Additive Separability and Excess Unobserved Heterogeneity: A Test for Hicks-neutral Productivity Shocks in U.S. Manufacturing Industry

Zhutong Gu*

Peking University, HSBC Business School

Abstract

Additive separability of unobservables is one of the commonly imposed assumptions of testing interest in structural estimation. This paper considers models with unobserved heterogeneity of unrestricted dimensions and proposes a test based on the equality of average structural functions from competing specifications. To implement, a Wald-type test statistic, combining empirical differences over quantile regions, is suggested. I investigate its statistical properties and extend it to various settings. In the application, I test the functional form of Hicks-neutral productivity shocks in estimating firm-level production functions. The results highlight a period of non-Hicks-neutral productions in the U.S. manufacturing industries during the late 90s.

Key Words: Additive Separability; Average Structural Function; Nonparametric Specification Test; Production Functions; Hicks-neutrality

JEL Classification: C12; C14; L60

*Corresponding author email: zgu@phbs.pku.edu.cn. Tel: 86-18617094914. Address: Office 613, Peking University HSBC Business School, Xili University Town, Shenzhen, China

1 Introduction

Nonseparability has important implications in structural econometric modeling and estimation. On the one hand, economic theory rarely specifies that the unobservables enter the structural equations in an additive manner. As a result, the imposition of additivity is equivalent to the absence of unobserved individual heterogeneity in marginal effects. On the other hand, identifiability of many economic objects relies critically on the additivity of disturbance. Roughly speaking, the consistency of many estimators such as instrumental variable (IV) estimators would not be established once unobservables enter in nonlinear fashions. Therefore, a specification test of structural separability should be empirically useful. However, a plethora of empirical research has only focused on additive models until recently. As noted, the additivity may hide the problem of multiplicity of unobservables. In practice, unobservables are often multi-dimensional and in most cases, even the dimension is unknown. The presence of *excess* unobserved heterogeneity complicates the testing problem, as it implies that structural functions cannot be identified without imposing substantial shape restrictions or distributional assumptions. As motivated by those facts, this paper is trying to clarify the benefits and costs of allowing fully flexible unobserved heterogeneity from a testing perspective and suggests a testing method based on the equality of average structural functions of competing specifications. Then I apply it to test the Hicks-neutral productivity shocks, a commonly assumed functional form in the identification and estimation of firm-level production functions.

Unobservables in microeconomics could represent heterogeneity in consumer tastes, product attributes, firm-specific productivity shocks, measurement and unexpected errors, etc. More importantly, a notable fact is that unobserved heterogeneity is rarely unit-dimensional and quite often even the number of dimensions is not even known *a priori*. Browning and Carro (2007) argue that most empirical models permit less heterogeneity than is actually present there. As the proliferation of individual and firm level data, empirical research has begun to take into account a richer set of heterogeneity. However, without

imposing substantial shape restrictions or distributional assumptions, structural functions are generally not identified. The non-identification of structural functions further renders additive separability non-testable. To circumvent this difficulty, I build the test around the average structural functions (ASFs), which is indeed an identified object in most settings. The identification of ASFs is achieved according to the control function literature (Blundell and Powell, 2004; Imbens, 2007; Imbens and Newey, 2009; Florens et al., 2008; Masten and Torgovitsky, 2014; D'Haultfœuille and Février, 2015; Torgovitsky, 2015, etc). In this paper, I show that testing the equality of ASFs generated by the two competing specifications can give rise to important implications on additive separability, though not being equivalent. In particular, when the true model is separable, ASFs under both models are consistent and identical; but the equality is likely to fail otherwise. For illustrative purpose, I give several counterexamples where the test has zero power. But even under those cases, the proposed test is still meaningful. As motivated by Blundell and Powell (2003), ASF should be the central object of estimation interest, since it suffices to answer many economic questions that empirical researchers care about. The testing result can be also informative in terms of the consistency and efficiency of ASF estimators and its related variants.

This paper contributes to the literature of testing additive separability while putting an emphasis on unrestricted unobservables. In particular, I try to fill this gap and propose an easy-to-implement test by relating structural separability to the equality between ASFs generated by the two competing models. Despite its empirical importance, testing additive separability has only received limited attention so far. Lu and White (2014) show that structural additivity can be transformed into a conditional independence assumption using a control function approach. They require the scalar monotonicity in an unobservable or some polynomial structure to establish the equivalence of tests. Su et al. (2015) provide a test against global alternatives by using the derivative of a normalized structural function, whereas the identification of which requires the scalar monotonicity in unobservables. Such assumption puts substantial restrictions on the form of unobserved heterogeneity

and might limit the scope of its applicability where flexible modeling of unobservables is necessary. Other related works include Huber and Mellace (2014) who propose a test in the context of sample selection, Heckman et al. (2010) who consider testing for the correlated random coefficient models. Hoderlein and Mammen (2009) mainly discuss the identification and estimation of local average structural derivatives and briefly mention that a test for separability can be conceived through the quantile structural functions. There are also nonparametric tests on scalar monotonicity, such as Su et al. (2016) in cross-sectional context and Hoderlein et al. (2011) for panel data models. Lewbel et al. (2015) consider a specification test of transformation models in an application of generalized accelerated failure-time models. An incomplete list of other related works include, but are not limited to, Heckman et al. (2010); Fan and Li (1996); Schennach et al. (2012); Sperlich et al. (2002), etc.

The secondary contribution is that an easy-to-implement nonparametric empirical quantile mean (EQM) test in the spirit of Klein (1993) has been developed. The main idea of the test is to compare average differences between two functions in quantile regions of observables. For each quantile region, the average difference would converge to a normal distribution in large samples. Combining those of all quantiles, one can obtain a Wald-type test statistic only using standard central limit theory. The proposed test statistic permits a closer investigation of the heterogeneous functions at each quantile region and can be informative on which quantile region conveys the most power. To control for the asymptotic bias, I employ the recursive bias correction technique recently developed by Shen and Klein (2017). In addition, as opposed to many nonparametric tests that rely on bootrapped inference, I find that the asymptotic variance of our test statistic performs reasonably well under moderate sample sizes. Next, I extend the fully nonparametric test to other empirically driven scenarios. In the first case, I address the problematic "curse of dimensionality" by imposing a single index restriction. To perform the semiparametric test, the finite-dimensional parameters are estimated in the first stage (Powell et al., 1989;

Ichimura and Lee, 1991; Ichimura, 1993; Klein and Spady, 1993, etc) and then the standard testing procedure then follows. In the second extension, I consider the two-stage test of nonparametric nonseparable triangular simultaneous equations models with the "generated" control variable estimated in the first stage.

Finally, the proposed method is being applied in testing a fundamental functional form assumption—Hicks-neutral technological shocks when estimating nonparametric production functions. In the cross-sectional content with firm-level data, Hicks-neutrality implies that firm-specific productivity is only multiplicative to the production function. It essentially eliminates unobserved heterogeneity in input substitution patterns across firms, ruling out labor or capital-augmented technological changes. More importantly, non-Hicks-neutrality, if present, would make most of the commonly used identification strategies invalid, including IV, dynamic panel and proxy variables, all of which depend on the log additivity of unobservables which is a direct implication of Hicks-neutral productivity shocks. As a result, inconsistent estimates of structural parameters, such as output-input elasticities and return-to-scale, could be produced. In this paper, I empirically test the log additivity of unobservables in the U.S. manufacturing industry from 1990 to 2011. The empirical results suggest that there are indeed some years of non-Hicks-neutral production in the late 90s. This finding coincides with a period of mass adoption of computer technology amongst manufacturing firms.

The rest of the paper is structured as follows. Section 2 further provides motivations for testing additive separability and discusses identification issues of nonseparable models with excess heterogeneity. Section 3 reviews the identification results of ASFs under competing specifications and derives the suggested testing implications. Section 4 provides the nonparametric test statistics in a heuristic manner along with the asymptotic properties. Next, finite sample performance is summarized in Section 6. Two extensions are presented in Section 7. In Section 8, I discuss the motivations for testing Hicks-neutral productivity and then cast the issue into the proposed test. The empirical results are also presented.

Section 9 concludes the paper. All proofs are given in the appendix.

2 Nonseparability and Unobserved Heterogeneity

2.1 Nonseparability

Nonparametric nonseparable models have been gaining popularity in theoretical econometrics for the past decades. The single equation nonseparable model considered in this paper, as seen in Eq. (2.1), allows arbitrary interactions between observed and unobserved covariates, e.g. X versus ε.

$$Y = m(X, \varepsilon) \tag{2.1}$$

where the unknown measurable function $m : \mathcal{X} \times \mathcal{E} \to \mathbb{R}$ is called the structural function representing some primitive economic relations. Such models are capable of capturing both observed and unobserved heterogeneity in the structural parameters of economic interest. For instance, model (2.1) can represent a nonparametric firm production function, where Y denotes the output level, X as amount of factor inputs and ε consisting of multi-dimensional unobservables including time-varying and time-invariant productivity shocks, input quality variations, measurement errors in output and inputs, and other disturbances pertaining to demand and cost conditions. Model (2.1) is also general enough to include an entire class of random coefficient models that are widely used in modeling unobserved individual heterogeneity, e.g. $Y = X'\varepsilon$, where X and ε are conformable vectors.

Being a special case of model (2.1), a competing class of specifications specially favored in empirical works, assumes the additive separable structure in which the unobservables can be collectively written as an added term,

$$Y = m_1(X) + m_2(\varepsilon) \tag{2.2}$$

where $m_1(\cdot)$ is an unknown measurable function of only observables defined on $\mathcal{X}$. It includes linear regression models, i.e. $Y = X'\beta + \varepsilon$ as a special case. The additive error, ε, is often taken to include measurement errors and omitted variables. The structural function $m_1(\cdot)$ is often identified by the conditional expectation function of Y when X are exogenous. In the presence of endogenous regressors, exogenous sources of variation are often needed for identification, such as instrumental variable (IV) (Newey and Powell, 2003; Carrasco et al., 2007; Horowitz, 2011, etc) or control function approach (Newey et al., 1999, etc).

Given the above nested specifications, I first define the following set of hypotheses that may be of testing interest to empirical researchers.

$$\begin{aligned} \mathbb{H}_0^* &: m(X,\varepsilon) = m_1(X) + m_2(\varepsilon), \text{ a.s.} \\ \mathbb{H}_1^* &: \text{Otherwise} \end{aligned}$$

According to previous literature, the motivations for testing hypotheses $\mathbb{H}_0^*$ against $\mathbb{H}_1^*$ are mainly fourfold. 1). It is a test on the absence of unobserved individual heterogeneity in structural functions. Once $\mathbb{H}_0^*$ holds, it implies the partial effect of X is deterministic given the level of observed covariates. For example, when estimating the wage equations, additivity implies that individual return-to-education is not affected by unobserved intellectual ability. In estimating production functions, the log additivity of errors amounts to the Hicks-neutral technology, suggesting the elasticity of substitution is not affected by firm-specific productivity shocks. 2). It is a test of the validity of some classes of estimators whose consistency relies crucially on the separability of disturbances, such as IV estimators. Hahn and Ridder (2011) show that the conditional mean restriction, often assumed in IV methods, only has identification power when the model is additive in unobservables. 3). There are more efficient estimators given the additional parametric structure under $\mathbb{H}_0^*$. Hahn et al. (1998) and Imbens and Wooldridge (2009) note that the asymptotic variance bounds of average treatment effect (ATE) can be made much smaller once additive separable unobservables

can be validated. 4). Testing separability could yield implications on endogeneity, a point mentioned already in Imbens (2007) and Imbens and Newey (2009). For instance, suppose a firm makes decisions on input choices X by maximizing the expected profit given its available private information η on the productivity shock ε.

$$X = \arg\max_{x} E[m(x, \varepsilon)|\eta] - C(x, Z) = h(Z, \eta)$$

where the output price is normalized to 1. For each z. $C(\cdot, z)$ is a regular cost function. Z can be cost shifters, such as the hourly labor wage. The solution X is endogenous because it is correlated with the structural error ε through the private information η. Now suppose $m(x, \varepsilon)$ is additive separable in which case the objective function becomes $m_1(x) + E[\varepsilon|\eta] - C(x, Z)$. Under this scenario, X is just a deterministic function of Z alone and therefore becomes exogenous whereas contradicting most empirical findings.

2.2 Non-identification under Excess Unobserved Heterogeneity

Unfortunately, testing $\mathbb{H}_0^*$ against $\mathbb{H}_1^*$ might not always have power when unobserved heterogeneity, ε, is modeled fully flexibly. Before outlining the suggested testing framework, I want to highlight the identification problem associated with multi-dimensional unobservables in nonseparable structural models. For this reason, I will modify the hypotheses of testing interest later.

This point can be made clear from the simple example given below in the spirit of Benkard and Berry (2006). Suppose X is univariate continuous variable independent of $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ that are independently distributed as standard normals. There is no way to distinguish between

$$Y = \frac{X}{\sqrt{X^2 + 1}}\varepsilon_1 + \frac{1}{\sqrt{X^2 + 1}}\varepsilon_2 \quad \text{versus} \quad Y = \varepsilon_3$$

in the sense that the above two models generate identical joint distributions of observables,

e.g. $F_{X,Y}$, which consists of all available information in the data. To see this, it is straightforward to show $Y \sim N(0,1)$ and is independent of X in both specifications. In the language of Roehrig (1988), let $F_{X,\varepsilon}$ be the distribution and $m(x,e)$ be the structure. This pair $(F_{X,\varepsilon}, m(x,e))$ defines the data generating process. As it can be seen, both the nonseparable and separable models implies the same joint distribution of observables, $F_{X,Y}$. Therefore, the above two structural functions $m(x,e)$ are observationally equivalent. Without further restrictions, $m(x,e)$ is not identified. As a consequence, it implies that testing the original hypothesis of $\mathbb{H}_0^*$ versus $\mathbb{H}_1^*$ may not have power in general because both nonseparable and separable models can deliver the same underlying data generating process.[1]

One solution is to impose additional structures in order to achieve identification (see Matzkin, 2003). In the context of testing for structural separability, previous works have focused on imposing shape restrictions such as scalar monotonicity in unobservables to attain identification of the structural function. Su et al. (2015) assumes that the error term is unidimensional and $m(x,\varepsilon)$ is strictly monotonic for each x. By taking the derivative of the identified structural function, they arrive at a consistent test which also has power against strictly monotonicity if it doesn't hold. Lu and White (2014) transform the original hypothesis into a conditional independence condition. However, they lose equivalence unless $m_1(\cdot)$ is some polynomial function or strictly monotonic in the unobservable.

