2 a_i = O_P(b^{\frac{d}{2}+2} (b^{-d_x-2} + c^{-d_x}))$ . If  $d_z > 0$ , then (b) also holds when we replace  $\mathbb{B}_{2n}$  by  $\bar{\mathbb{B}}_{2n} \equiv n^{-4} b^{\frac{d}{2}+2} \sum_{i=1}^n \sum_{j=1}^n \left\| \sum_{k=1}^n \zeta_j(Y_k; X_i, Z_k) \circ r_{2i} \right\|^2 a_i$ .

**Lemma B.6** *Suppose that Assumptions C.1-C.7 hold. Then*

- (a)  $\Gamma_{2n,5} = 2b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_i - r_i) \circ r_0)' (r_i \circ (\hat{r}_0 - r_0)) a_i = o_P(1)$ ,

$$(b) \Gamma_{2n,6} = -2b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_i - r_i) \circ r_0)' ((\hat{r}_{1i} - r_{1i}) \circ r_{2i}) a_i = -2\mathbb{B}_{3n} + o_P(1),$$

$$(c) \Gamma_{2n,7} = -2b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_i - r_i) \circ r_0)' (r_{1i} \circ (\hat{r}_{2i} - r_{2i})) a_i = o_P(1),$$

$$(d) \Gamma_{2n,8} = -2b^{\frac{d}{2}+2} \sum_{i=1}^n (r_i \circ (\hat{r}_0 - r_0))' ((\hat{r}_{1i} - r_{1i}) \circ r_{2i}) a_i = o_P(1),$$

$$(e) \Gamma_{2n,9} = -2b^{\frac{d}{2}+2} \sum_{i=1}^n (r_i \circ (\hat{r}_0 - r_0))' (r_{1i} \circ (\hat{r}_{2i} - r_{2i})) a_i = o_P(1),$$

$$(f) \Gamma_{2n,10} = 2b^{\frac{d}{2}+2} \sum_{i=1}^n ((\hat{r}_{1i} - r_{1i}) \circ r_{2i})' (r_{1i} \circ (\hat{r}_{2i} - r_{2i})) a_i = o_P(1),$$

where  $\mathbb{B}_{3n} \equiv n^{-3} b^{\frac{d}{2}+2} \sum_{i=1}^n \left[ (\sum_{l=1}^n \zeta_l (W_i) \circ r_0)' \left( \sum_{j=1}^n \sum_{k=1}^n \zeta_k (Y_j; X_i, Z_j) \circ r_2 (Y_i) \right) \right] a_i = O_P(b^{(d_z - d_x)/2} + b^{\frac{d}{2}+2} c^{-d_x})$ .

**Proof of Theorem 3.2.** The proof follows closely from that of Theorems 3.1. By (B.1) and the proof of Theorem 3.1. Now  $\hat{\Gamma} = n^{-1} b^{-(\frac{d}{2}+2)} \Gamma_{1n} + n^{-1} b^{-(\frac{d}{2}+2)} \Gamma_{4n} + n^{-1} b^{-(\frac{d}{2}+2)} \Gamma_{5n} + o_P(1)$ . It is easy to show that  $n^{-1} b^{-(\frac{d}{2}+2)} \Gamma_{1n} = n^{-1} \sum_{i=1}^n \|r_i \circ r_0 - r_{1i} \circ r_{2i}\|^2 a_i = \mu_A + o_P(1)$  and  $n^{-1} b^{-(\frac{d}{2}+2)} \Gamma_{sn} = o_P(1)$  under  $\mathbb{H}_A$  for  $s = 4, 5$ . In addition, under  $\mathbb{H}_A$ , we have  $n^{-1} b^{-(\frac{d}{2}+2)} \hat{\mathbb{B}}_n = o_P(1)$  and  $\hat{\sigma}_n^2 \xrightarrow{P} \sigma_A^2$ . It follows that  $n^{-1} b^{-(\frac{d}{2}+2)} T_n = n^{-1} b^{-(\frac{d}{2}+2)} [nb^{(\frac{d}{2}+2)} \hat{\Gamma} - \hat{\mathbb{B}}_n] / \sqrt{\hat{\sigma}_n^2} = \mu_A / \sigma_A + o_P(1)$  and the result follows. ■

