Additive Separability and Unobserved Heterogeneity: A Test for Hicks-neutral Productivity Shocks

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 properties of average structural functions of nonseparable and separable models. In particular, a Wald-type test statistic, combining empirical differences over quantile regions, is suggested. I investigate its statistical properties and extend it to various settings. The method is then applied on testing the functional form of Hicks-neutral productivity shocks when estimating fully nonparametric firm-level production functions. The results highlight a period of non-Hicks-neutral productions in the U.S. manufacturing industries during the late 90s.

Keywords: Additive Separability; Average Structural Function; Nonparametric Test; Production Functions; Hicks-neutrality

*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. 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 fail to hold once unobservables under nonseparability. Therefore, a specification test of structural separability can be empirically useful. However, a plethora of empirical research has only focused on additive models until recently. Further, additivity may hide the problem of multiplicity of unobservables. In practice, unobservables are often multi-dimensional of unknown dimension. Such fact complicates the testing problem because it implies that structural functions cannot be identified without imposing substantial shape restrictions or distributional assumptions. 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 properties of average structural functions of competing specifications, namely additive separable versus nonseparable. 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 usually represent heterogeneity in consumer tastes, product attributes, firm-specific productivity shocks, measurement and unexpected errors, etc. As such, the unobserved heterogeneity is rarely unit-dimensional and quite often even the number of dimensions is not known unsurprisingly. Browning and Carro (2007) argue that most empirical models permit less heterogeneity than actually present. As the proliferation of micro level data, empirical research has begun to take into account a richer set of heterogeneity. However, without substantial shape or distributional assumptions, nonseparable structural functions are generally not identified. To circumvent this difficulty, I build the test around average structural functions (ASFs). The identification of ASFs is wellstudied in the control function literature (Blundell and Powell, 2004; Imbens, 2007; Imbens and Newey, 2009; Florens et al., 2008, etc). 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. In this paper, I show that testing the equality of ASFs generated by the two competing specifications can deliver important implications on additive separability, though not perfect. In particular, once the equality is being rejected, it gives immediate support for nonseparable models. But caution need to be exercised when non-rejection is found. Further discussion is presented in the main text.

This paper contributes to the literature of testing additive separability, paying special attention on models with unrestricted unobservables. Despite its empirical importance, testing additive separability has only received limited attention so far. In particular, Lu and White (2014) show that structural additivity can be cast into a conditional independence condition. Either the scalar monotonicity in unobservables or some polynomial structure is required 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, also relying on the scalar monotonicity assumption. The imposed restriction of unobserved heterogeneity can be substantial and might limit the scope of their applicability. There are also tests proposed in specific settings such as sample selection models (Huber and Mellace, 2014), random coefficient models (Heckman et al., 2010), transformation models Lewbel et al. (2015), etc. Nonparametric tests on scalar monotonicity are considered in Hoderlein et al. (2011) and Su et al. (2016). An incomplete list of other related works include, but are not limited to Sperlich et al. (2002); Hoderlein and Mammen (2009); Heckman et al. (2010), etc.

To conduct the test, a nonparametric empirical quantile mean (EQM) test, similar to Klein (1993), is developed. The main idea is to compare average differences between two functions in quantile regions of observables. For each quantile region, the average difference converges to a normal distribution in large samples. By combining all quantile regions, one can obtain an easy-to-implement Wald-type test statistic. To control for the asymptotic bias, a recursive bias correction technique recently developed by Shen and Klein (2017) is applied. In addition, I find that the asymptotic variance of our test statistic performs reasonably well under moderate sample sizes. Straightforwardly, the nonparametric test can be adapted to many different settings. In this paper, I consider two extensions—semiparametric single-index models and models with nonparametrically "generated" control variates—both of which are commonly seen in empirical studies.

Finally, the proposed method is being applied in testing a fundamental functional form assumption—Hicks-neutral technological shocks when estimating fully nonparametric production functions. Hicks-neutrality implies that firm-specific productivity is only multiplicative to the production function, or additive in logarithms forms. 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 firm-level unobservables. Consequently, inconsistent estimates of structural parameters, such as output-input elasticities, return-to-scale, and productivity distribution could be produced. In this paper, I empirically test this primal assumption 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. Such conjecture may worth future investigations. Hence by providing methods and empirical support, this paper also contributes to the recent resurgence of interest in estimating firm-level non-Hicks-neutral production functions (Kasahara et al., 2015; Li and Sasaki, 2017; Hu et al., 2017, etc).