In many situations, it is undesirable to impose assumptions such as scalar monotonicity, aforementioned as they are often subject to test in its own right. For example, in the empirical application of this paper, production functions often involve multiple unobserved shocks, including productivity, ex post shocks as well as other idiosyncratic errors. As a consequence, restricting it to single dimension can be hard to justify theoretically. Another direction is to determine what can be identified without compromising the dimensionality of heterogeneity (Blundell and Powell, 2004; Imbens and Newey, 2009, etc). In many cases,

[1]This is a fallacy pointed out by Benkard and Berry (2006). They revisit the identification results of simultaneous equations models from Brown (1983) and Roehrig (1988) and show that a supporting lemma (called derivative condition) is incorrect.

it would be unnecessary to recover the structural functions if the identified parameters are sufficient to answer the economic questions of interest. And this is the exactly approach that this paper undertakes. In the next section, I will derive the testable implication based only on the ASF that is identified even in the presence of excess heterogeneity. For multivariate unobservables, there are papers dealing with identification and estimation via restrictions like single index property (Benkard and Berry, 2006; Matzkin, 2007, 2008; Chernozhukov et al., 2007; Chesher, 2009, etc), which may be exploited to develop other testing procedures.

3 Testing with Average Structural Functions

In this section, I first review the existing identification results of ASFs and then derive testable implications for structural separability. Define the ASF at $X = x$ of nonseparable models (3.1) as

$$ASF(x) \equiv g(x) = \int_{\mathcal{E}} m(x, e) dF_{\varepsilon}(e), \quad \forall x \in \mathcal{X} \tag{3.1}$$

where calligraphic letters denote the support on which F_{ε}, the distribution function of ε is defined. The function $g(\cdot)$ is structural in the sense that X can be manipulated arbitrarily without changing the marginal distribution of ε, the counterfactuals of which may be of policy interest. Also motivated in Blundell and Powell (2003), ASFs should be the central object of estimation interest, by which many useful structural objects can be easily constructed. For instance, when X is binary, the ATE can be obtained by the difference between $g(1)$ and $g(0)$. The identification of ASF as well as a number of related objects have been studied in Imbens and Newey (2009).

Now recall the nonseparable model (2.1), where the unknown structural function is defined on $\mathcal{X} \times \mathcal{E}$ where $\mathcal{X} \subset \mathbb{R}^{d_X}$ and $\mathcal{E} \subset \mathbb{R}^{\infty}$.[2] The identification of ASF without endogeneity is trivial, as suggested by the reduced form regression $g(x) = E(Y|X = x), \forall x \in$

[2]The test proposed also works for discrete X. For brevity, I only demonstrate the continuous case.

$\mathcal{X}$. Unfortunately, many economic models would include at least one endogenous regressor. For instance, the "simultaneity bias" could arise from the dependency between choices of variable inputs and unobserved productivity shocks when estimating firm-level production functions. To handle endogeneity, this paper employs the control function approach, widely used in the literature. Suppose that the control variables $V \in \mathcal{V} \subset \mathbb{R}^{d_V}$, where $\mathcal{V} = \text{supp}(V)$, satisfy Assumption I.1 and I.2.

Assumption I.1 Conditional independence. $X \perp \varepsilon | V$, where X and ε are not measurable with respect to σ-field generated by V.

Assumption I.2 Large support. $\mathcal{V} = \mathcal{V}^x, \ \ \forall x \in \mathcal{X}, \ \$ a.s. where $\mathcal{V}^x = \text{supp}(V|X = x)$.

Assumption I.1 parallels the unconfoundedness condition in the treatment effect literature, assuming independence between X and ε conditional on V. Loosely speaking, it also requires that X and ε cannot be exact functions of V; otherwise, they would be degenerate given V. Admittedly, Assumption I.2 is a relatively strong condition. In the absence of conditional large support of V, ASF is only partially identified with sharp bounds (Imbens and Newey, 2009). On the other hand, the large support condition might hold only over some region of X, say $\mathcal{X}_0$, instead of the whole support. In this case, ASF is identified only over the region, $\mathcal{X}_0$ and fortunately, the test is still valid, though the effective sample used to construct the test statistic needs to be shrunken accordingly.

There are many ways to obtain the control covariates. In some cases, V might be readily available and observed in the dataset. For example, IQ test scores are often employed to control for the omitted intellectual ability in estimating returns to education. Once panel data is available, within group summary statistics may be adequate to control for the endogeneity. Moreover, the control variables can be "generated" through the triangular simultaneous equations frameworks. In the application of testing Hicks-neutral productivity shocks, I use the "proxy" variable derived from the intermediate input demand function and to control for the unobserved productivity shocks. For now, I just presume the control

variables V satisfying Assumption I-1 and I-2 available so as to simplify the explication of the testing idea. Next, I focus on the identification of ASFs for nonseparable models and additive separable models, respectively, preceding the discussion of the testable implications.

3.1 Identification of ASF of Nonseparable Models

Proposition 3.1 is borrowed from the nonseparable model literature (Blundell and Powell, 2004; Imbens and Newey, 2009, etc). It can be shown that ASF is identified by integrating out the conditional expectation function (CEF) with respect to the marginal distribution of control variables. The proof is given in Appendix A.

Proposition 3.1. *Under Assumption I.1 and I.2, $g(\cdot)$ defined in Eq.* (3.1) *is identified at each $x \in \mathcal{X}$,*

$$g(x) = \int_{\mathcal{V}} C(x, v) dF_V(v) \tag{3.2}$$

where the CEF is defined as $C(x, v) \equiv E(Y|X = x, V = v)$ and F_V is the CDF of V on $\mathcal{V}$.

Note that both $C(x, v)$ and F_V can be estimated from the data. All available information is summarized by the joint distribution of observables, i.e. $F_{Y,X,V}$. By Proposition 3.1, related "structural" parameters such as average partial effects, are subsequently obtained, provided existence.

3.2 Identification of ASF of Separable Models

A popular subclass of models admits an additive structure between observables and unobservables as in Eq. (2.2). Such models impose substantial restrictions on the way how unobserved heterogeneity enters. It indicates constant partial effects conditional on observed covariates. Despite those shortcomings, it has received most attention in both theoretical and empirical works. Under Assumption I.1 and I.2, the ASF of model (2.2) is

immediately identified through Proposition 3.1 since additive models belong to a subclass of nonseparable models. However, a weaker set of assumptions suffices to identify $a(\cdot)$, as stated in Assumption I.1′and I.2′. The identification of nonparametric additive models has been studied in Newey et al. (1999).

Assumption I.1′ Conditional mean independence. $E(m_2(\varepsilon)|X,V) = E(m_2(\varepsilon)|V) \equiv h(V)$, a.s.

Assumption I.1′ doesn't require full independence conditional on the control variates as the mean independence condition is sufficient. Intuitively, X would not provide any additional information on the average of unobservables given the knowledge of V. Also note that under Assumption I.1′, the CEF becomes additive in the unknown functions of X and V,

$$C(x,v) = m_1(x) + h(v), \forall (x,v) \in \mathcal{X} \times \mathcal{V} \tag{3.3}$$

Assumption I.2′ Nonexistence of additive functional dependence. $\Pr(\delta(X) + \gamma(V) = 0) = 1$ implies there is a constant c that $\Pr(\delta(X) = c) = 1$, for any differentiable functions $\delta : \mathcal{X} \to \mathbb{R}$ and $\gamma : \mathcal{V} \to \mathbb{R}$.

Assumption I.2′ rules out the possibility of exact additive functional dependence between $m_1(x)$ and $h(v)$. To see this, suppose there is another set of functions, $\widetilde{m}(x)$ and $\widetilde{h}(v)$ such that $\Pr(\widetilde{m}(X)+\widetilde{h}(V) = m(X)+h(V)) = \Pr(\delta(X)+\gamma(V) = 0) = 1$, where $\delta(\cdot) = \widetilde{m}(\cdot) - m(\cdot)$ and $\gamma(\cdot) = \widetilde{h}(\cdot) - h(\cdot)$. Then $m(\cdot)$ and $h(\cdot)$ are generally not point identified unless both are degenerate. The formal proof is given in Newey et al. (1999) and the identification result is summarized in Proposition 3.2.

Proposition 3.2. *Under Assumption I.1′and I.2′, a).* $m_1(\cdot)$ *and* $h(\cdot)$ *in Eq.* (3.3) *is identified up to an additive constant for each* $(x,v) \in \mathcal{X} \times \mathcal{V}$*. b).* $g(\cdot)$ *defined in Eq.* (3.1) *is identified at each* $x \in \mathcal{X}$*,*

$$g(x) = m_1(x) + c_h, \quad \textit{where } E[h(V)] = c_h$$

Without loss of generality, one can normalize that $E[h(V)] \equiv c_h = 0$, attributing all constants into $m_1(\cdot)$. I adopt this normalization to ease the following exposition. And under this case, it is true that $m_1(\cdot) = g(\cdot)$. In addition, it implies that $h(\cdot)$ can be identified in Eq. (3.4).[3]

$$h(v) = \int_{\mathcal{X}} C(x, v) dF_X(x) - E(Y) \tag{3.4}$$

The additive structure of model (2.1) provides us with the additional information which can be exploited to recover the ASF through the one-step backfitting procedure as proposed in Linton (1997). Nonetheless, for nonseparable models, it need not hold in general. Alternatively, define the conditional expectation of $Y - h(V)$ to be $a(\cdot)$ given $X = x$ in Eq. (3.5),

$$a(x) = E(Y - h(V)|X = x), \ \forall x \in \mathcal{X} \tag{3.5}$$

In Proposition 3.3, it states that $a(\cdot)$ identifies ASFs for additive models, i.e. $a(x) = m_1(x)$. But it is generally not ture for nonseparable models. This also explains why tests that based on ASFs might have some power on testing additive separability, though not being equivalent.

Proposition 3.3. *Under Assumption I.1 and I.2, for each $x \in \mathcal{X}$, a). for additive models of (2.2), $a(x) = g(x)$; b). for nonseparable models of (2.1), $a(x) = g(x)$ if and only if the following condition holds a.s.*

$$\int_{\mathcal{V}} C(x, v)(dF_{V|X}(v|x) - dF_V(v)) = \Delta(x), \quad \forall x \in \mathcal{X}$$

where $\Delta(x) = \int C(x', v) dF_X(x') dF_{V|X}(v, x)$.

[3]It is straightforward to verify that $E[h(V)] = 0$ because $\int C(x, v) dF_X(x) dF_V(v) = \int C(x, v) dF_{X,V}(x, v)$ when $c(x, v)$ is additive.

The equality in Proposition 3.3 a), is trivial to hold. Nonetheless, for nonseparable models, it would be hard to come up with any intuitive interpretation for the condition in b). It does not seem possible to characterize the entire class of models satisfying this property. To illustrate this condition, I present 3 examples of nonseparable models below that can produce the equality. Those are exemplary cases where the ASF-based test only has zero power on testing additive separability. Example 1 is to illustrate that a nonseparable model can be rewritten as an additive model in general without endogenous regressors. Example 2 manifests that a nonseparable model is able to generate an additive CEF that in turn can produce a ASF equal to some additive model. Example 3 shows that despite the non-additive CEF, the ASF of a nonseparable model could still be equal to that of some additive structural model after integration.

Example 1. Suppose that $X \perp \varepsilon$ and V is of null dimension, a situation of no endogeneity. Then, $g(x) = a(x)$ for all x. Even if the true model is nonseparable, it can always be written as the additive one,

$$Y = E(Y|X) + \epsilon, \quad \epsilon = m(X, \varepsilon) - E(Y|X)$$

where $E(\epsilon|X) = 0$. In the case of only exogenous observables, $E(Y|X = x, V = v) = \widetilde{C}(x)$, so the condition in Proposition 3.3 holds and $a(\cdot)$ recovers the ASF for nonseparable models.

Example 2. This example demonstrates that a nonseparable model can generate an additive CEF, thus producing a ASF equivalent to that of some additive model. Suppose the nonseparable model is given as follows,

$$Y = X\varepsilon_1 + \varepsilon_2$$

where $E(\varepsilon_1|X, V) = c$ for some constant c_1 and $E(\varepsilon_2|X, V) = h(V)$. The CEF then becomes additive in x and v, i.e. $C(x, v) = cx + h(v)$. Then it is not hard to see $a(x) = g(x), \forall x$, a.s.. This example is taken from Lu and White (2014) who argue that testing separability

according to CEF has no power in situations like this.

Example 3. This example shows even though the CEF is not additive in X and V, $a(\cdot)$ may still be equal to $g(\cdot)$ due to the integration. Suppose $V = \varepsilon$,

$$Y = X\varepsilon, \quad E(\varepsilon|X) = 0$$

Be aware that the mean independent condition doesn't imply the full independence between X and ε. The CEF generated by this structural function is $C(x, v) = xv$. The ASF is therefore $g(x) = xE(V) = 0$ if $E(V) = E(\varepsilon) = 0$, then

$$a(x) = xE(\varepsilon|X = x) - E(X)E(V|X = x) = 0 = g(x), \quad \forall x$$

One can also verify the condition in Proposition 3.3 does hold in all the above examples.

3.3 Testing Implications

As discussed earlier, this paper attempts to test additive separability with unrestricted unobservables. Unfortunately, the original set of hypotheses turns out to be non-testable due to the non-identification of structural functions once excess heterogeneity has to be allowed. Hence in this paper, instead of testing $\mathbb{H}_0^*$ against $\mathbb{H}_1^*$, I consider another set of testable hypotheses below,

$$\begin{aligned} \mathbb{H}_0 &: D(X) \equiv g(X) - a(X) = 0, \quad \text{a.s.} \\ \mathbb{H}_1 &: \text{Otherwise.} \end{aligned}$$

This is essentially to see whether the ASFs obtained under the two competing specifications are identical. The power of this test comes from the fact that $g(\cdot)$ in Eq. (3.2) recovers the ASF for both models whereas $a(\cdot)$ in Eq. (3.5) only recovers the ASF for additive models (and a small class of nonseparable models satisfying the condition stated in Proposition 3.3).

Admittedly, $\mathbb{H}_0$ versus $\mathbb{H}_1$ is no longer equivalent to $\mathbb{H}_0^*$ versus $\mathbb{H}_1^*$. However, the benefits of doing so are threefold. First, $\mathbb{H}_0$ is indeed a testable hypothesis with minimal assumptions (no shape restrictions or distributional assumptions) in contrast to the non-testable original hypotheses. Such relaxation on unobservables is deemed useful in many real situations where the amount of heterogeneity could not be fully controlled through observed characteristics. Second, the test still has reasonable power against additive separability in many applications, though not against global alternatives for $\mathbb{H}_0^*$, as can be seen from our finite sample simulations in Section 6 and empirical results in Section 8. Finally, were ASFs and its variants sufficient to answer the research questions, there would be no need to test the original set of hypotheses. Besides, once $\mathbb{H}_0$ cannot be rejected, more efficient estimators could be available by incorporating this additional information and treating the model as if it had an additive error structure. So are the variants of ASFs, such as the estimators of average marginal effects which are usually of primary interest in many empirical microeconomic studies.

Note that the inequality of ASFs, i.e. $g(x) \neq a(x), \forall x$ indicates nonadditive of CEF, subsequently indicating a nonseparable structural function, $m(\cdot, \cdot)$. However, the reverse is not true in general. Example 1-3 can be taken as counterexamples. This might be a shortcoming of the suggested test as the equivalence is lost, therefore reflecting the trade-off of incorporating maximal heterogeneity. Hence, researchers should be advised when making a conclusion on structural separability when $\mathbb{H}_0$ cannot be rejected. On the other hand, the specification test of ASFs is also of great importance in its own right as it can shed light on the consistency and efficiency of ASF estimators. From now on, I will only focus on the suggested set of hypotheses—$\mathbb{H}_0$ versus $\mathbb{H}_1$.