**Proof of Theorem 3.3.** By assumption,  $\delta_n(x, \cdot)$  is strictly increasing and continuously differentiable for each  $x \in \mathcal{X}_n$ . Let  $\delta_n^{-1}(x, \cdot)$  denote the inverse function of  $\delta_n(x, \cdot)$ . Let  $\delta_{n,u}(x, \cdot)$  denote the derivative of  $\delta_n(x, u)$  with respect to  $\cdot$ , i.e.,  $\delta_{n,u}(x, u) = 1 + \gamma_n \delta_u(x, u)$  where  $\delta_u(x, u) = \partial \delta(x, u) / \partial u$ . Then by the inverse function theorem

$$\frac{d\delta_n^{-1}(x, t)}{dt} = \frac{1}{\delta_{n,u}(x, \delta_n^{-1}(x, t))} = \frac{1}{1 + \gamma_n \delta_u(x, \delta_n^{-1}(x, t))}.$$

It follows that

$$\delta_n^{-1}(x, t) = \eta_n(x) + \int_0^t \frac{1}{1 + \gamma_n \delta_u(x, \delta_n^{-1}(x, s))} ds$$

for some function  $\eta_n(x)$  that does not depend on  $t$ . Noting that  $\frac{1}{1+a} = 1 - a + o(a)$  when  $a = o(1)$ , we have

$$\begin{aligned} \delta_n^{-1}(x, t) &= \eta_n(x) + \int_0^t [1 - \gamma_n \delta_u(x, \delta_n^{-1}(x, s))] ds + o(\gamma_n t) \\ &= \eta_n(x) + t - \gamma_n \bar{\eta}_n(x, t) + o(\gamma_n t), \end{aligned} \tag{B.4}$$

where  $\bar{\eta}_n(x, t) = \int_0^t \delta_u(x, \delta_n^{-1}(x, s)) ds$ .

Let  $F_{U_n|Z_n}(\cdot, z)$  and  $f_{U_n|Z_n}(\cdot, z)$  denote the conditional CDF and PDF of  $U_{ni}$  given  $Z_{ni} = z$ , respectively. Let  $t(x, y) \equiv G^{-1}(y) - H_1(x)$ . Then by (3.2), (B.4) and the strict monotonicity of  $G$  and  $\delta_n$ ,

$$\begin{aligned} F_n(y | x, z) &\equiv \Pr[Y_{ni} \leq y | X_{ni} = x, Z_{ni} = z] \\ &= \Pr[G[H_1(X_{ni}) + U_{ni} + \gamma_n \delta(X_{ni}, U_{ni})] \leq y | X_{ni} = x, Z_{ni} = z] \\ &= \Pr[\delta_n(x, U_{ni}) \leq t(x, y) | Z_{ni} = z] \\ &= \Pr[U_{ni} \leq \delta_n^{-1}(x, t(x, y)) | Z_{ni} = z] \\ &= F_{U_n|Z_n}[\delta_n^{-1}(x, t(x, y)), z]. \end{aligned}$$