In what follows, Section 2 discusses identification problems and then derives testing implications. Section 3 provides the nonparametric test statistics. Asymptotic and finite sample properties are summarized in Section 4. Two extensions are presented in Section 5. Section 6 connects to the application and provides the motivations. The empirical results are also presented in Section 7. Section 8 concludes. All proofs are given in the appendix and supplemental materials.

2 Identification and Testing Implications

2.1 Nonseparability and Non-identification

Nonparametric nonseparable models have been gaining popularity in theoretical econometrics for the past decades, as in Eq. (2.1).

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

where the unknown function $m : \mathcal{X} \times \mathcal{E} \to \mathbb{R}$ denotes the structural function that represents some primitive economic relations. Such models are capable of capturing both observed and unobserved heterogeneity in the structural parameters of economic interest by allowing arbitrary interaction between X and ε. In our empirical application, model (2.1) represents a nonparametric firm production function, where Y denotes the output level, X as amount of factor inputs and ε consisting of unobservables including time-varying and time-invariant productivity shocks, input quality variations, measurement errors in output and inputs, etc., potentially multi-dimensional.

While a special class of model (2.1), more favored in empirical works, takes the additive separable structure in the unobservables,

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

where both $m_1 : \mathcal{X} \to \mathbb{R}$ and $m_2 : \mathcal{E} \to \mathbb{R}$ are unknown functions. The additive error, ε, is often taken to include measurement errors and omitted variables such as in linear models. Previous literature has paid particular attention to recovering $m_1(\cdot)$ in the presence of endogenous regressors (Newey et al., 1999; Newey and Powell, 2003; Carrasco et al., 2007;

Horowitz, 2011, etc). According to previous literature, motivations for testing the structural separability below are manifold.

$$\mathbb{H}_0^* : m(X, \varepsilon) = m_1(X) + m_2(\varepsilon), \quad \mathbb{H}_1^* : \text{Otherwise}$$

Firstly, 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. Second, it is a test of a particular class of estimators whose identification power comes only from the conditional mean restriction, such as IV estimators shown in Hahn and Ridder (2011). Third, more efficient estimators are available 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 by exploiting the additive restriction.[1]

Unfortunately, testing $\mathbb{H}_0^*$ against $\mathbb{H}_1^*$ might not always have power when unobserved heterogeneity, ε, is modeled fully flexibly. This is due to the non-identification problem of the nonseparable functions, a point that can be simply illustrated in the following well-known example. Suppose X is univariate continuous variable independent of ε_i which are independent standard normal variates for $i = 1, 2$ and 3.

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

The above data generating processes are observationally equivalent in the sense that the joint distributions of all observables, i.e. $F_{X,Y}$, would be identical. 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), the above two structural functions $y = m(x, e)$ are observationally equivalent and thus not identified. As a consequence, it implies that testing the original hypothesis of

[1] Besides, testing separability could also yield implications on endogeneity, a point mentioned in Imbens (2007) and Imbens and Newey (2009).

$\mathbb{H}_0^*$ versus $\mathbb{H}_1^*$ may not always have power in general.[2] Benkard and Berry (2006) give a more detailed discussion on this issue.

One solution is to impose additional structures (Matzkin, 2003, etc). In the context of testing for structural separability, previous works have focused on imposing shape restrictions such as scalar monotonicity in unobservables such as Su et al. (2015) and Lu and White (2014), like discussed in Section 1. Nonetheless, 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 the example in 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 difficult to justify. Another direction is to determine what can be identified without compromising the dimensionality of heterogeneity (Blundell and Powell, 2004; Imbens and Newey, 2009, etc). Under many situations, 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, an identifiable object, even in the presence of unrestricted heterogeneity.[3]

2.2 Identification of ASF

Before deriving testable implications, I first consider the identification of ASF under $\mathbb{H}_0$ and $\mathbb{H}_1$, respectively. Define the ASF at $X = x$ of nonseparable models in Eq. (2.3)

$$ASF(x) \equiv g(x) = \int_{\mathcal{E}} m(x,e)dF_{\varepsilon}(e), \quad \forall x \in \mathcal{X} \tag{2.3}$$

[2] 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.

[3] 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.

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 constructed. When X is binary, the ATE can be obtained by the difference between $g(1)$ and $g(0)$.