4 Estimation and Testing

4.1 Estimation

I first discuss the nonparametric estimator for CEF which is the central building block for the test statistic. In this paper, I focus on the Nadaraya-Watson (or local constant) estimator (Nadaraya, 1964) to estimate conditional mean functions. In principle, other nonparametric smoothers such as local polynomials and sieve estimators can be applied as well. To facilitate the proof of asymptotic theory, leave-one-out estimators are used throughout and subscripts of the leave-one-out indicators are suppressed for notational brevity whenever the context is self-evident. Recall that $C(x,v) \equiv E(Y|X=x,V=v)$. Given any non-boundary set of points, $(x,v) \in \mathcal{X} \times \mathcal{V}$, the preliminary kernel estimator is defined in Eq. (4.1),

$$\widehat{C}_0(x,v) = \frac{\sum_{i=1}^N K_{h_1}(X_i - x)K_{h_1}(V_i - v)Y_i}{\sum_{i=1}^N K_{h_1}(X_i - x)K_{h_1}(V_i - v)} \tag{4.1}$$

where admitted some of abuse of notation, $K_h(\cdot) = \prod_d [k(\cdot/h)/h]$ represents the d-dimensional product of kernel functions. Bandwidths are allowed to be different for X and V.

To make sure that the asymptotic bias vanishes faster than $\sqrt{N}$, I suggest to use the recursive nonparametric conditional mean estimator recently proposed by Shen and Klein (2017), due to its bias-reducing property.[4] Simply put, I firstly construct the local bias from the preliminary kernel estimator, e.g. $\widehat{\delta}_i(x,v) \equiv \widehat{C}_0(X_i,V_i) - \widehat{C}_0(x,v)$ and then apply the local constant estimator again on the "bias-free" dependent variable, $Y_i - \widehat{\delta}_i(x,v)$. So the bias-reducing conditional mean estimator can be thus obtained in Eq. (4.2).

$$\widehat{C}(X_l,V_j) = \frac{\sum_{i\neq j,l}^N K_{h_1}(X_i - X_l)K_{h_1}(V_i - V_j)[Y_i - \widehat{\delta}_i(X_l,V_j)]}{\sum_{i\neq j,l}^N K_{h_1}(X_i - X_l)K_{h_1}(V_i - V_j)} \tag{4.2}$$

[4]Other bias reducing methods such as higher order kernels, local smoothing also work in theory. However, it is found that using higher order kernels are likely to produce unreasonably large limiting variances of the test statistic under the null in the finite sample simulations.

where the leave-one-out kernel estimator is used when it is evaluated at (X_l, V_j), $l, j \in \{1, 2, \cdots, N\}$.

Next I consider the estimation of ASF. Linton and Nielsen (1995) suggest a marginal integration method and Newey (1994) consider the partial mean estimator. This paper employs the latter approach since taking the partial mean is more computationally straightforward when V is multi-dimensional. When evaluated at X_l, the nonseparable ASF, $g(X_l)$, is estimated with the leave-one-out partial mean estimator $\widehat{g}(X_l)$ in Eq. (4.3),

$$\widehat{g}(X_l) = \frac{1}{N-1} \sum_{j \neq l}^{N} \widehat{C}(X_l, V_j), \quad \forall l = 1, \cdots, N \tag{4.3}$$

Likewise, $h(\cdot)$ defined in Eq. (3.4) can be estimated in the similar way in Eq. (4.4)

$$\widehat{h}(V_j) = \frac{1}{N-1} \sum_{i \neq j}^{N} \widehat{C}(X_i, V_j) - N^{-1} \sum_{i=1}^{N} Y_i, \quad \forall j = 1, \cdots, N \tag{4.4}$$

where the mean of Y subtracted resembles the sample analog of the unconditional expectation, $E(Y)$. This subtraction ensures the normalization that the unconditional mean of $h(\cdot)$ is 0.[5]

Now consider the ASF estimator of the "additive" model, $\widehat{a}(\cdot)$. I borrow the idea from Linton (1997) who considers the one-step backfitting procedure implied by the constructive identification strategy in the previous section. The ASF estimator here differs from Linton's in that the partial mean of kernel estimator is employed rather than the marginal integration of the local linear estimator. Linton also argues that the one-step backfitting estimator is preferred to the alternating conditional expectation (ACE) approach in estimating the nonparametric additive regression models in a multitude of aspects. ACE, also known as the "backfitting" procedure, has a long-standing history in statistics literature (Hastie and Tibshirani, 1990) and is thought to yield the most efficient estimator since it finds the

[5]Note that the CEF estimator in constructing $\widehat{h}(\cdot)$ could be potentially different from the one in Eq. (4.3) in terms of bandwidth and kernel choices.

unique orthogonal projection of Y onto the space of additive functions providing the best mean square error approximation. However, such iterative nature often requires intensive computational effort. Moreover, closed-form solutions are hard to derive and this prevents further study of its asymptotic properties. So from now on, I adopt its simple one-step counterpart unique to the additive models, presuming that $h(\cdot)$ is known. And the infeasible estimator (or oracle estimator) of $a(\cdot)$ is given in Eq. (4.5).

$$\widetilde{a}(X_l) = \widehat{E}_{h_2}(Y_i - h(V_i)|X_l), \quad l = 1, \cdots, N \tag{4.5}$$

where $\widehat{E}(\cdot)$ is the bias-reducing recursive conditional mean estimator similar to Eq. (4.2), with the bandwidth, $h_2 \to 0$ as $N \to \infty$. By simply substituting $\widehat{h}(V_i)$ for the unknown function $h(V_i)$, one can obtain the feasible estimator $\widehat{a}(\cdot)$ in Eq. (4.6),

$$\widehat{a}(X_l) = \widehat{E}_{h_2}(Y_i - \widehat{h}(V_i)|X_l) = \frac{\sum_{i \neq l}^{N} K_{h_2}(X_i - X_l)[Y_i - \widehat{h}(V_i) - \widehat{\delta}_i(X_l)]}{\sum_{i \neq l}^{N} K_{h_2}(X_i - X_l)}, \quad l = 1, \cdots, N \tag{4.6}$$

where $\widehat{\delta}_i(X_l)$ is the "bias" of preliminary estimators defined in the similar way when estimating $\widehat{C}(\cdot, \cdot)$. The estimator here differs from Linton (1997) in threefolds. First, kernel estimator is being applied instead of the local linear estimator. Second, partial mean estimator rather than marginal integration is used to estimate the pilot nonparametric function $\widehat{h}(\cdot)$ aforementioned. Finally, Linton seeks the optimal nonparametric rate in the estimation context by setting the bandwidth of order $O(N^{-1/5})$ when getting $\widehat{a}(\cdot)$. In contrast, I am targeting the root-N rate in the testing environment while bias reduction techniques are being used. Nevertheless, our estimator of the "additive" ASF does share the same merit in terms of efficiency. In particular, the one-step backfitting method provides a more efficient estimator of ASF when $\mathbb{H}_0$ is true.[6]

[6]However, our ASF estimator under $\mathbb{H}_0$ is not the most efficient estimator. For further discussion, see Linton (2000).

4.2 Test Statistics

The specification test of $\mathbb{H}_0$ versus $\mathbb{H}_1$ falls into the class of testing the distance between two functions. To this end, the Kolmogorov-Smirnov or Cramer-von Mises test statistics are often applied. But in this paper, I adopt an even simpler test that combines information from empirical quantile means (EQM). The test idea is firstly mentioned in Klein (1993) in specification tests of parametric error distributions versus semiparametric single-index binary choice models.

For the purpose of illustration, consider the univariate continuous variable X with $d_X = 1$ for now, but generalizations to multivariate X is similar. Denote the empirical ASF difference by

$$D(X_i) \equiv g(X_i) - a(X_i), \quad \forall i = 1, \ldots, N \tag{4.7}$$

Under $\mathbb{H}_0$, $D(X_i) = 0$ for each i almost surely. To proceed, divide the whole sample into P_N number of even subsamples or quantile regions thereafter, over the support of X. It can be postulated that for each quantile region, the average empirical difference, is centered at 0 under the null. For multivariate $X_i = (X_{1i}, X_{2i}, \cdots, X_{d_X i})'$, each quantile region can be thought as the intersection of quantiles of each variables. The number of quantiles can be any positive integer so long as $P_N/N = o(1)$ in theory.

Next, define the pth-quantile empirical mean difference as the following,

$$T_N^p \equiv N^{-1} \sum_{i=1}^{N} t_i^p D(X_i), \quad p = 1, \ldots, P_N \tag{4.8}$$

where the quantile-trimming indicator is defined in Eq. (4.9),

$$t_i^p \equiv \mathbf{1}\left\{\min[c_L, q_X(p - 1/P_N)] \leq X_i < \max[q_X(p/P_N), c_U]\right\} \tag{4.9}$$

where $q_X(\cdot)$ is the quantile function of X, i.e. $q_X(\tau) = \inf\{x : F_X(x) \geq \tau\}$. c_L, c_U

are predetermined fixed lower and upper bounds, respectively, to ensure non-existence of boundary biases. Specifically, $t_i^p = 1$ if X_i falls in the pth-quantile region and 0 otherwise. Let $T_N = (T_N^1, \ldots, T_N^P)'$ be a vector of quantile mean differences. Because each T_N^p is simply the sample average centered at 0 under the null, one would expect that T_N converges at the rate of $\sqrt{N}$ to a multivariate normal distribution as $N \to \infty$ according to the standard central limit theorem. A Wald-type statistic in Eq. (4.10) could be thus constructed,

$$W_N \equiv N T_N' \Omega^{-1} T_N \tag{4.10}$$

where Ω is the positive definite weighting matrix and is often taken to be the variance of T_N, i.e. $\Omega \equiv E(T_N T_N')$, see Theorem 5.2 for explicit expressions.

Note that dividing sample into subregions enables researchers to have a closer look across quantiles and discover anomalies hidden in the data and be informative about where the power of the test comes from. Quite often, with a little modification, the test can be performed on a specific region of researchers' interest, instead of over the whole population. For example, policymakers might want to know if unobserved intellectual ability affects the return-to-education for people with only high school diploma. In so doing, the test permits a rich and in-depth characterization based on observed covariates.

A feasible test statistic is made possible by substituting unknown objects with estimators, like in Eq. (4.11)

$$\widehat{W}_N = N \widehat{T}_N' \widehat{\Omega}_N^{-1} \widehat{T}_N \tag{4.11}$$

where

$$\widehat{T}_N = (\widehat{T}_N^1, \ldots, \widehat{T}_N^P)' \tag{4.12}$$

and $\widehat{\Omega}$ is the consistent estimator of Ω to be given later in Eq. (5.8) in the Corollary 5.2.1,

$$\widehat{T}_N^p = N^{-1} \sum_{i=1}^{N} \widehat{t}_i^p \widehat{D}(X_i), \quad p = 1, \ldots, P_N \tag{4.13}$$

where $\widehat{D}(X_i) = \widehat{g}(X_i) - \widehat{a}(X_i)$ and $\widehat{t}_i^p$, a consistent estimator of the trimming indicator, is given in Eq. (4.14).

$$\widehat{t}_i^p \equiv \mathbf{1}\left\{\min[c_L, \widehat{q}_X(p - 1/P_N)] \leq X_i < \max[\widehat{q}_X(p/P_N), c_U]\right\} \tag{4.14}$$

where the quantile function is defined as $\widehat{q}_X(\tau) = \inf\left\{x : N^{-1}\sum_{i=1}^{N} \mathbf{1}(X_i > x) \geq \tau\right\}$.

A final remark is concerning the choice of number of quantile regions P_N. In theory, as long as $P_N/N = o(1)$, the results would hold. But providing the optimal choice of P_N is beyond the scope of this paper. The theory below only considers a fixed number of quantiles for simplicity, i.e. $P_N = P$. In practice, one is advised to experiment with various values for robustness check, e.g. $P_N = 4, 6$ or 8 in the following Monte Carlo studies, though the testing results are found to be relatively robust in Section 6.

5 Asymptotic Results

Before stating the asymptotic assumptions, notations are simplified in the following way and will be carried through the rest of the paper. Let $U_i = (X_i', V_i') \in \mathcal{U} \subset \mathbb{R}^d$, where $d = d_X + d_V$. Let $\mathcal{U}_0$ be the compact subset of $\mathcal{U}$ on which the density of U, f_U is well defined. Also let $f^*(x, v) \equiv f_X(x) f_V(v)/f_U(u)$ for any $(x, v) \in \mathcal{U}_0$, where $f(\cdot)$ denotes the density distribution function of the corresponding continuous variable.

Assumption A.1. DGP. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space on which are defined the random vectors, $(Y, X, V, \varepsilon) : \Omega \to \mathcal{Y} \times \mathcal{X} \times \mathcal{V} \times \mathcal{E}$. $\mathcal{Y} \in \mathbb{R}, \mathcal{X} \in \mathbb{R}^{d_X}, \mathcal{V} \in \mathbb{R}^{d_V}, \mathcal{E} \in \mathbb{R}^{\infty}$ i). $\{(Y_i, X_i, V_i, \varepsilon_i)\}_{i=1}^N$ are i.i.d. ii). $\mathrm{Var}(Y|U) < \infty$. iii). $Y = m(X, \varepsilon)$ where $m : \mathcal{X} \times \mathcal{E} \to \mathcal{Y}$ is a Borel measurable function defined on $\mathcal{F}$.

Assumption A.2. Smoothness. The conditional distribution $F_{Y|U}$ has the uniformly continuous and bounded Radon-Nikodym second order density derivatives. i). f_U is continuous in u and $f_{Y|U}$ is continuous in (y, u). ii). There exists $C > 0$ such that $\inf_{\mathcal{U}_0} f_U > C$ and $\inf_{\mathcal{Y}\times\mathcal{U}_0} f_{Y|U} > C$.

Assumption A.3. Kernel. For some even integer ν, the kernel K is the product of symmetric bounded kernel $k : \mathbb{R} \to \mathbb{R}$, satisfying $\int_{\mathbb{R}} u^i k(u)du = \delta_{i0}$, for $i = 1, 2, \cdots, \nu - 1$, $\int_{\mathbb{R}} u^\nu k(u)du < \infty$ and $k(u) = O((1 + u^{\nu+1+\varepsilon})^{-1})$, for some $\varepsilon > 0$, where δ_{ij} is the Kronecker's delta.

Assumption A.4. Dominance. For any $u \in \mathcal{U}_0$, $E(Y|U = u)$ has all partial derivatives up to νth order. Let $D^j E(Y|U = u) \equiv \frac{\partial^{|j|E(Y|U=u)}}{\partial^{j_1}u_1\cdots\partial^{j_d}u_d}$ where $u = (u_1, \cdots, u_d)'$ and $|j| = \nu$. $D^j E(Y|U = u)$ is uniformly bounded and Lipschitz continuous on $\mathcal{U}_0$: for all $u, \widetilde{u} \in \mathcal{U}_0$, $|D^j E(Y|U = u) - D^j E(y|U = \widetilde{u})| \leq C||u - \widetilde{u}||$, for some constant $C > 0$, where $||\cdot||$ is the Euclidean norm.