It follows that

$$\begin{aligned}
r_n^0(y; x, z) &\equiv \frac{\partial F_n(y | x, z) / \partial x}{\partial F_n(y | x, z) / \partial y} = \frac{\partial [\delta_n^{-1}(x, t(x, y))] / \partial x}{\partial [\delta_n^{-1}(x, t(x, y))] / \partial y} \\
&= \frac{\partial [\eta_n(x) + t(x, y) - \gamma_n \bar{\eta}_n(x, t(x, y)) + o(\gamma_n t(x, y))] / \partial x}{\partial [\eta_n(x) + t(x, y) - \gamma_n \bar{\eta}_n(x, t(x, y)) + o(\gamma_n t(x, y))] / \partial y} \\
&\simeq \frac{\partial (\eta_n(x) + t(x, y)) / \partial x - \gamma_n [\bar{\eta}_{n,1}(x, t(x, y)) + \bar{\eta}_{n,2}(x, t(x, y)) \cdot \partial t(x, y) / \partial x]}{\partial t(x, y) / \partial y - \gamma_n \bar{\eta}_{n,2}(x, t(x, y)) \cdot \partial t(x, y) / \partial y} \\
&\simeq \frac{\partial [\eta_n(x) - H_1(x)] / \partial x}{\partial G^{-1}(y) / \partial y} + \gamma_n \Delta_n(y; x, z),
\end{aligned}$$

where  $\simeq$  is used to indicate that we have suppressed high-order remainder terms,  $\bar{\eta}_{n,1}$  and  $\bar{\eta}_{n,2}$  are the partial derivative of  $\bar{\eta}_n$  with respect to its first and second arguments, respectively, and

$$\Delta_n(y; x, z) = \frac{1}{\partial G^{-1}(y) / \partial y} [\eta_n'(x) \bar{\eta}_{n,2}(x, t(x, y)) - \bar{\eta}_{n,1}(x, t(x, y))]. \quad (\text{B.5})$$

So the functions  $s_{1n}$  and  $s_2$  exist and are given by

$$s_{1n}(x) = \frac{C \partial [\eta_n(x) - H_1(x)]}{\partial x} \quad \text{and} \quad s_2(y) = \frac{1}{C \partial G^{-1}(y) / \partial y}, \quad (\text{B.6})$$

where  $C \neq 0$  is an arbitrary constant. Clearly,  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_+$  if  $C > 0$  and  $s_2 : \mathbb{R} \rightarrow \mathbb{R}_-$  if  $C < 0$ . The measurable functions  $S_{1n}$  and  $S_2$  are given by  $C [\eta_n(x) - H_1(x)]$  and  $CG^{-1}$ , respectively. ■

**Proof of Corollary 3.4.** Let  $\mathbf{1}_{ni} = \mathbf{1}\{Y_{ni} \in \mathcal{Y}_0\}$  and  $\mathbf{1}_y = \mathbf{1}\{y \in \mathcal{Y}_0\}$ . By Theorem 3.3,

$$\begin{aligned}
r_{0n} &= E[s_{1n}(X_{ni})] E[s_2(Y_{ni}) \mathbf{1}_{ni}] + \gamma_n E_{Y_n} E_{X_n Z_n} [\Delta_n(Y_{ni}; X_{ni}, Z_{ni}) \mathbf{1}_{ni}] + o(\gamma_n), \\
r_{1n}(x) &= s_{1n}(x) E[s_2(Y_{ni}) \mathbf{1}_{ni}] + \gamma_n E[\Delta_n(Y_{ni}; x, Z_{ni}) \mathbf{1}_{ni}] + o(\gamma_n), \\
r_{2n}(y) &= E[s_{1n}(X_{ni})] s_2(y) \mathbf{1}_y + \gamma_n E[\Delta_n(y; X_{ni}, Z_{ni}) \mathbf{1}_y] + o(\gamma_n)
\end{aligned}$$

It follows that  $r_n(y; x, z) \circ r_{0n} - r_{1n}(x) \circ r_{2n}(y) = \gamma_n \bar{\Delta}_n(y; x, z) + o(\gamma_n)$  for all  $(y, x, z) \in \mathcal{W}_{0n}$ , where