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 $\mathcal{E} \subset \mathbb{R}^{\infty}$.[4] The identification of ASF without endogeneity is trivial and is simply $E(Y|X = x), \forall x \in \mathcal{X}$. Thereafter, I will focus on models where X and ε are not necessarily independent. Following the literature, suppose that there are control variables $V \in \mathcal{V} \subset \mathbb{R}^{d_V}$, satisfying Assumption I.1 and I.2.

Assumption I.1 Conditional independence. $X \perp \varepsilon | V$.

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

Assumption I.1 parallels the unconfoundedness condition in the treatment effect literature.[5] Admittedly, Assumption I.2 is a relatively strong condition. On the other hand, the large support condition might hold only over some subregion of X, say $\mathcal{X}_0$. As a result, our test may only work on $\mathcal{X}_0$. There are many ways to obtain V. In some cases, V might be readily available and observed. In other cases, the control variables can be "generated", for instance, in a triangular equations frameworks like the application of this paper. Next, I focus on the identification of ASFs for nonseparable models and additive separable models, respectively, preceding the discussion of the testable implications.

Proposition 2.1 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.

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

[5] It is also ruled out that X and ε be exact functions of V; otherwise, they would be degenerate given V.

Proposition 2.1. *Under Assumption I.1 and I.2, $g(\cdot)$ defined in Eq.* (2.3) *is identified at each $x \in \mathcal{X}$,*

$$g(x) = \int_{\mathcal{V}} C(x, v) dF_V(v) \tag{2.4}$$

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 $\mathcal{V}$.[6]

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. Under Assumption I.1 and I.2, the ASF of model (2.2) is immediately identified through Proposition 2.1 since additive models belong to a subclass of nonseparable models. However, a weaker set of assumptions suffices to identify $a(\cdot)$.

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

Assumption I.2′ Nonexistence of 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.1′ doesn't require full independence conditional on the control variates, therefore weaker than Assumption I.1. 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,

$$C(x, v) = m_1(x) + h(v), \forall (x, v) \in \mathcal{X} \times \mathcal{V} \tag{2.5}$$

Assumption I.2′ rules out the possibility of exact additive functional dependence between $m_1(x)$ and $h(v)$. The formal proof is given in Newey et al. (1999) and the identification result is summarized in Proposition 2.2.

[6] The proof is given in Theorem 1 in Imbens and Newey (2009).

Proposition 2.2. *Under Assumption I.1′ and I.2′, a). $m_1(\cdot)$ and $h(\cdot)$ in Eq. (2.5) is identified up to an additive constant for each $(x, v) \in \mathcal{X} \times \mathcal{V}$. b). $g(\cdot)$ defined in Eq. (2.3) is identified at each $x \in \mathcal{X}$, then $g(x) = m_1(x) + c_h$ where $E[h(V)] = c_h$.*[7]

Without loss of generality, I normalize that $E[h(V)] \equiv c_h = 0$, attributing all constants into $m_1(\cdot)$. 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. (2.6).[8]

$$h(v) = \int_{\mathcal{X}} C(x, v) dF_X(x) - E(Y) \tag{2.6}$$

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 $X = x$ as if $h(\cdot)$ were known,

$$a(x) \equiv E(Y - h(V)|X = x), \ \forall x \in \mathcal{X} \tag{2.7}$$

In Proposition 2.3, it states that $a(\cdot)$ identifies ASFs for additive models, i.e. $a(x) = m_1(x)$, yet not for nonseparable models in general. This also explains why tests that based on ASFs could have some power on testing additive separability, though not being equivalent.

Proposition 2.3. *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*

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

[7] The proof is given in Theorem 2.1 and 2.2 in Newey et al. (1999).

[8] 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.

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

The equality in Proposition 2.3 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.

2.3 Testing Implications

Now come back to the testing problem aforementioned. Unfortunately, the original set of hypotheses turns out to be non-testable due to the non-identification of structural functions. Hence in this paper, instead of testing $\mathbb{H}_0^*$ against $\mathbb{H}_1^*$, I consider another set of testable hypotheses below,

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

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. (2.4) recovers the ASF for both models whereas $a(\cdot)$ in Eq. (2.7) only recovers the ASF for additive models (and a class of nonseparable models satisfying the condition stated in Proposition 2.3). 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 observables. 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 4.2 and empirical results in Section 7. Finally, were ASFs and its variants, such as average marginal effects, 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 the restriction and treating the model as if it had an additive error structure. In other words, the specification test of ASFs is important in its own right as it sheds light on the consistency and efficiency of ASF estimators.