Assumption A.5. Bandwidth. As $N \to \infty$, i). $h_1, h_2 \to 0$, $Nh_1^d \to \infty$, $Nh_2^{d_X} \to \infty$, $Nh_1^8 \to 0$, $Nh_2^8 \to 0$, ii). $d = d_X + d_V < 4$.

Assumption A.6. Invertability. $|\det(\Omega)| > 0$ w.p.1

Assumption A.1-A.4 are regularity conditions frequently employed in the literature of nonparametric estimation and testing. Assumption A.1 formally states the data generating process (DGP) and requires the boundedness of conditional variances. The i.i.d. assumption is standard in cross-sectional studies. Assumption A.2 is standard in nonparametric kernel estimation of conditional mean and density. If $\mathcal{U}$ is compact, it is possible to let $\mathcal{U}_0 = \mathcal{U}$; otherwise, trimming could be used to ensure the compactness of the support. Assumption A.3 puts restrictions on the kernel functions. In the rest of this paper, only the second order kernel ($\nu = 2$), such as the standard normal, is required in conjunction with the recursive

bias-reducing procedure.[7] Assumption A.4 provides additional smoothness conditions for derivatives of the conditional mean functions. Assumption A.5 restricts the choices of bandwidth. It implies that the window parameters $(h_i = O(N^{-r_i}), i = 1, 2)$ need to satisfy $1/8 < r_1 < 1/d, 1/8 < r_2 < 1/d_X$. Nevertheless, those restrictions rule out the optimal bandwidth that minimizes the asymptotic mean squared error. Assumption A.5 also restricts the dimension of conditioning variables to be less than 4. In empirical settings where the number of control covariates is of large dimension, I suggest a semiparametric version of the test in Section 7. Assumption A.6 states that the weighting matrix defined in Eq. (5.7) is invertible.

In what follows, I show that the asymptotic null distribution of pth-quantile average difference in Theorem 5.1, with the scratch of the proof outlined below. All details and supporting lemmas are given in Appendix B.

Theorem 5.1. *Suppose that Assumption I.1-I.2 and A.1-A.6 hold, under* $\mathbb{H}_0$*, for any* $p \in (1, 2, \cdots, P)$*, it is true that*

$$\sqrt{N}\widehat{T}_N^p \xrightarrow{\text{D}} N(0, \Omega_p)$$

where

$$\Omega_p = E(\xi_i^p \xi_i^{p'}) \tag{5.1}$$

and the influence function, ξ_i^p*, is defined in Eq.* (5.2)

$$\xi_i^p \equiv [t_i^p + E(t^p|V_i)]f^*(X_i, V_i) - t_i^p]\epsilon_i + E(t^p)h(V_i) \tag{5.2}$$

where $\epsilon_i = Y_i - C(X_i, V_i)$.

[7] I find that the performance of higher order kernels ($\nu = 4$) is unstable in finite samples even though they are valid in theory.

Theorem 5.1 says the quantile average difference in Eq. (4.13) converging to a normal distribution at the parametric rate under $\mathbb{H}_0$. The proof of the above theorem can be roughly divided in three steps. Firstly, I show that the estimated quantile (and trimming) indicator can be replaced by its true counterpart plus remaining terms converging faster than $\sqrt{N}$. Secondly, I show that the empirical difference can be decomposed into three components by substituting for the infeasible estimator, $\widehat{a}_I(\cdot)$. Finally, I utilize a U-statistic theorem to represent the pth-quantile average difference, $\widehat{T}_N^p$ in the format of a sample average plus higher order remainders and then the standard CLT applies. All the follow-up theorems and corollaries rely critically on Theorem 5.1.

Step 1: To be specific, consider the pth-quantile sample average difference

$$\widehat{T}_N^p = \underbrace{N^{-1}\sum_{i=1}^{N} t_i^p \widehat{D}(X_i)}_{I_1^p} + \underbrace{N^{-1}\sum_{i=1}^{N} (\widehat{t}_i^p - t_i^p)(\widehat{D}(X_i) - D(X_i))}_{I_2^p} + \underbrace{N^{-1}\sum_{i=1}^{N} (\widehat{t}_i^p - t_i^p) D(X_i)}_{I_3^p}$$

where $\widehat{D}(\cdot) = \widehat{g}(\cdot) - \widehat{a}(\cdot)$ and t_i^p and $\widehat{t}_i^p$ are define in Eq. (4.9) and Eq. (4.14). Note that $D(X_i) = 0$, for any X_i under $\mathbb{H}_0$, so $I_3^p = 0$ is trivial. As for I_2^p, Lemma 1 in Appendix B shows it is equal to $o_p(N^{-1/2})$. Therefore, one only need to deal with I_1^p.

Step 2: Now I further decompose I_1^p into three components by first adding $a(X_i)$ and subtracting $g(X_i)$ without changing its value as $g(X_i) = a(X_i)$ under $\mathbb{H}_0$.

$$I_1^p = N^{-1}\sum_{i=1}^{N} t_i^p [(\widehat{g}(X_i) - g(X_i)) - (\widehat{a}(X_i) - a(X_i))]$$

Recall $\widehat{a}(\cdot)$ in Eq. (4.6) suffers from the problem of "generated" variables $\widehat{h}(\cdot)$. Therefore I replace $\widehat{a}(\cdot)$ with its infeasible counterpart $\widetilde{a}(\cdot)$ which assumes the knowledge of $h(\cdot)$,

$$\widehat{a}(X_i) = \widetilde{a}(X_i) - \widehat{E}(\Delta_i | X_i),$$

where $\Delta_i \equiv \widehat{h}(V_i) - h(V_i)$ and $\widehat{E}(\Delta_i|X_i)$ is the leave-one-out conditional mean estimator of

Δ given X_i with reduced bias.[8] Substituting this expression into I_1^p and using results from step 1, it would suffice to work with $\widetilde{T}_N^p$ since $\widehat{T}_N^p = \widetilde{T}_N^p + o_p(N^{-1/2})$, with $\widetilde{T}_N$ defined in Eq. (5.3)

$$\widetilde{T}_N^p \equiv D_N^g + D_N^a + D_N^h \tag{5.3}$$

where by definition

$$D_N^g = N^{-1}\sum_{i=1}^N t_i^p(\widehat{g}(X_i) - g(X_i)) \tag{5.4}$$

$$D_N^a = -N^{-1}\sum_{i=1}^N t_i^p(\widetilde{a}(X_i) - a(X_i)) \tag{5.5}$$

$$D_N^h = N^{-1}\sum_{i=1}^N t_i^p\widehat{E}(\Delta(V_i)|X_i) \tag{5.6}$$

Step 3: By the U-statistic theorems of various orders, D_N^g, D_N^a and D_N^h can be represented as serveral sample means, characterized by Lemma 2, 3 and 5, respectively.[9] Therefore, we can rewrite $\widehat{T}_N^p$ as an influence function plus those asymptotically negligible terms at $\sqrt{N}$-rate, like the following.

$$\sqrt{N}\widehat{T}_N^p = N^{-1/2}\sum_{i=1}^N \xi_i^p + o_p(1), \quad \forall p = 1, \cdots, P$$

Then the standard CLT applies to the sample average while the remainder vanishes in the limit. Intuitively, the variance of pth quantile average difference would come from the variation of estimation of $g(\cdot)$, variation of estimation of $a(\cdot)$ as well as the estimation of the unknown function $h(\cdot)$. To have a cleaner expression, I apply the recursive biad-reducing estimator to ensure that the asymptotic biases vanish faster than $\sqrt{N}$ so that the vector of quantile mean differences will recenter at 0.

[8] For the sake of brevity, the subscript $-i$ is suppressed.

[9] One technical simplification is to replace the estimated density denominator with the truth, guaranteed by the intermediate lemma A2 in the supplemental materials.

Theorem 5.2 below combines the information of the vector $\widehat{T}_N$ which, under the null, follows the asymptotic multivariate normal distribution with a positive definite diagonal covariance matrix. The final test statistic then converges asymptotically to the χ^2_P distribution with the degree of freedom equal to the predetermined number of quantile regions, P. In Corollary 5.2.1, the asymptotic null distribution of the feasible test statistic is given by simply plugging-in some consistent covariance estimator of Ω.

Theorem 5.2 The infeasible test statistic $\widehat{W}^0_N$**.** *Suppose that Assumption I.1-I.2 and A.1-A.6 hold, under* $\mathbb{H}_0$*, it follows that*

$$\widehat{W}^0_N \xrightarrow{\text{D}} \chi^2_P$$

where $\widehat{W}^0_N = N\widehat{T}'_N\Omega^{-1}\widehat{T}_N$*, with* $\widehat{T}_N$ *in Eq.* (4.12) *and* Ω *in Eq.* (5.7)

$$\Omega = E(\xi_i\xi_i') \tag{5.7}$$

where $\xi_i \equiv (\xi_i^1, \xi_i^2, \cdots, \xi_i^P)'$.

Corollary 5.2.1 The feasible test statistic $\widehat{W}_N$**.** *Suppose that Assumption I.1-I.2 and A.1-A.6 hold, under* $\mathbb{H}_0$*, it follows that*

$$\widehat{W}_N \xrightarrow{\text{D}} \chi^2_P$$

where $\widehat{W}_N = N\widehat{T}'_N\widehat{\Omega}_N^{-1}\widehat{T}_N$*, with* $\widehat{T}_N$ *in Eq.* (4.12) *and* $\widehat{\Omega}_N$ *in Eq.* (5.8).

The consistent estimator of covariance matrix $\widehat{\Omega}_N$ in Theorem 5.1 is therefore obtained in the following way,

$$\widehat{\Omega}_N = N^{-1}\sum_{i=1}^{N}\widehat{\xi}_i\widehat{\xi}_i' \tag{5.8}$$

where $\widehat{\xi}_i \equiv (\widehat{\xi}_i^1, \widehat{\xi}_i^2, \cdots, \widehat{\xi}_i^P)'$. To be specific, $\widehat{\xi}_i^p$ is obtained by substituting with the consistent

estimators for unknown functions and densities for each p.

$$\widehat{\xi}_i^p \equiv \{[\widehat{t}_i^p + \widehat{E}(t^p|V_i)]\widehat{f}^*(X_i,V_i) - t_i^p]\}\hat{\epsilon}_i + \overline{t}^p\widehat{h}(V_i) \tag{5.9}$$

where $\overline{t}^p \equiv N^{-1}\sum_{i=1}^N \widehat{t}_i$ and

$$\begin{aligned}\widehat{\epsilon}_i &= Y_i - \widehat{a}(X_i) - \widehat{h}(V_i) \\ \widehat{f}^*(X_i,V_i) &= \widehat{f}_X(X_i)\widehat{f}_V(V_i)/\widehat{f}_U(U_i).\end{aligned}$$

6 Finite Sample Results

In this section, I obtain the power and size results using simulations under various data generating processes (DGPs). In DGP 1, the simple additive model is tested against nonseparable models with polynomials. In DGP 2, it is against much general nonseparable functional forms while having the same null hypothesis as in DGP 1. In DGP 3, I allow for multi-dimensional unobservables, featured in this paper.

In each DGP, the number of quantiles is allowed to vary as the choice of P is known to affect the asymptotic local power functions but is empirically unclear. Therefore, I experiment with different values such as $P_N = \{4, 6, 8\}$, in order to check the robustness of the results with respect to this parameter. I also introduce a "nonseparability" measure δ. When $\delta = 0$, the model is purely additive. It becomes, in a sense, more nonseparable as δ increases. The varying δ corresponds to the rate at which a series of local alternatives converge to the null. The rule-of-thumb bandwidth of Silverman, i.e. $h = 1.06 \times s.e.(U) \times N^{-r}$, has been implemented for practical purposes. Furthermore, I trim on U with trimming parameters $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$. As aforementioned, I use observations in the range $(\kappa_1, 1-\kappa_1)$ to control for boundary biases when recursively estimating the nonparametric conditional expectations and those in narrower range $(\kappa_2, 1-\kappa_2)$ to construct the test statistic. I consider a moderate number of replications, $N_{mc} = 250$ to make computational time manageable and

I have tried sample size N at both 250 and 500 for each DGP.

6.1 DGP 1

The data is generated from the following DGP,

$$\mathbb{H}_1^1 : Y = (X + \varepsilon) + \delta (X\varepsilon)^2 \tag{6.1}$$

where δ represents the level of nonseparability and if $\delta = 0$, the model becomes completely additive. DGP 1 models the nonseparability arising from the interactions of X and ε as follows,

$$X = \frac{1}{4} + V - \frac{1}{4}V^2 + u_2, \quad \varepsilon = \frac{1}{2}V + u_1$$

where V, u_1 and u_2 are generated independently from the uniform distribution, $U[0, 1]$.

Table 1 displays the results of empirical size studies under the null $\mathbb{H}_0$, which sets $\delta = 0$ under various number of observations, e.g. $N = 250$ or 500, and smoothing options, r_1. Table 1 also presents the power analysis under $\mathbb{H}_1$ of DGP 1, as the nonseparability parameter δ varies. Column 4-6 of Table 1 display the results of the empirical sizes while powers are in column 7-12. The first three columns give empirical size results coming from small sample sizes. The test statistics are likely to be undersized but such phenomena are mitigated when sample size is increased to 500. The test statistic almost captures the correct sizes and I expect these minor discrepancies would go away as the number of replications increases. Next turn to the power analysis. When there is a little nonseparable portion, like $\delta = 0.5$, the rejection rates are uniformly below 50% for small sizes. When N doubles, one observes that powers increase by around 0.3 for each design. On the other hand, as nonseparability strengthens to $\delta = 1$, one rejects the null hypothesis over 90% of times on average. Then take a look at another tuning parameter, P over $\{4, 6, 8\}$. And it is evident that the power results are somewhat robust to the choice of number of quantiles. As P varies, the rejection rates

are relatively stable. To sum up, even under small samples, the empirical sizes produced by our test statistics look very close to what theory predicts under the null. Whereas under the scenario of $\mathbb{H}_1$, tests with analytic variances could deliver reasonable powers, but may depend on the nature of nonseparability in the DGP.

6.2 DGP 2

DGP 2 considers more general form of nonlinearity other than the polynomials, specified in (6.2).

$$\mathbb{H}_1^2 : Y = X + \varepsilon + \delta \frac{\exp(2X)}{2 + \sin(\varepsilon)} \tag{6.2}$$

where X and ε are generated in the same way as in DGP 1. Likewise, δ measures the nonseparability and in the following simulation experiments, I let δ vary from $\{0.1, 0.25\}$. When $\delta = 0$, the model goes back to the $\mathbb{H}_0$ of DGP 1, so only the alternatives need to be studied. When $\delta = 0.1$, one can check the performance of the test to the extent that only very weak nonseparability presents.

The small sample power results are presented in Table 2, it is not hard to see that the rejection rates depend critically on how separable the DGP is. As with $\delta = 0.1$, rejection rates are generally low. In contrast, as the nonseparable part gains more weight, e.g. $\delta = 0.25$, the rejection rates become reasonably large, exhibiting good power. This simulation shows that the magnitude of nonseparability can make a difference for the power of the test for very small sample sizes.