$$\begin{aligned}
\bar{\Delta}_n(y; x, z) &= \{\Delta_n(y; x, z) \mathbf{1}_y\} \circ \{E[s_{1n}(X_{ni})] E[s_2(Y_{ni}) \mathbf{1}_{ni}]\} \\
&\quad + \{s_{1n}(x) s_2(y) \mathbf{1}_y\} \circ \{E_{Y_n} E_{X_n Z_n} [\Delta_n(Y_{ni}; X_{ni}, Z_{ni}) \mathbf{1}_{ni}]\} \\
&\quad - \{s_{1n}(x) E[s_2(Y_{ni}) \mathbf{1}_{ni}]\} \circ \{E[\Delta_n(y; X_{ni}, Z_{ni}) \mathbf{1}_y]\} \\
&\quad - \{E[\Delta_n(Y_{ni}; x, Z_{ni}) \mathbf{1}_{ni}] \circ E[s_{1n}(X_{ni})] s_2(y)\} \{\mathbf{1}_y\}. \quad \blacksquare \quad (\text{B.7})
\end{aligned}$$

**Proof of Theorem 3.5.** The proof follows closely from that of Theorem 3.1, now keeping the additional terms that do not vanish under  $\mathbb{H}_A(\gamma_n)$  with  $\gamma_n = n^{-1/2} b^{-\frac{d}{4}-1}$ . Let  $r_{ni} \equiv r_n(Y_{ni}; X_{ni}, Z_{ni})$ ,  $r_{1ni} \equiv r_{1n}(X_{ni})$ ,  $r_{2ni} \equiv r_{2n}(Y_{ni})$ ,  $\hat{r}_i \equiv \hat{r}(Y_{ni}; X_{ni}, Z_{ni})$ ,  $\hat{r}_{1i} \equiv \hat{r}_1(X_{ni})$ , and  $\hat{r}_{2i} \equiv \hat{r}_2(Y_{ni})$ . Noting that  $\mathbb{B}_n = \mathbb{B}_n + o_P(1)$  and  $\hat{\sigma}_n^2 = \sigma_0^2 + o_P(1)$  under  $\mathbb{H}_A(\gamma_n)$ , it suffices to show that under  $\mathbb{H}_A(\gamma_n)$ , (i)  $\Gamma_{1n} \xrightarrow{P} \mu_0$ , (ii)  $\Gamma_{4n} = o_p(1)$  and (iii)  $\Gamma_{5n} = o_p(1)$ , where  $\Gamma_{1n}$ ,  $\Gamma_{4n}$ , and  $\Gamma_{5n}$  are defined after (B.1) with  $r_0$ ,  $r_i$ ,  $r_{1i}$ , and  $r_{2i}$  now replaced by  $r_{n0}$ ,  $r_{ni}$ ,  $r_{1ni}$ , and  $r_{2ni}$ , respectively. The results in previous lemmas continue to hold when the single-index IID observations  $(Y_i, X_i, Z_i)$  are replaced by the double-array IID observations  $(Y_{ni}, X_{ni}, Z_{ni})$ . Let  $V_{ni} \equiv (X'_{ni}, Z'_{ni})'$  and  $f_{ni} \equiv f_n(Y_{ni} | V_{ni})$ .

(i) holds under  $\mathbb{H}_A(\gamma_n)$  because  $\Gamma_{1n} = b^{\frac{d}{2}+2} \sum_{i=1}^n \|r_{ni} \circ r_{n0} - r_{1ni} \circ r_{2ni}\|^2 a_i = n^{-1} \sum_{i=1}^n \|\bar{\Delta}_n(Y_{ni}; X_{ni}, Z_{ni})\|^2 a_i = \mu_0 + o_P(1)$  by the law of large numbers. For (ii), we decompose  $\Gamma_{4n}$  as  $\Gamma_{4n} = \Gamma_{4n,1} + \Gamma_{4n,2} -$