6.3 DGP 3

DGP 3 incorporates multiple unobservables as featured in the test. For simplicity, assume the true model is like (6.3),

$$\mathbb{H}_0^3 : Y = X\eta + \varepsilon \tag{6.3}$$

where $\eta \sim U[0.5, 1, 5]$ and (X, V, ε) are generated in the same way as DGP 1. To analyze the power property, a nonseparable portion is incorporated such as (6.4), denoted by DGP 3.1.

$$\mathbb{H}_1^3 : Y = X\eta + \varepsilon + \delta \exp(X\varepsilon) \tag{6.4}$$

The empirical size and power results are in Table 3. $\mathbb{H}_0^3$ is one of the examples where the condition in Proposition 3.3 holds. It is a nonseparable structural function with more than one unobservable. For $\delta = 0$, it is true that the test has no power against structural separability. From $\mathbb{H}_0^3$, it indicates that the ASF generated is equivalent to that from some additive model, though the true DGP is a nonseparable model. The test provides enough confidence for us to compute the ASF as if the true model is additive separable and obtain a more efficient estimator of ASF. $\mathbb{H}_1^3$ showcases the situation when more nonseparable forms are being added, the proposed test becomes much powerful against both $\mathbb{H}_0$ (equality of ASFs) and $\mathbb{H}_0^*$ (additive separability). Therefore, inconsistent estimates of ASFs would be likely to be produced unless the nonseparable nature is properly taken into account.

Turn to the empirical power analysis. The setup of this experiment copies that of DGP 1 where I vary the nonseparability parameter δ from 0.5 to 1. Now I summarize key results over the following three dimensions. First, as sample size increases from 250 to 500, the rejection rates have increased, yielding reasonable powers. Second, the powers do not change much across selected number of quantiles. This property gives me more confidence on the

robustness of test results with respect to this tuning parameter. Third, doubling the weight of the nonseparable component, on average, increase rejection probabilities by 20% or so.

Table 1: Empirical Size and Power Results of DGP 1

		$\delta = 0$			$\delta = 0.5$			$\delta = 1$		
N	P_N	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
250	4	0.000	0.006	0.024	0.204	0.368	0.468	0.524	0.652	0.732
250	6	0.000	0.016	0.024	0.212	0.344	0.440	0.508	0.632	0.712
250	8	0.012	0.056	0.092	0.208	0.324	0.396	0.504	0.604	0.664
500	4	0.012	0.040	0.074	0.620	0.772	0.820	0.908	0.948	0.968
500	6	0.012	0.028	0.036	0.624	0.760	0.812	0.908	0.940	0.960
500	8	0.012	0.028	0.040	0.612	0.760	0.800	0.900	0.940	0.948

Note: Number of replications $N_{mc} = 250$. Smoothing parameters $r_1 = 1/7.9, r_2 = 1/7.9$. Trimming parameters $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$.

Table 2: Empirical Power Results under of DGP 2

N	δ	P_N	0.01	0.05	0.1
250	0.1	4	0.104	0.232	0.372
250	0.1	6	0.144	0.252	0.344
250	0.1	8	0.144	0.252	0.336
250	0.25	4	0.912	0.992	1.000
250	0.25	6	0.924	0.992	1.000
250	0.25	8	0.928	0.992	0.996

Note: Number of replications $N_{mc} = 250$. Smoothing parameters $r_1 = 1/7.9, r_2 = 1/7.9$. Trimming parameters $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$.

Table 3: Empirical Power Results of DGP 3

		$\delta = 0$			$\delta = 0.5$			$\delta = 1$		
N	P_N	0.010	0.050	0.100	0.010	0.050	0.100	0.010	0.050	0.100
250	4	0.065	0.160	0.215	0.400	0.540	0.616	0.615	0.785	0.810
250	6	0.050	0.110	0.165	0.340	0.470	0.570	0.550	0.720	0.760
250	8	0.035	0.070	0.105	0.270	0.415	0.460	0.480	0.645	0.700
500	4	0.070	0.155	0.230	0.665	0.770	0.805	0.825	0.930	0.955
500	6	0.035	0.100	0.135	0.585	0.710	0.765	0.800	0.885	0.940
500	8	0.020	0.060	0.105	0.500	0.655	0.690	0.725	0.830	0.885

Note: Number of replications $N_{mc} = 250$. Smoothing parameters $r_1 = 1/7.9, r_2 = 1/7.9$. Trimming parameters $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$.

7 Extensions

In this section, two extensions commonly encountered in empirical settings are presented. Firstly, despite all attractive properties that nonparametric testing entails, in practice, semiparametric models are often invoked due to the high dimensional nature of conditioning variables. I therefore coax the test into a two-stage semiparametric single-index framework with the finite-dimensional parameters estimated firstly by weighted semiparametric least square (WSLS) by Ichimura and Lee (1991) and Ichimura (1993). In the second extension, I consider the nonparametric nonseparable triangular simultaneous equations models. Following Imbens and Newey (2009), the marginal distribution of the first stage disturbance suffices to serve as a control variable. Fortunately, the asymptotic properties of the test statistic are robust to the problem of "generated" regressors.

7.1 Semiparametric Test

When the dimension of X (or V) is large like in many real settings, dimension reduction techniques are necessary. Following the semiparametric literature, I assume the multi-dimensional covariates to comply with a linear index structure, e.g. $I_0 \equiv X'\beta_0$, where β_0 is a conformable vector of finite-dimensional true parameters. Such models have been well studied (Powell et al., 1989; Ichimura, 1993; Ai and Chen, 2003; Klein and Spady, 1993; Klein and Shen, 2015, etc).

Now redefine the model (2.1) as the semiparametric single index nonseparable model in Eq. (7.1).

$$Y = m(X'\beta_0, \varepsilon) \tag{7.1}$$

where it is assumed that there exists at least one continuous variable in X for identification purpose. From now on, I modify the hypotheses of interest by incorporating the

semiparametric structure.

$$\mathbb{H}_0 : g(x'\beta_0) = a(x'\beta_0), \text{ a.s., for each } x \in \mathcal{X}; \quad \mathbb{H}_1 : \mathbb{H}_0 \text{ is not true}$$

where the ASFs become the following in Eq. (7.2) and Eq. (7.3), respectively.

$$g(x'\beta_0) = \int m(x'\beta_0, e)dF_\varepsilon(e) = \int E(Y|x'\beta_0, v)dF_V(v) \tag{7.2}$$

$$a(x'\beta_0) = E(Y - h(V)|x'\beta_0), \text{where } h(v) = \int_{\mathbb{R}} E(Y|x'\beta_0, v)dF_{X'\beta_0}(x'\beta_0) - E(Y) \tag{7.3}$$

To conduct the semiparametric inference, one can apply a two-stage procedure. In the first stage, a consistent estimator of β is obtained by employing the multiple-index WSLS in Ichimura and Lee (1991). Next, replace the true single index $I_0 = X'\beta_0$ with $\widehat{I} = X'\widehat{\beta}$ and then follow the exact procedure outlined in Section 4.

It has been well recognized that β_0 is only identified up to location and scale. Common normalizations include setting $\beta_{10} = 1$, where β_{10} is the coefficient associated with any continuous variable or $||\beta_0|| = 1$, where $||\cdot||$ is the Euclidean norm.[10] Note that when X and ε are not correlated of any sort, model (7.1) can be rewritten as the semiparametric single index regression with an additive error like Ichimura (1993). To see this, $E(Y|X) = E[m(X'\beta_0, \varepsilon)|X] = m_1(X'\beta_0)$, implying $Y = m_1(X'\beta_0) + U$, where $E(U|X) = 0$. In the presence of endogenous regressors, with the imposition of single index Assumption S-1 in conjunction, one can work with the conditional mean representation such as (7.4).

$$Y = E(Y|X'\beta_0, V) + \epsilon \tag{7.4}$$

where $E(\epsilon|X'\beta_0, V) = 0$ by construction. Assumption S.1 only assumes that the equality

[10] General parametric forms of indexes, e.g. $I(X, \beta_0)$, are allowed but identification of β_0 has to be conducted on a case-by-case basis.

$E(Y|X,V) = E(Y|X'\beta_0, V)$ holds at the true parameter values.

Assumption-S.1 Index identification. There is a unique interior point $\beta_0 \in \mathcal{B}$ such that

$$E(Y|X,V) = E(Y|X'\beta_0, V), \quad \text{a.s.}$$

Consistent estimation of β_0 can be made possible by WSLS by Ichimura and Lee (1991).[11] $\widehat{\beta}$ can be obtained by minimizing the sum of squares of residuals weighted by

$$\widehat{\beta} = \arg\min_{\beta \in \mathcal{B}} N^{-1} \sum_{i=1}^{N} \widehat{W}_i(\widehat{\beta}^0)[Y_i - \widehat{E}(Y|X_i'\beta, V_i)]^2$$

where the bias-reducing conditional expectation estimator is defined in Eq. (4.2). $\widehat{W}_i(\widehat{\beta}^0) = 1/\widehat{E}(\hat{\epsilon}^2|X_i'\widehat{\beta}^0, V_i)$ with $\widehat{\beta}^0$ being a preliminary consistent estimator, such as unweighted SLS estimator and $\hat{\epsilon}$ is the corresponding residual estimator. As additive semiparametric models (7.4) are nested by nonseparable models (7.1), $\widehat{\beta}$ is also consistent even if the true model is additive separable. $\sqrt{N}$-consistency are immediately established in Proposition 7.1 which is a direct implication of the theorem in Ichimura and Lee (1991). See Ichimura (1993) and Klein and Shen (2015) for details.

Proposition 7.1 Consistency of $\widehat{\beta}$**.** *Under Assumption I.1 and S.1, then it follows that*

$$|\widehat{\beta} - \beta_0| = o_p(N^{-1/2})$$

To apply the EQM test statistic, one can replace β_0 with $\widehat{\beta}$ and restrict X as an estimated single index. Fortunately, the semiparametric covariance estimator, $\widehat{\Omega}_N(\widehat{\beta})$ takes exactly the same form as the nonparametric counterpart. This is because the first-stage estimation variability of β_0 has no impact on the second stage empirical mean differences of ASFs. To see the intuition, recall the difference estimator $\widehat{D}(x'\widehat{\beta}) = \widehat{g}(x'\widehat{\beta}) - \widehat{a}(x'\widehat{\beta})$ at any $x \in \mathcal{X}$ and

[11] Model (7.1) coincides with the generalized regression model in Han (1987) if strict monotonicity of m in $X'\beta_0$ is assumed. Han estimate β_0 by maximum rank correlation.

by the Delta method around β_0, assuming differentiability, e.g. $\widehat{g}'(\cdot)$ and $\widehat{a}'(\cdot)$.

$$\widehat{g}(x'\widehat{\beta}) - \widehat{a}(x'\widehat{\beta}) = [\widehat{D}(x'\beta_0)] + [\widehat{g}'(x'\beta_0) - \widehat{a}'(x'\beta_0)]x'(\widehat{\beta} - \beta_0) + o_p(N^{-1/2})$$

The second term is also $o_p(N^{-1/2})$ as $|\widehat{g}'(x) - \widehat{a}'(x)| \to 0$ and $\sqrt{N}(\widehat{\beta} - \beta_0) = o_p(1)$. As a consequence, Theorem 5.1, Theorem 5.2 and Corollary 5.2.1 would immediately apply by substituting the single index estimators, $\widehat{\beta}$. The formal result is stated in Theorem 7.1.

Theorem 7.1 Asymptotic null distribution**.** *Under* $\mathbb{H}_0$ *and Assumption S.1, I.1-I.2 and A.1-A.6, then* $\widehat{W}_N(\widehat{\beta}) \xrightarrow{\text{D}} \widehat{W}_N(\beta_0)$.

Remark 1. Efficient estimation of β_0. Often times, the finite dimensional parameters are of estimation and testing interest in its own right. For example, β_0 might measure the relative importance of regressors and their substitution patterns. Therefore, hypotheses of economic interest can be directly formulated upon β_0. Not only is our separability test informative on the consistency of estimators, but also it can shed light on the efficiency. In the case of not rejecting $\mathbb{H}_0'$, one can utilize such additional information by solving a nested minimization problem below,

$$\widehat{\beta}^a = \underset{\beta \in \mathcal{B}^0}{\arg\min}\, N^{-1} \sum_{i=1}^{N} \widehat{W}_i(\beta)[Y_i - \widehat{h}(V_i) - \widehat{E}(Y - \widehat{h}(V)|X_i'\beta)]^2$$

where $\widehat{h}(v)$ is the consistent estimator of $h(v)$ based on $\widehat{\beta}$. $\widehat{W}_i$ is the optimal weight estimator. Iteratively, updating with $\widehat{\beta}^a$ would give a more efficient ASF estimator in $\widehat{a}^e(X_i'\widehat{\beta}^a) \equiv \widehat{E}(Y - \widehat{h}^a(V)|X_i'\widehat{\beta}^a)$, where $\widehat{h}^a(v)$ is estimated using the more efficient estimator $\widehat{\beta}^a$.

Remark 2. A more powerful test. When the first-stage finite parameters are present, it opens up possibilities to increase the power of our ASF test. For instance, one can incorporate the information of β_0 into a new set of joint hypotheses as below,

$$\mathbb{H}_0 : \beta_a = \beta_0, g(x'\beta_0) = a(x'\beta_0), \;\; \text{a.s., for each } x \in \mathcal{X}; \;\; \mathbb{H}_1 : \mathbb{H}_0 \text{ is not true}$$

where β_0 and β_a are unique solutions to the conditional mean restrictions, respectively.

$$E[Y - E(Y|X'\beta_0, V)|X, V] = 0; E[Y - h(V) - E(Y - h(V)|X'\beta_a)|X, V] = 0$$

Generally speaking, $\beta_a = \beta_0$ may not necessarily hold unless $m(\cdot, \cdot)$ is additive. Natural estimators of β_0 and β_a are their respective WSLS estimators aforementioned, $\widehat{\beta}$ and $\widehat{\beta}^a$. The derivation of asymptotic null distribution need to take into account the correlation between finite and infinite-dimensional estimators and is left for future research.

7.2 Triangular Simultaneous Equations Models

In many cases, control variables V are not directly observable. However, there may be auxiliary equations or excluded instrumental variables Z that be can exploited for external variation. Suppose that the endogenous regressors are determined by first stage equations given in (7.5)

$$X_k = h_k(Z, \eta_k), \quad \forall k = 1, \cdots, d_X \tag{7.5}$$

where $h_k(\cdot)$ is an unknown measurable function and Z is a vector of instrumental variables subject to the exogeneity condition in Assumption T.1. Let $\eta \equiv (\eta_1, \cdots, \eta_{d_X})'$.

Assumption-T.1 Exogeneity. $Z \perp (\varepsilon, \eta)$.

Assumption-T.2 Scalar monotonicity. For each $z \in \mathcal{Z}$, $h_k(z, \cdot)$ is a strictly monotonic function, for $k = 1, 2, \cdots, d_X$.

Assumption T.1 requires the full independence between instrumental variables and unobservables. Assumption T.2 is more substantive than it looks, despite its popularity in nonseparable literature. First, it restricts the dimension of unobservables η to be unit. Then it imposes shape restrictions in terms of monotonicity. Furthermore, it rules out discrete endogenous variables. Imbens and Newey (2009) consider the above assumptions

in nonparametric nonseparable triangular simultaneous equations models and prove the following proposition.

Proposition 7.2 Theorem 1, Imbens and Newey (2009)**.** *Under Assumption T.1 and T.2,* $X \perp \varepsilon | V$*, where* $V = [F_{X_1|Z}(X_1, Z), \cdots, F_{X_{d_X}|Z}(X_{d_X}, Z)] = [F_{\eta_1}(\eta_1), \cdots, F_{\eta_{d_X}}(\eta_{d_X})]$.

Proposition 7.2 is an existing result in the nonparametric identification literature, so I would not reiterate the proof here. If one knows the true conditional distribution of X given Z, nothing would change in the testing procedure and one can simply replace V with the "generated" control variable. In a nonparametric situation, an additional step is needed to estimate $F_{X|Z}(X, Z)$ first by the recursive conditional expectation estimator defined in Eq. (7.6)

$$\widehat{V}_k = \widehat{F}_{X_k|Z}(x, z) = \frac{\sum_{i=1}^N K_h(Z_i - z)]\{\mathbf{1}[X_{ki} \leq x] - \widehat{\delta}_i(z)\}}{\sum_{i=1}^N K_h(Z_i - z)}, \, k = 1, 2, \cdots, d_X \tag{7.6}$$

Fortunately, asymptotic results of the test statistic are not influenced by the first stage estimation. Theorem 7.2 gives formal results on the asymptotic null distribution and it basically states that it is permitted to use the true V in place of $\widehat{V}$ regardless of the "generated" regressors. This theorem is based on the result from Mammen et al. (2012) who study nonparametric regression with nonparametrically "generated" covariates. In the example of estimating ASFs in the nonparametric nonseparable triangular simultaneous equations models, they establish that the limiting variances are not affected when $\widehat{V} = \widehat{F}_{X|Z}(X|Z)$ need to be estimated in the first stage, under very mild conditions. Let $\widehat{W}_N(\widehat{V})$ denote the test statistic in Eq. (4.11) with all V replaced by $\widehat{V}$ and $\widehat{\Omega}(\widehat{V})$ obtained in the similar fashion.

Theorem 7.2 Asymptotic null distribution**.** *Under* $\mathbb{H}_0$ *and Assumption T.1-T.2, I.1-I.2 and A.1-A.6, then* $\widehat{W}_N(\widehat{V}) \xrightarrow{\text{D}} \widehat{W}_N$.

The proof of Theorem 7.2 exploits the empirical processes arguments. Supporting lemmas can be found in Mammen et al. (2012).

8 An Application: Testing Hicks-neutral Productivity

Understanding how inputs are related to outputs is a fundamental issue in empirical industrial organization and other fields of economics, see (Ackerberg et al., 2015). In empirical trade and macroeconomics, researchers are often interested in estimating production functions to obtain a measure of total factor productivity, to examine the impact of trade policy and FDI, and to analyze the role of resource allocation on aggregate productivity. In empirical IO and public economics, firm-level production functions are usually estimated to evaluate the effect of deregulation and industrial policies, and to predict market outcomes of mergers and R&D, etc. The concept of Hicks neutrality was first seen in 1932 by John Hicks in the book *The Theory of Wages*. In essence, a change is considered to be Hicks-neutral if it does not affect the balance of labor and capital in a production function. When it comes to firm-level production functions, Hicks-neutrality unavoidably puts substantial restrictions on how unobserved firm-level heterogeneity in productivity is modeled.

The Hicks-neutral functional form has at least three implications as below. One of the restrictions dictates that the input substitution pattern be free of unobserved technological shocks across firms, implying that output-input elasticities are identical for firms that use the same amount of inputs. Second, the wrongly imposed Hicks-neutrality would make most of the commonly employed identification strategies invalid, leading to inconsistent structural parameter estimates from an econometric perspective. Those identification methods include, but are not limited to, IV approaches using input prices (Griliches and Mairesse, 1995), dynamic panel (Blundell and Bond, 2000), "proxy" variables approach (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015) and so on. Last but not least, once the Hicks-neutrality does not hold, the interpretation of the "Solow" residual as a measure of total factor productivity would be incorrect or even misleading. Motivated by the above facts, an empirical test of Hicks-neutrality is necessary. Since such a test can be converted into testing additive separability between input choices and firm heterogeneity, the

nonparametric test proposed in this paper is directly applicable due to its flexible treatment of functional forms.

To be concrete, if a firm's production in year t is Hicks-neutral, then it is true that its production function is multiplicative separable or log additive separable as below[12]

$$Y = F_t(K, L, \omega, \varepsilon) = F_t^1(K, L)A_t(\omega, \varepsilon) \text{ or } y = f_t^1(K, L) + a_t(\omega, \varepsilon) \tag{8.1}$$

where F_t is the unknown measurable firm production function at year t and it is usually assumed that both F_t^1 and A_t map into strictly positive sets in this context. Following the notational convention of the production function literature, Y denotes the value-added output, K as capital amount, L as labor input and ω as productivity and ε as idiosyncratic shocks or measurement error in output. (ω, ε) are unobservables with ω being supposed to correlate with variable inputs such as labor whereas ε is usually seen as independent of observable inputs.

To control for the endogeneity, I extend the "proxy" variable approach to fully nonparametric nonseparable settings, by assuming the availability of intermediate material input M like in Levinsohn and Petrin (2003) and the functional form, timing and support assumptions needed. Assumption F.1 is the functional form assumption. Almost no shape restrictions or distributional assumptions are imposed except for the scalar value of ω_t. However, even this restriction can be relaxed, once multiple intermediate inputs are observed. Assumption F.2 reiterates the static nature of the current model and highlights the source of endogeneity.[13] Assumption F.3 is standard in the proxy variable literature.[14]

Assumption F.1 Functional forms. $Y_t = F_t(K_t, L_t, \omega_t, \varepsilon_t)$, where $\omega_t \in \mathbb{R}, \varepsilon_t \in \mathbb{R}^\infty$.

[12]Lower letters denote variables (or functions) after log transformation.

[13]This amounts to treating capital as predetermined and unrelated with the contemporaneous productivity shock. For example, capital could be accumulated deterministically according to $K_t = k(I_{t-1}, K_{t-1})$.

[14]Admittedly, scalar monotonicity unobservable in the intermediate demand function can be substantive in situations where local demand conditions, market power, input quality/price differentials and measurement errors might matter for the choice of intermediate inputs. Huang and Hu (2011); Kim et al. (2013) have considered cases when capital is measured with errors.

Assumption F.2 Timing and shocks. K_t is fixed input, $K_t \perp \omega_t$; L_t is flexible input, $L_t \not\perp \omega_t$; $(K_t, L_t) \perp \varepsilon_t$.

Assumption F.3 Intermediate demand. There exists an unknown function $M_t = \mathbb{M}_t(K_t, \omega_t)$ where $\mathbb{M}_t(k, \cdot) : \mathbb{R} \to \mathbb{R}^+$ is strictly increasing for each $k \in \mathcal{K}$.

Assumption F.4 Large support. $\text{supp}(V_t) = \text{supp}(V_t|k, l) = [0, 1]$, for each $(l, k) \in \mathcal{L} \times \mathcal{K}$ and t.

Proposition 8.1 Conditional Independence**.** *Under Assumption F.1-F.4, then* $(K_t, L_t) \perp (\omega_t, \varepsilon_t)|V_t$ *and* $V_t = F_t(M_t|K_t)$ *and* $F_t(\cdot|\cdot)$ *is the conditional distribution of* M_t *given* K_t *for each period* t*.*

Proposition 8.1 states that the conditional distribution of the intermediate input M_t given capital level K_t, i.e. $V_t = F_{M|K,t}(M|K)$ is able to serve as the control variate. The proof of Proposition 8.1, given in Appendix A, resembles that of Proposition 1 in Imbens and Newey (2009) who consider general nonparametric triangular simultaneous equations. The "generated" control variable can be estimated by any nonparametric estimators in principle. Here I consider a bias-reduced local constant estimator.

This paper employs an unbalanced panel of 5,088 manufacturing firms, 40,560 total observations, from Compustat North America fundamental annual database during the period 1990-2011. I also supplement it with deflators and industry-level depreciation rates from Becker et al. (2013), available from NBER website and industry-level annual average wages from Quarterly Census of Employment Wage (QCEW) collected by BLS. The value-added output Y is obtained by subtracting material cost, to be defined later, from net sales deflated by industry-level price index for shipments.[15] Capital input, K, is computed using a Perpetual Inventory Method (PIM), i.e. $K_{t+1} = (1 - \delta)K_t + I_t$. The initial capital, K_0 is the value of property, plant and equipments deflated by the new investments price index. I

[15]Compustat database only provides sales information instead of quantity. Using sales to represent output may introduce output price bias. To remedy it, one may supplement demand system as in De Loecker (2011), but significant parametrization cannot be avoided.

is the capital expenditures deflated by the new investments price index; δ is the depreciation rate for assets, which is backed out by the PIM from Becker et al. (2013). Following Olley and Pakes's method, I use the lagged investment when computing capital input.[16] Labor input is taken as number of workers per firm. For material input, it is equal to the costs of goods sold plus administrative and selling expenses minus depreciation and wages, then deflated by its corresponding deflator.[17] Table 4 provides some descriptive statistics of the whole industry and five selected sectors. For output and input variables, each cell reports the average value. The average length of firm appearance in our sample is around 12 years, reflecting the unbalanced nature of the panel data. To single out the testing problem, I consider the static model year by year, in order to separate out the sample selection and dynamic issues.

Table 4: Some Descriptive Statistics of Selected Sectors

NAICS-3	Name	No. Obs.	Avg. Year	Y	K	L	M
All	Manufacturing	40,560	12.91	1462.55	1676.76	7.16	1318.63
311	Food product	1,822	13.88	848.87	1074.02	12.16	1527.99
325	Chemical	4,965	12.79	1044.30	1852.64	8.02	1162.67
332	Fabricated Metal	1,822	13.99	311.98	453.81	4.52	450.64
333	Machinery	4,119	13.74	449.94	667.18	5.46	757.34
336	Transportation	2,330	13.86	2723.54	4935.40	23.14	4973.30

Note: 1. All manufacturing industry encompasses 21 sectors with NAICS code 31-33. 2. Avg. Year is the average number of years of presence in the sample period. 3. Y, K and M are measured in thousand dollars and L is measured in thousand units.

Before jumping to the main testing results, I perform a preliminary test because I want to confirm the power of the proposed testing procedure in an obvious empirical content as follows.

$$\mathbb{H}_0 \quad : \quad F_t(L, K, \omega, \varepsilon) = F_t^1(L, K) + A_t(\omega, \varepsilon), \quad \text{a.s.}$$

$$\mathbb{H}_1 \quad : \quad \text{Otherwise}$$

[16] Only the deflator for new capital expenditures (investment flows) is available, rather than that for capital stock.

[17] Wages are computed as the multiplication of total employment and industry-level average annual total compensation.

The null hypothesis says that the production function is additive. This is clearly a false statement as many theoretical and empirical works have proved. Table 5 provides empirical testing results in terms of test statistics and p-values. The results of the preliminary test are presented in the specification (1). To better summarize the findings, I also visualize the results in the following figures. From Figure 1, one rejects the null hypotheses of additive production functions in all years with 1% significant level as it should be. The test statistics are large in most years and the test has reasonablly good power. This further gives us confidence in applying the test in more interesting scenarios.

In Figure 2, I present the testing results for the Hicks-neutrality for each year for the specification (2) in Table 5. As shown before, it is equivalent to testing the additive separability of the log-transformed production function,

$$\begin{aligned} \mathbb{H}_0 &: f_t(L, K, \omega, \varepsilon) = f_t^1(L, K) + a_t(\omega, \varepsilon), \text{ a.s.} \\ \mathbb{H}_1 &: \text{Otherwise} \end{aligned}$$

The results show that non-Hicks neutral production happened during 1990-2002 and thereafter became Hicks-neutral until 2011. It is interesting that the rejection years correspond to a period of fast-growing of the manufacturing industries. Many empirical evidences have found that the most important driver of this growth is the mass adoption of computer technologies from 1993 to 1998. If I choose a higher significant level, then the tests would precisely capture those years where non-Hicks neutral production occurred. This finding is very intuitive as when firms adopt new technologies and innovate on production processes, this change is usually on the firm-level, rather than the whole industry. As firm are heterogeneous, there are "first-adopters" who begin reforms earlier and thus the impact of productivity shocks on their essential technologies can be very differently from their slower competitors. Thus, the differences in the speed of reforming (or adoption of new technology) are very likely to cause the differences in the "essential" technologies, even

within the same sector. After 2000, most of firms have finished this transformation so that their substitution patterns start to converge again, as evidenced by the non-rejections of Hicks-neutral technological shocks.

However, one can conjecture that the rejection of Hicks-neutrality might be due to the failure of controlling for sector heterogeneity. For an industry as large as manufacturing, it consists of various subsectors such as transportation, machinery, textiles product, etc. So it can be perceived that firms across sectors could employ totally different technologies or production functions. For example, it would make no sense to expect that a labor union strike to affect machinery and food product equally as their substitution pattens of labor and capital could be very different. When there is sufficient amount of observations in each sector, one might mitigate this problem by placing sector dummy variables to control for this disparity, i.e. $Y = F_t(L, K, S, \omega, \varepsilon)$, where S represents sector dummies or sector specific effect. However, one is facing with the notorious "curse of dimensionality" problem due to the high-dimensionality of sector-specific effects. In my sample, there are totally 21 sectors and for some sectors and as a result only a few observations are available in some years. Therefore, the variances of the test statistic could be extremely large and renders the test of almost no power. The limitation of data forces us to compromise on the full nonparametric sector-specific effects. To resolve this problem, we test with linear sector dummies, as commonly used in empirical works.