$\Gamma_{4n,3} - \Gamma_{4n,4}$ , where

$$\begin{aligned}\Gamma_{4n,1} &\equiv b^{\frac{d}{2}+2} \sum_{i=1}^n (r_{ni} \circ r_{0n} - r_{1ni} \circ r_{2ni})' ((\hat{r}_i - r_{ni}) \circ r_{0n}) a_i, \\ \Gamma_{4n,2} &\equiv b^{\frac{d}{2}+2} \sum_{i=1}^n (r_{ni} \circ r_{0n} - r_{1ni} \circ r_{2ni})' (r_{ni} \circ (\hat{r}_0 - r_{0n})) a_i, \\ \Gamma_{4n,3} &\equiv b^{\frac{d}{2}+2} \sum_{i=1}^n (r_{ni} \circ r_{0n} - r_{1ni} \circ r_{2ni})' ((\hat{r}_{1i} - r_{1ni}) \circ r_{2ni}) a_i, \\ \Gamma_{4n,4} &\equiv b^{\frac{d}{2}+2} \sum_{i=1}^n (r_{ni} \circ r_{0n} - r_{1ni} \circ r_{2ni})' (r_{1ni} \circ (\hat{r}_{2i} - r_{2ni})) a_i.\end{aligned}$$

It suffices to prove  $\Gamma_{4n,s} = o_P(1)$  for  $s = 1, 2, 3, 4$ . We only prove  $\Gamma_{4n,1} = o_P(1)$  as the other cases are similar. Let  $\bar{\Delta}_{ni} \equiv \bar{\Delta}_n(Y_{ni}; X_{ni}, Z_{ni})$ . Under  $\mathbb{H}_A(\gamma_n)$  we apply Lemma B.3(a) to obtain

$$\Gamma_{4n,1} = n^{-\frac{1}{2}} b^{\frac{d}{4}+1} \sum_{i=1}^n \delta'_{ni} ((\hat{r}_i - r_{ni}) \circ r_{0n}) a_i = \bar{\Gamma}_{4n,1} + n^{\frac{1}{2}} b^{\frac{d}{4}+1} O_P(\nu_{bc}) = \bar{\Gamma}_{4n,1} + o_P(1),$$

where  $\bar{\Gamma}_{4n,1} \equiv n^{-\frac{1}{2}} b^{\frac{d}{4}+1} \sum_{i=1}^n \bar{\Delta}'_{ni} \{ [b^{-1} e_1 \bar{\mathbf{S}}_b(V_{ni})^{-1} \mathbf{V}_b(Y_{ni}; X_{ni}) f_{ni}^{-1} - D_{xi} e_2' \bar{\mathbf{S}}_c(V_{ni})^{-1} \mathbf{V}_c^{(L)}(V_{ni}) f_{ni}^{-2}] \circ r_{0n} \} a_i$ . Write  $\bar{\Gamma}_{4n,1} = n^{-\frac{3}{2}} b^{\frac{d}{4}+1} \sum_{i=1}^n \sum_{j=1}^n \bar{\Delta}'_{ni}(\zeta_j(Y_{ni}; V_{ni}) \circ r_{0n}) a_i$ . Then  $E \|\bar{\Gamma}_{4n,1}\|^2 = O((b^{\frac{d}{2}} + n^{-1} b^{\frac{d}{2}+2} (b^{-d-2} + c^{-d-1}) + n^{-2} b^{\frac{d}{2}+2} (b^{-2d-2} + c^{-2d-1})) = o(1)$ , implying that  $\bar{\Gamma}_{4n,1} = o_P(1)$ . It follows that  $\Gamma_{4n,1} = o_P(1)$ .

We now show (iii). Decompose  $\Gamma_{5n} = \Gamma_{5n,1} - \Gamma_{5n,2}$  where  $\Gamma_{5n,1} = (\hat{r}_0 - r_0)' \bar{\Gamma}_{5n,1}$ ,  $\Gamma_{5n,2} = \gamma_n b^{\frac{d}{2}+2} \sum_{i=1}^n \bar{\Delta}'_{ni} [(\hat{r}_{1i} - r_{1ni}) \circ (\hat{r}_{2i} - r_{2ni})] a_i$ , and  $\bar{\Gamma}_{5n,1} \equiv b^{\frac{d}{2}+2} \sum_{i=1}^n [r_{ni} \circ r_{0n} - r_{1ni} \circ r_{2ni}] \circ (\hat{r}_i - r_{ni}) a_i$ . Analogous to the study of  $\Gamma_{4n,1}$ , one can readily show that  $\bar{\Gamma}_{5n,1} = o_P(1)$ . Then by Lemma B.3(b),  $\Gamma_{5n,1} = O_P(\nu_{bc} + n^{-1/2} b^{-1}) o_P(1) = o_P(1)$ . Analogous to the proof of Lemma B.6(f), we can show that  $\Gamma_{5n,2} = o_P(\gamma_n)$ . Thus  $\Gamma_{5n} = o_P(1)$ .