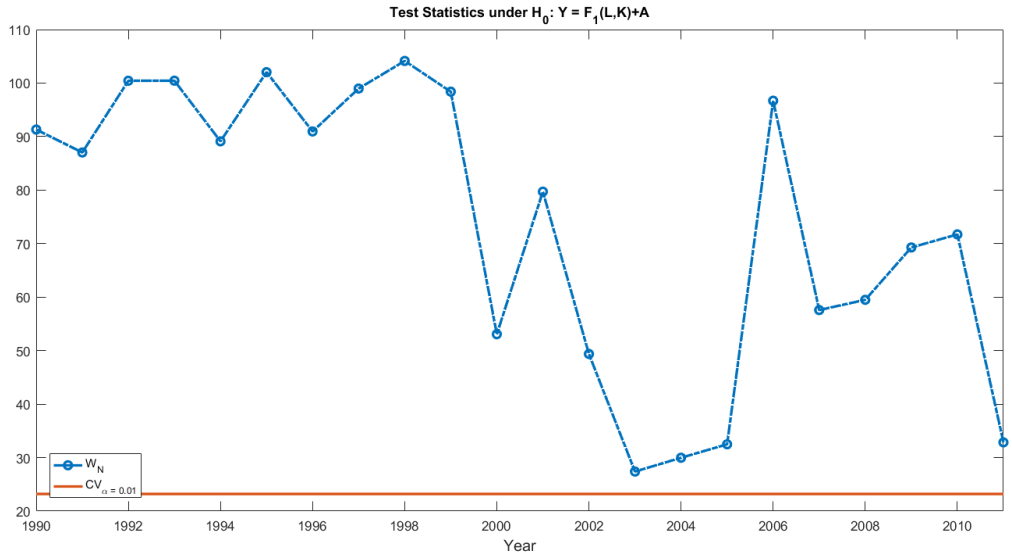
$$\begin{aligned} \mathbb{H}_0 &: f_t(L, K, \omega, \varepsilon) + S_t = f_t^1(L, K) + S_t + a_t(\omega, \varepsilon), \quad \text{a.s.} \\ \mathbb{H}_1 &: \text{Otherwise} \end{aligned}$$

The linearity of sector dummies implies that the sector-specific effects impact value-added output only through a multiplicative or scaled effect, rather than altering the functional form of the essential technology. The results are presented in the specification (3) in Table 5 and displayed in Figure 3. Now the rejection of Hicks-neutrality is more obvious in the 90s and

early 2000s and they are more pronounced than those without controlling for sector-specific effects. More importantly, it indicates that the substitution patterns could be heterogeneous across firms even within the same sectors.

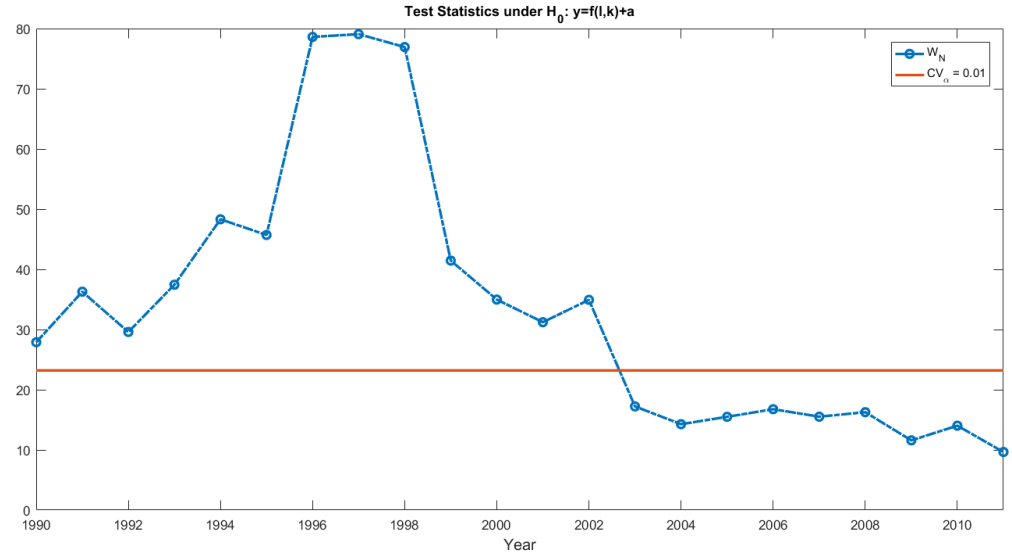
The empirical results show that there are indeed periods of non-Hicks-neutral productions in the U.S. manufacturing industries, which coincide with the rapid adoption of computer technology that occurred in the 90s. Clearly, the proposed test is able to single out those years of non-Hicks-neutral productions, even after controlling for sector specific effects.

Figure 1: Test Statistics by Year of Nonparametric Production Functions



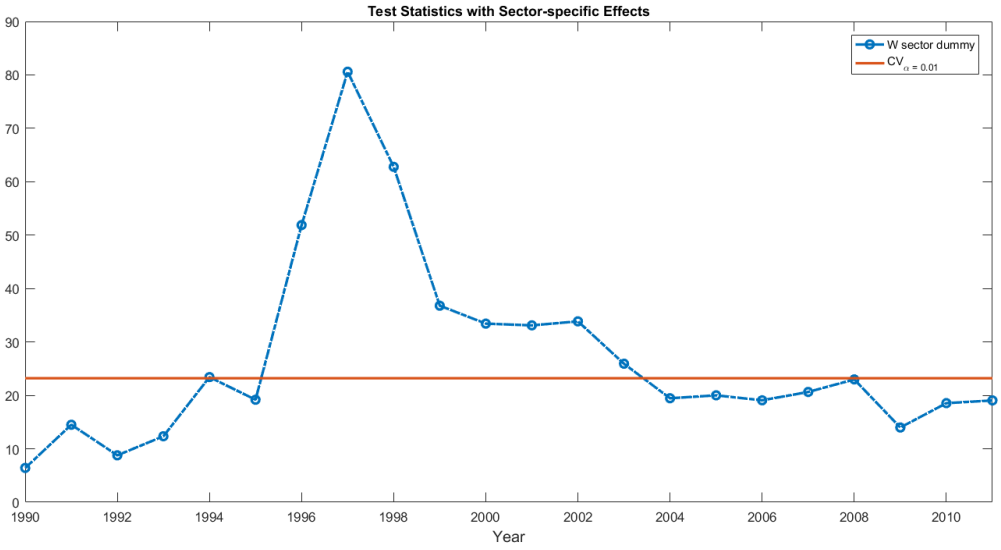
Note: The horizontal line stands for critical values at 1% significant level. Circled markers are test statistics at each year.

Figure 2: Test Statistics by Year of Log Transformed Models



Note: The horizontal line stands for critical values at 1% significant level. Circled markers are test statistics at each year.

Figure 3: Test Statistics by Year of Log Transformed Models with Sector Dummies



Note: The horizontal line stands for critical values at 1% significant level. Circled markers are test statistics at each year.

Table 5: Empirical Testing Results by Year 1990-2011

		(1)		(2)		(3)	
Year	N	W	p-value	W	p-value	W	p-value
1990	1818	91.315	0.000	27.889	0.000	6.414	0.635
1991	1893	87.028	0.000	36.282	0.000	14.506	0.053
1992	1997	100.432	0.000	29.591	0.000	8.779	0.384
1993	2146	100.443	0.000	37.444	0.000	12.354	0.125
1994	2219	89.113	0.000	48.315	0.000	23.426	0.000
1995	2350	102.037	0.000	45.709	0.000	19.188	0.005
1996	2419	90.953	0.000	78.615	0.000	51.880	0.000
1997	2391	98.973	0.000	79.071	0.000	80.590	0.000
1998	2251	104.138	0.000	76.927	0.000	62.781	0.000
1999	2136	98.389	0.000	41.452	0.000	36.796	0.000
2000	1975	53.060	0.000	34.978	0.000	33.438	0.000
2001	1818	79.707	0.000	31.219	0.000	33.107	0.000
2002	1766	49.370	0.000	34.929	0.000	33.863	0.000
2003	1736	27.392	0.000	17.225	0.015	25.927	0.000
2004	1683	29.974	0.000	14.275	0.058	19.464	0.005
2005	1594	32.488	0.000	15.499	0.034	20.012	0.003
2006	1527	96.703	0.000	16.768	0.019	19.081	0.006
2007	1465	57.577	0.000	15.522	0.033	20.642	0.002
2008	1345	59.490	0.000	16.270	0.024	22.987	0.001
2009	1276	69.274	0.000	11.588	0.165	13.985	0.066
2010	1278	71.716	0.000	14.028	0.065	18.548	0.007
2011	1477	32.853	0.000	9.652	0.304	19.053	0.006

Note: 1. Test statistics are reported under W along with the p-values. 2. Number of quantile $P = 10$. 3. Smoothing parameters, $r_1 = 1/7.9, r_2 = 1/7.9$. Trimming parameters, $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$.

9 Conclusions

In this paper, I propose an easy-to-implement test for structural separability of fully nonparametric models, explicitly allowing maximal unobserved heterogeneity. The test is motivated by recent advances in the literature of structural modeling and nonparametric identification. In particular, one of the distinct features is that no shape restrictions or distributional assumptions need to be imposed. But in so doing, one has to overcome the non-identification problem in the presence of multi-dimensional unobservables. As opposed to the previous methods, the test relates the ASF to the additivity of unobservables, as ASFs contain important information on additive separability and could be exploited for testing purpose. The performance of the test statistics is confirmed through the Monte Carlo experiments. In particular, even in fairly small samples, the test possesses good power to some extent. With only slight modifications, it can be adapted to more practical scenarios, including semiparametric single-index models and triangular simultaneous equations. The test is relatively robust to the choice of number of quantile regions. However, developing the optimal number of quantile regions is beyond the scope of this paper and could be a future direction.

The secondary purpose of this paper is to apply the proposed test to examine a commonly imposed functional form—Hicks-neutral technological shocks in production function estimation. As argued, not only does such assumption significantly restrict the input substitution patterns but also make many identification approaches questionable once it fails to hold. With a firm-level dataset on U.S. manufacturing industry over 22 years, the test has detected a period of strong non-Hicks-neutral productions in the late 90s, which may be attributed to the mass adoption of computer technology.

Appendix A Proof of Identification Results

Proof of Proposition 3.1. The proof is also given in Blundell and Powell (2003); Imbens and Newey (2009). Given $x \in \mathcal{X}$, it follows that

$$\begin{aligned} g(x) &= \int_{\mathcal{E}} m(x,e)dF_{\varepsilon}(e) = \int m(x,e)dF_{\varepsilon|V}(e|v)dF_V(v)(e) \\ &= \int_{\mathcal{V}^x} \int m(x,e)dF_{\varepsilon|V,X}(e|v,x)dF_V(v) \\ &= \int_{\mathcal{V}} E(Y|X=x,V=v)dF_V(v) \end{aligned}$$

The last equality invokes the large support assumption to obtain point identification. □

Proof of Proposition 3.2. By Assumption I.1′, $E(U|X=x,V=v) = E(U|V=v) \equiv h(v)$, $\forall (x,v) \in (\mathcal{X} \times \mathcal{V})$ and under model (2.2) note that $C(x,v) = m_1(x) + h(v)$. Suppose there is another set of functions such that $C(x,v) = \widetilde{m}_1(x) + \widetilde{h}(v)$. By Assumption I.2′, $m_1(x) = \widetilde{m}_1(x) + c_\delta$ and $h(v) = \widetilde{h}(v) - c_\delta$. Then it is obvious that $m_1(\cdot)$ and $h(\cdot)$ are identified up to an additive constant. See Newey et al. (1999) for details.

Now integrate marginally with respect to v on both sides of $C(x,v)$,

$$g(x) = \int_{\mathcal{V}} C(x,v)dF_V(v) = m_1(x) + E(h(V)) \equiv m_1(x) + c_h$$

Assumption I.2 guarantees that $C(x,\cdot)$ is well-defined on $\mathcal{V}$ for each x. Since $g(\cdot)$ is identified from Proposition 1, $m_1(\cdot)$ is identified up to a constant. □

Proof of Proposition 3.3. a) is obvious from Proposition 2 once $c_h = 0$ by normalization since $a(x) = m_1(x) = E(Y - h(V)|X=x)$. To show b), given $X = x$,

$$\begin{aligned} & a(x) = g(x) \Leftrightarrow E(Y - h(V)|X=x) = g(x) \\ \Leftrightarrow\ & E\left(Y - \int C(x',V)dF_X(x')|X=x\right) + E(Y) = g(x) \\ \Leftrightarrow\ & E(Y|X=x) - \int_{\mathcal{V}}\int_{\mathcal{X}} C(x',v)dF_X(x')dF_{V|X}(v|x) + E(Y) = g(x) \\ \Leftrightarrow\ & \int_{\mathcal{V}} C(x,v)dF_{V|X}(v,x) = \int_{\mathcal{V}} C(x,v)dF_V(v) + \Delta(x) \end{aligned}$$

and where $\Delta(x) = \int C(x',v)dF_X(x')dF_{V|X}(v,x)$. □

Proof of Proposition 8.1. In the first step, we prove that $V = F(M|K) = F_\omega(\omega) \sim U[0,1]$. In the second step, it is sufficient to show that conditioning on V is equivalent to conditioning on ω. Therefore (K,L) is independent of (ω,ε).

First. Let $\omega = \mathbb{M}^{-1}(K, M)$

$$\begin{aligned} F(M|K) = \Pr(\mathbb{M}(K,\omega) \leq m|K) &= \Pr(\omega \leq \mathbb{M}^{-1}(K,m)|K) \\ &= \Pr(\omega \leq \mathbb{M}^{-1}(K,m)) \\ &= \Pr(\omega \leq w) = F_\omega(\omega) \end{aligned}$$

Second, conditioning on $F_\omega(\omega)$ is equivalent to conditioning on ω, so it is obvious from A-PF.S2 that $(K, L) \perp (\omega, \varepsilon)|V$. □

Appendix B Asymptotic Proof

Notation. $C(x,v) = E(Y|X = x, V = v)$, $\Delta_i = \widehat{h}(V_i) - h(V_i)$, $f^*(x,v) = f(x)f(v)/f(x,v)$, where $f(\cdot)$ denotes marginal/joint densities. $t_i = \mathbf{1}\{X_i \in \mathcal{X}_0\}$. Remember that under $\mathbb{H}_0$, $Y = m_1(X) + h(V) + \epsilon$. Also let $\sum_{i,j}^N \equiv \sum_{i=1}^N \sum_{j=1}^N$, $\sum_{i,j,k}^N \equiv \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N$, $\sum_{j>i}^N \equiv \sum_{i=1}^{N-1} \sum_{j>i}^N$ and $\sum_{k>j>i}^N \equiv \sum_{i=1}^{N-2} \sum_{j>i}^{N-1} \sum_{k>j}^N$. Let $K_{i,j}^X = K_h(X_i - X_j) = k(X_i - X_j/h)/h$, $K_{i,\cdot}^X = K_h(X_i - x)$ and similar for other variables. I suppress superscript (or subscript) p for quantile(and trimming) indicators.

B.1 Main proofs

Proof of Theorem 5.1. Recall $\widehat{T}_N^p$ in Eq. (4.12) and $\widetilde{T}_N^p$ in Eq. (5.3). In the main text, $\widehat{T}_N^p$ is decomposed first into three components

$$\widehat{T}_N^p = I_1 + I_2 + I_3$$

As noted in the text, $I_3 = 0$ under $\mathbb{H}_0$. The following Lemma 1 aids to prove $I_2 = o_p(N^{-1/2})$. To analyze I_1, it suffices to study $\widetilde{T}_N^p$,

$$\widetilde{T}_N^p = D_N^g + D_N^a + D_N^h$$

as $\widehat{T}_N^p = \widetilde{T}_N^p + o_p(N^{-1/2})$, where D_N^g, D_N^a and D_N^h are defined in Eq. (5.4), (5.5) and (5.6), respectively. Lemma 2 handles D_N^a using second order U-statistic theorem and Lemma 3 deals with D_N^g using third order U-statistic theorem. Lemma 4 provides the intermediate result for Lemma 5 on D_N^h. For illustrative purpose, only the proof of Lemma 2 is shown in the following. Other proofs can be found in the supplemental materials. Then it is shown that

$$\sqrt{N}\widehat{T}_N^p = N^{-1/2} \sum_{i=1}^N (\xi_{gi}^p + \xi_{ai}^p + \xi_{hi}^p) + o_p(1)$$

where

$$\begin{aligned}\xi_{gi}^p &= t_i^p f^*(X_i, V_i)\epsilon_i + E(t^p)h(V_i)\\ \xi_{ai}^p &= -t_i^p\epsilon_i\\ \xi_{hi}^p &= E(t^p|V_i)f^*(X_i, V_i)\epsilon_i\end{aligned}$$

Combine those three terms, $\xi_i^p = \xi_{gi}^p + \xi_{ai}^p + \xi_{hi}^p$.

$$\xi_i^p = [t_i^p + E(t^p|V_i)]f^*(X_i, V_i) - t_i^p]\epsilon_i + E(t^p)h(V_i)$$

By the CLT, Theorem 5.1 is established with the limiting variance $\Omega_p = E(\xi_i^p \xi_i^{p\prime})$.

□

Proof of Theorem 5.2 and Corollary 5.2.1. According to Theorem 5.1, it is true that $\widehat{T}_N$ follows a P-dimensional multivariate normal distribution.

$$\sqrt{N}\widehat{T}_N \xrightarrow{\text{D}} N(\mathbf{0}, \Omega)$$

So $W_N = NT_N'\Omega^{-1}T_N \xrightarrow{\text{D}} \chi_P^2$. By Slutsky's theorem, for any $\widehat{\Omega}_N \xrightarrow{\text{P}} \Omega$, then it holds that $\widehat{W}_N = NT_N'\widehat{\Omega}_N^{-1}T_N \xrightarrow{\text{D}} \chi_P^2$. □

Lemma 1 I_2**.** *Suppose H_0 is true, under Assumption A.1-A.6, for each p*

$$\sqrt{N}I_2 \equiv \sqrt{N}\sum_{i=1}^{N}(\widehat{t}_i - t_i)(\widehat{D}(X_i) - D(X_i)) = o_p(1)$$

Proof. For any X_i,

$$|\widehat{D}(X_i) - D(X_i)| = \Big|[\widehat{g}(X_i) - g(X_i)] + [\widehat{a}(X_i) - a(X_i)]\Big| \leq |\widehat{g}(X_i) - g(X_i)| + |\widehat{a}(X_i) - a(X_i)| = O_p((Nh)^{-1/2})$$

According to this, it is true that,

$$N^{-1}\sum_{i=1}^{N}|\widehat{D}(X_i) - D(X_i)|^2 = O((Nh)^{-1})$$

By Cauchy-Schwartz inequality,

$$\begin{aligned}\sqrt{N}\sum_{i=1}(\widehat{t}_i - t_i)(\widehat{D}(X_i) - D(X_i))/N &\leq \sqrt{N}\sqrt{\sum_{i=1}(\widehat{t}_i - t_i)^2/N}\sqrt{\sum_{i=1}[\widehat{D}(X_i) - D(X_i)]^2/N}\\ &= \sqrt{N}o_p(N^{-1/2})O_p(N^{-1/2}h^{-1/2}) = o_p(1)\end{aligned}$$

□

Lemma 2 D_N^a**.** *Suppose that Assumption A.1-A.6 hold, under $\mathbb{H}_0$, then D_N^a in Eq.* (5.5) *can*

be written as the following,

$$D_N^a = -N^{-1}\sum_{i=1}^{N} t_i[Y_i - h(V_i) - a(X_i)] + o_p(N^{-1/2})$$

Proof. Let $Y_i^+ \equiv Y_i - h(V_i)$. Recall that

$$D_N^a = -N^{-1}\sum_{i=1}^{N} t_i[\widehat{E}(Y^+|X_i) - E(Y^+|X_i)]$$

Apply the intermediate Lemma A 2 in the supplemental material, it is true that $\widehat{E}(Y^+|X_i) - E(Y^+|X_i) = (N-1)^{-1}\sum_{j\neq i}^{N} K_{j,i}^X[Y_j^+ - E(Y^+|X_i)]/f(X_i) + o_p(N^{-1/2})$. Substitute this into D_N^a and note that $D_N^a = \widetilde{D}_N^a + o_p(N^{-1/2})$. From now on, it suffices to only work with $\widetilde{D}_a$ below,

$$\begin{aligned}
\widetilde{D}_N^a &= -\frac{1}{N(N-1)}\sum_{i,j}^{N} t_i f(X_i)^{-1} K_{j,i}^X[Y_j^+ - E(Y^+|X_i)] \\
&= -\binom{N}{2}^{-1}\sum_{j>i}^{N}(a_{ij} + a_{ji})/2
\end{aligned}$$

To apply the U-statistic theorem, we rewrite $\widetilde{D}_N^a$ as symmetric in i and j and where

$$\begin{aligned}
a_{ij} &= t_i f(X_i)^{-1} K_{j,i}^X[Y_j^+ - E(Y^+|X_i)] \\
a_{ji} &= t_j f(X_j)^{-1} K_{i,j}^X[Y_i^+ - E(Y^+|X_j)]
\end{aligned}$$

Moreover, by intermediate lemmas,

$$E(a_{ij}|X_i) = O(h^4); E(a_{ji}|X_i) = t_i[Y_i^+ - E(Y^+|X_i)] + O(h^4)$$

By Assumption A.1-A.3, it is true that $E|a_{ji}|^2 + E|a_{ij}|^2 = O(1) = o(N)$ as every multiplicative term is bounded. Therefore, the standard second order U-statistic applies,

$$\widetilde{D}_N^a = -N^{-1}\sum_{i=1}^{N} t_i[Y_i^+ - E(Y^+|X_i)] + O(h^4) + o_p(N^{-1/2})$$

Under $\mathbb{H}_0$ and Assumption A-5, $O(h^4) = o(N^{-1/2})$, it can be simplified to

$$D_N^a = -N^{-1}\sum_{i=1}^{N} t_i[Y_i - h(V_i) - a(X_i)] + o_p(N^{-1/2}) = -N^{-1}\sum_{i=1}^{N} t_i\epsilon_i + o_p(N^{-1/2})$$

□

Due to its similarity, the proofs of Lemma 3, Lemma 4 and Lemma 5 are skipped here and

presented in the supplemental materials as their proofs are nothing but more complicated as they involve third-order U-statistics theorems. All intermediate lemmas are also moved to supplemental materials to save space.[18]

Lemma 3 D_N^g**.** *Suppose that Assumption A.1-A.6 hold and under* $\mathbb{H}_0$*, then* D_N^g *in Eq.* (5.4) *can be written as the following,*

$$D_N^g = N^{-1} \sum_{i=1}^{N} \{t_i f^*(X_i, V_i)\epsilon_i + E(t)h(V_i)\} + o_p(N^{-1/2})$$

Lemma 4 Δ_i**.** *Given* X_i *and* $V_i = v$*, let* $\Delta(v) = \widehat{h}(v) - h(v)$*, then it follows that*

$$\Delta(v) = \Delta_1(v) + \Delta_2(v) + \Delta_3 + o_p(N^{-1/2})$$

where

$$\begin{aligned} \Delta_1(v) &= \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)K_{i,\cdot}^V}{f(X_i, v)}[Y_i - C(X_i, v)] \\ \Delta_2 &= \frac{1}{N} \sum_{i=1}^{N} m_1(X_i) - E[m_1(X)] \\ \Delta_3 &= E(Y) - \overline{Y} \end{aligned}$$

Lemma 5 D_N^h**.** *Suppose that Assumption A.1-A.6 hold and under* $\mathbb{H}_0$*, then* D_N^h *in Eq.* (5.6) *can be written as the following,*

$$D_N^h = N^{-1} \sum_{i=1}^{N} E(t|V_i) f^*(X_i, V_i)\epsilon_i + o_p(N^{-1/2})$$

Acknowledgement

This is one of the major chapters of the author's doctoral dissertation titled "Essays on Nonparametric Structural Econometrics: Theory and Applications". The author gratefully thanks his advisor, Roger W. Klein, for his sincere and valuable help. The author is also grateful for all constructive comments from Norman R. Swanson, Chia-Shang J. Chu, Yuan Liao, Xiye Yang, John Landon-Lane and others. All errors are purely the author's.

[18] Intermediate Lemma A1 presents the U-statistic theorem from Serfling (2009). Lemma A2 gives results of recursive bias correction from Shen and Klein (2017). Lemma A3 handles indicator functions.

References

Daniel A Ackerberg, Kevin Caves, and Garth Frazer. Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451, 2015.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Randy Becker, Wayne Gray, and Jordan Marvakov. Nber-ces manufacturing industry database: Technical notes. *NBER Working Paper*, 5809, 2013.

C Lanier Benkard and Steven Berry. On the nonparametric identification of nonlinear simultaneous equations models: Comment on brown (1983) and roehrig (1988). *Econometrica*, 74(5):1429–1440, 2006.

Richard Blundell and Stephen Bond. Gmm estimation with persistent panel data: an application to production functions. *Econometric reviews*, 19(3):321–340, 2000.

Richard Blundell and James L Powell. Endogeneity in nonparametric and semiparametric regression models. *Econometric Society Monographs*, 36:312–357, 2003.

Richard W Blundell and James L Powell. Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679, 2004.

Bryan W Brown. The identification problem in systems nonlinear in the variables. *Econometrica*, 51(1):175–96, 1983.

Martin Browning and Jesus Carro. Heterogeneity and microeconometrics modeling. *Econometric Society Monographs*, 43:47, 2007.

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

Victor Chernozhukov, Guido W Imbens, and Whitney K Newey. Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4–14, 2007.

Andrew Chesher. Excess heterogeneity, endogeneity and index restrictions. *Journal of Econometrics*, 152(1):37–45, 2009.

Jan De Loecker. Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. *Econometrica*, 79(5):1407–1451, 2011.

Xavier D'Haultfœuille and Philippe Février. Identification of nonseparable triangular models with discrete instruments. *Econometrica*, 83(3):1199–1210, 2015.

Yanqin Fan and Qi Li. Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica: Journal of the econometric society*, pages 865–890, 1996.

Jean-Pierre Florens, James J Heckman, Costas Meghir, and Edward Vytlacil. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica*, 76(5):1191–1206, 2008.

Zvi Griliches and Jacques Mairesse. Production functions: the search for identification. Technical report, National Bureau of Economic Research, 1995.

Jinyong Hahn and Geert Ridder. Conditional moment restrictions and triangular simultaneous equations. *Review of Economics and Statistics*, 93(2):683–689, 2011.

Jinyong Hahn et al. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–332, 1998.

Aaron K Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2):303–316, 1987.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

James J Heckman, Daniel Schmierer, and Sergio Urzua. Testing the correlated random coefficient model. *Journal of Econometrics*, 158(2):177–203, 2010.

Stefan Hoderlein and Enno Mammen. Identification and estimation of local average derivatives in non-separable models without monotonicity. *The Econometrics Journal*, 12(1):1–25, 2009.

Stefan Hoderlein, Liangjun Su, and Halbert White. Specification testing for nonparametric structural models with monotonicity in unobservables. *Working Paper*, 2011.

Joel L Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, 2011.

Guofang Huang and Yingyao Hu. Estimating production functions with robustness against errors in the proxy variables. *Available at SSRN 1805213*, 2011.

Martin Huber and Giovanni Mellace. Testing exclusion restrictions and additive separability in sample selection models. *Empirical Economics*, 47(1):75–92, 2014.

Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120, 1993.

Hidehiko Ichimura and Lung-Fei Lee. Semiparametric least squares estimation of multiple index models: single equation estimation. In *Nonparametric and semiparametric methods in econometrics and statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics. Cambridge*, pages 3–49, 1991.

Guido W Imbens. Nonadditive models with endogenous regressors. *Econometric Society Monographs*, 43:17, 2007.

Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009.

Kyoo Kim, Amil Petrin, and Suyong Song. Estimating production functions when capital input is measured with error. 2013.

Roger Klein and Chan Shen. Semiparametric instrumental variable estimation in an endogenous treatment model. 2015.

Roger W Klein. Specification tests for binary choice models based on index quantiles. *Journal of Econometrics*, 59(3):343–375, 1993.

Roger W Klein and Richard H Spady. An efficient semiparametric estimator for binary response models. *Econometrica*, 61(2):387–421, 1993.

James Levinsohn and Amil Petrin. Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies*, 70(2):317–341, 2003.

Arthur Lewbel, Xun Lu, and Liangjun Su. Specification testing for transformation models with an application to generalized accelerated failure-time models. *Journal of Econometrics*, 184(1):81–96, 2015.

Oliver Linton and Jens Perch Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):93, 1995.

Oliver B Linton. Miscellanea efficient estimation of additive nonparametric regression models. *Biometrika*, 84(2):469–473, 1997.

Oliver B Linton. Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory*, 16(04):502–523, 2000.

Xun Lu and Halbert White. Testing for separability in structural equations. *Journal of Econometrics*, 182(1):14–26, 2014.

Enno Mammen, Christoph Rothe, Melanie Schienle, et al. Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, 40(2):1132–1170, 2012.

Matthew Masten and Alexander Torgovitsky. Instrumental variables estimation of a generalized correlated random coefficients model. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 2014.

Rosa L Matzkin. Nonparametric estimation of nonadditive random functions. *Econometrica*, 71(5):1339–1375, 2003.

Rosa L Matzkin. Nonparametric identification. *Handbook of Econometrics*, 6:5307–5368, 2007.

Rosa L Matzkin. Identification in nonparametric simultaneous equations models. *Econometrica*, 76(5):945–978, 2008.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9 (1):141–142, 1964.

Whitney K Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(02):1–21, 1994.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Whitney K Newey, James L Powell, and Francis Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

G Steven Olley and Ariel Pakes. The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297, 1996.

James L Powell, James H Stock, and Thomas M Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430, 1989.

Charles S Roehrig. Conditions for identification in nonparametric and parametric models. *Econometrica: Journal of the Econometric Society*, pages 433–447, 1988.

Susanne Schennach, Halbert White, and Karim Chalak. Local indirect least squares and average marginal effects in nonseparable structural systems. *Journal of Econometrics*, 166 (2):282–302, 2012.

Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Chan Shen and Roger Klein. Market recursive differencing: Bias reduction with regular kernels. Working paper, 2017.

Stefan Sperlich, Dag Tjøstheim, and Lijian Yang. Nonparametric estimation and testing of interaction in additive models. *Econometric Theory*, 18(02):197–251, 2002.

Liangjun Su, Yundong Tu, and Aman Ullah. Testing additive separability of error term in nonparametric structural models. *Econometric Reviews*, 34(6-10):1057–1088, 2015.

Liangjun Su, Stefan Hoderlein, and Halbert White. Testing monotonicity in unobservables with panel data. 2016.

Alexander Torgovitsky. Identification of nonseparable models using instruments with small support. *Econometrica*, 83(3):1185–1197, 2015.