Consequently,  $P(T_n \geq z | \mathbb{H}_A(n^{-1/2} b^{-\frac{d}{4}-1})) \rightarrow 1 - \Phi(z - \mu_0 / \sigma_0)$ . This concludes the proof of the theorem.  $\blacksquare$

## C Proof of the main results in Section 5

**Proof of Proposition 5.1.** We first prove the ‘‘if’’ part. By the definition of the hazard function, for any values  $(y, x, \varsigma)$  on the support of  $(Y, X, \xi)$ ,  $h(y, x, \varsigma) = \frac{f(y|x, \varsigma)}{1 - F(y|x, \varsigma)}$ , where  $f(y|x, \varsigma)$  and  $F(y|x, \varsigma)$  are conditional PDF and CDF of  $Y$  given  $(X, \xi) = (x, \varsigma)$ , respectively. Then,

$$\begin{aligned}F(y|x, \varsigma) &= P \left[ G \left( H_1(X) + \ln \left( \frac{-\ln(1-\varepsilon)}{\xi} \right) \right) \leq y \mid X = x, \xi = \varsigma \right] \\ &= P \left[ \frac{-\ln(1-\varepsilon)}{\xi} \leq \exp [G^{-1}(y) - H_1(X)] \mid X = x, \xi = \varsigma \right] \\ &= P [\varepsilon \leq 1 - \exp \{ -\varsigma \exp [G^{-1}(y) - H_1(x)] \}] \\ &= 1 - \exp \{ -\varsigma \exp [G^{-1}(y) - H_1(x)] \}.\end{aligned}$$

Thus  $f(y|x, \varsigma) = \varsigma \exp \{ -\varsigma \exp [G^{-1}(y) - H_1(x)] \} \exp [G^{-1}(y) - H_1(x)] \frac{d[G^{-1}(y)]}{dy}$  and

$$h(y, x, \varsigma) = \frac{f(y|x, \varsigma)}{1 - F(y|x, \varsigma)} = \left\{ \frac{\exp [G^{-1}(y)]}{g(G^{-1}(y))} \right\} \exp [-H_1(x)] \varsigma = \lambda(y) \cdot \theta(x) \cdot \varsigma,$$

where  $\lambda(y) = \exp[G^{-1}(y)]/g(G^{-1}(y))$ ,  $g(s) = dG(s)/ds$ , and  $\theta(x) = \exp[-H_1(x)]$ . This holds for all  $(y, x, \varsigma)$  on the support of  $(Y, X, \xi)$ , thus the “if” part is proved.

Next, we prove the “only if” part. Define the integrated hazard function  $H(Y, X, \xi) = \int_0^Y h(y, X, \xi) dy$ . Then

$$H(Y, X, \xi) = \int_0^Y h(y, X, \xi) dy = \int_0^Y \lambda(y) dy \cdot \theta(X) \cdot \xi = \Lambda(Y) \cdot \theta(X) \cdot \xi,$$

where  $\Lambda(Y) = \int_0^Y \lambda(y) dy$ . Let  $F(Y | X, \xi)$  be the conditional CDF of  $Y$  given  $X$  and  $\xi$ . For any distribution function  $F$ , the integrated hazard function is related to its distribution function by  $H(Y, X, \xi) = -\ln(1 - F(Y | X, \xi))$ . Therefore  $\Lambda(Y)\theta(X)\xi = -\ln(1 - F(Y | X, \xi))$ . Define the random variable  $\varepsilon = F(Y | X, \xi)$ . By construction  $\varepsilon$  is uniformly distributed on  $[0, 1]$  and  $\varepsilon \perp (X, \xi)$  and  $\Lambda(Y)\theta(X)\xi = -\ln(1 - \varepsilon)$ . Thus  $\ln[\Lambda(Y)] = \ln\left[\frac{-\ln(1-\varepsilon)}{\xi}\right] + \ln\left[\frac{1}{\theta(X)}\right]$ . That is,  $Y = G[H_1(X) + U]$  where  $G(\cdot)$  is the inverse function of  $\ln[\Lambda(\cdot)]$ ,  $H_1(X) = -\ln[\theta(X)]$  and  $U = \ln\left[\frac{-\ln(1-\varepsilon)}{\xi}\right]$ . ■

## REFERENCES

- Abbring, J. H., Chiappori, P., Zavadil, T., 2008. Better safe than sorry? Ex ante and ex post moral hazard in dynamic insurance data. Working Paper, VU University Amsterdam and Columbia University.
- Altonji, J., Matzkin, R., 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* 73, 1053-1102.
- Barrientos-Marin, J., Sperlich, S., 2010. The size problem of bootstrap tests when the null is non- or semiparametric. *Revista Colombiana de Estadística* 33, 307-319.
- Blundell, R., Powell, J., 2003. Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics*, eds. M. Dewatripont, L. Hansen, and S. Turnovsky, 294-311, Cambridge: Cambridge University Press.
- Bierens, H., Ploberger, W., 1997. Asymptotic theory of integrated conditional moment tests. *Econometrica* 65, 1129-1151.
- Blundell, R., Powell, J., 2004. Endogeneity in semiparametric binary response models. *Review of Economic Studies* 71, 581-913.
- Cameron, C., Trivedi, P., 2005. *Microeconometrics: methods and applications*. Cambridge University Press, New York.
- Chiappori, P-A., Komunjer, I., Kristensen, D., 2013. Nonparametric identification and estimation of transformation models. Working paper, Department of Economics, Columbia University.
- Cox, D. R., 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 34, 187-220.
- Dawid, A. P., 1979. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B* 41, 1-31.
- Ekeland, I., Heckman, J. J., Nesheim, L., 2004. Identification and estimation of hedonic models. *Journal of Political Economy* 112, 60-109.
- Engle, R. F., 2000. The econometrics of ultra-high-frequency data. *Econometrica* 68, 1-22.
- Fan, J., Jiang, J., 2005. Nonparametric inference for additive models. *Journal of American Statistical Association* 100, 890-907.
- Fan, J., Yao, Q., Tong, H., 1996. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* 83, 189-206.
- Fan, J., Zhang, C., Zhang, J., 2001. Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics* 29, 153-193.
- Fernandes, M., Grammig, J., 2005. Nonparametric specification tests for conditional duration models. *Journal of Econometrics* 127, 35-68.
- Gozalo, P., Linton, O., 2001. Testing additivity in generalized nonparametric regression models with estimated parameters. *Journal of Econometrics* 104, 1-48.

- Greene, W., 2011. *Econometric Analysis*. 7th edition. NJ: Prentice Hall.
- Hansen, B. E., 2008. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* 24, 726-748.
- Härdle, W., Mammen, E., 1993. Comparing nonparametric versus parametric regression fits. *Annals of Statistics* 21, 1926-1947.
- Heckman, J. J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric Models for duration data. *Econometrica*, 52, 271-320.
- Heckman, J. J., Matzkin, R., Nesheim, L., 2005. Estimation and simulation of hedonic models. In T. Kehoe, T. Srinivasan, and J. Whalley (eds), *Frontiers in Applied General Equilibrium*, pp. 277-340. Cambridge University Press, Cambridge.
- Heckman, J.J., Robb, R., 1986. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In: Wainer H., editor. *Drawing Inferences from Self-Selected Samples*. New York: Springer-Verlag; Mahwah, NJ: Lawrence Erlbaum Associates; pp. 63-107.
- Hoderlein, S., Mammen, E., 2007. Identification of marginal effects in nonseparable models without monotonicity. *Econometrica* 75, 1513-1518.
- Hoderlein, S., Su, L., White, H., Yang, T., 2014. Testing for monotonicity in unobservables under unconfoundedness. Working paper, Dept. of Economics, Boston College.
- Horowitz, J. L., 1996. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica* 64, 103-137.
- Horowitz, J. L., 2001. Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica* 69, 499-513.
- Horowitz, J. L., 2014. Nonparametric additive models. In J. Racine, L. Su and A. Ullah (eds), *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, pp. 129-148. Oxford University Press, Oxford.
- Horowitz, J. L., Mammen, E., 2004. Nonparametric estimation of an additive model with a link function. *Annals of Statistics* 36, 2412-2443.
- Horowitz, J. L., Mammen, E., 2007. Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Annals of Statistics* 35, 2589-2619.
- Horowitz, J. L., Mammen, E., 2011. Oracle-efficient estimation of an additive model with an unknown link function. *Econometric Theory* 27, 582-608.
- Ichimura, H., Lee, S., 2011. Identification and estimation of a nonparametric transformation model. Working paper, Department of Economics, Tokyo University.
- Imbens, G., Newey, W., 2009. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77, 1481-1512.
- Jacho-Chávez, D., Lewbel, A., Linton, O., 2010. Identification and nonparametric estimation of a transformed additively separable model. *Journal of Econometrics* 156, 392-407.
- Kennan, J. E., 1985. The duration of contract strikes in US manufacturing. *Journal of Econometrics* 28, 5-28.
- Keifer, N. M., 1988. Economic duration data and hazard functions. *Journal of Economic Literature* 26, 646-679.
- Lancaster, T., 1979. Econometric models for the duration of unemployment. *Econometrica* 47, 939-956.
- Li, Q., Racine, J., 2003. Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86, 266-292.
- Lu, X., White, H., 2014. Testing for separability in structural equations. *Journal of Econometrics* 182, 14-26.
- Mata, J., Portugal, P., 1994. Life duration of new firms. *Journal of Industrial Economics*, 42, 227-245.
- Masry, E., 1996. Multivariate local polynomial regression for time series: uniform strong consistency rates. *Journal of Time Series Analysis* 17, 571-599.
- Matzkin, R. L., 2003. Nonparametric estimation of nonadditive random functions. *Econometrica* 71, 1339-1375.
- Matzkin, R. L., 2007. Nonparametric identification. In J.J. Heckman and E.E. Leamer (eds), *Handbook of Econometrics*, Vol. 6, pp.5307-5368. Elsevier, B.V.

- Politis, D., Romano, J., Wolf, M., 1999. *Subsampling*. New York: Springer-Verlag.
- Ridder, G., 1990. The nonparametric identification of generalized accelerated failure-time models. *Review of Economic Studies* 57, 167-181.
- Rodriguez-Poo, J. M., Sperlich, S., Vieu, P., 2014. Nonparametric specification testing when the null hypothesis is non- or semiparametric. *Econometric Theory*, forthcoming.
- Su, L., Tu, Y., Ullah, A., 2014. Testing additive separability of error term in nonparametric structural models. *Econometric Reviews*, forthcoming.
- Van den Berg, G., 2001. Duration Models: Specification, Identification and Multiple Durations. In J.J. Heckman and E.E. Leamer (eds), *Handbook of Econometrics*, Vol. 5, pp.3381-3460. Elsevier, B.V.
- White, H., Lu, X., 2011. Causal diagrams for treatment effect estimation with application to selection of efficient covariates. *Review of Economics and Statistics* 93, 1453-1459.