So our test can be thought as relatively conservative. Rejection of $\mathbb{H}_0$ immediately implies rejection of $\mathbb{H}^*$, indicating a nonseparable structural function. However, the reverse is not true in general. This might be a shortcoming of the suggested test since the equivalence is lost, instead 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. Finally, examples below are to show cases where our proposed testing method has no power against additive separability.

Example 1. No Endogeneity. Suppose that $X \perp \varepsilon$ and V is of null dimension. Then, $g(x) = a(x)$ for all x. Even if the true model is nonseparable, it can always be written as the additive one, i.e. $Y = E(Y|X) + \epsilon,\ \epsilon = m(X, \varepsilon) - E(Y|X)$ where $E(\epsilon|X) = 0$, so the condition in Proposition 2.3 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 $Y = X\varepsilon_1 + \varepsilon_2$ 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 $a(x) = g(x)$.

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 integration. Suppose $V = \varepsilon$ and $Y = X\varepsilon,\quad E(\varepsilon|X) = 0$. 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 $a(x) = g(x) = 0$.

3 Estimation and Testing

3.1 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 to estimate conditional mean functions. In principle, other nonparametric smoothers such as local polynomials and sieve estimators can be applied as well.[9] 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. (3.1),

$$\widehat{C}_0(x, v) = \frac{\sum_{i=1}^{N} K_{h_1}(X_i - x)K_{h_1}(V_i - v)Y_i}{\sum_{i=1}^{N} K_{h_1}(X_i - x)K_{h_1}(V_i - v)} \tag{3.1}$$

where admitted some of abuse of notation, $K_h(\cdot) = \prod_d [k(\cdot/h)/h]$ represents the d-dimensional product of kernel functions.[10]

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.[11] 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. (3.2).

$$\widehat{C}(X_l, V_j) = \frac{\sum_{i \neq j,l}^{N} K_{h_1}(X_i - X_l)K_{h_1}(V_i - V_j)[Y_i - \widehat{\delta}_i(X_l, V_j)]}{\sum_{i \neq j,l}^{N} K_{h_1}(X_i - X_l)K_{h_1}(V_i - V_j)} \tag{3.2}$$

Next consider the estimation of ASF. The partial mean estimator, similar to that in

[9] 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.

[10] Bandwidths are allowed to be different for X and V.

[11] 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.

Newey (1994) is used due to its simplicity.[12] 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. (3.3),

$$\widehat{g}(X_l) = \frac{1}{N-1}\sum_{j\neq l}^{N}\widehat{C}(X_l, V_j), \quad \forall l = 1, \cdots, N \tag{3.3}$$

Likewise, $h(\cdot)$ defined in Eq. (2.6) can be estimated in the similar way in Eq. (3.4)

$$\widehat{h}(V_j) = \frac{1}{N-1}\sum_{i\neq j}^{N}\widehat{C}(X_i, V_j) - N^{-1}\sum_{i=1}^{N} Y_i, \quad \forall j = 1, \cdots, N \tag{3.4}$$

where the mean of Y subtracted resembles the sample analog of the unconditional expectation, $E(Y)$, for normalization.[13]

Now consider the ASF estimator of the "additive" model, $\widehat{a}(\cdot)$. I borrow the idea from Linton (1997) and Mammen et al. (1999) 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 used rather than the marginal integration of the local linear estimator. Now presuming that $h(\cdot)$ is known, the infeasible estimator of $a(\cdot)$ is given in Eq. (3.5).

$$\widetilde{a}(X_l) = \widehat{E}_{h_2}(Y_i - h(V_i)|X_l), \quad l = 1, \cdots, N \tag{3.5}$$

where $\widehat{E}(\cdot)$ is the bias-reducing recursive conditional mean estimator similar to Eq. (3.2), 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. (3.6),

$$\widehat{a}(X_l) = \widehat{E}_{h_2}(Y_i - \widehat{h}(V_i)|X_l) = \frac{\sum_{i\neq l}^{N} K_{h_2}(X_i - X_l)[Y_i - \widehat{h}(V_i) - \widehat{\delta}_i(X_l)]}{\sum_{i\neq l}^{N} K_{h_2}(X_i - X_l)}, \quad l = 1, \cdots, N \tag{3.6}$$

[12] Linton and Nielsen (1995) consider marginal integration and smooth backfitting projection.

[13] Note that the CEF estimator in constructing $\widehat{h}(\cdot)$ could be potentially different from the one in Eq. (3.3) in terms of bandwidth and kernel choices.

Linton (1997) 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 with bias reduction techniques being used. Nevertheless, our estimator of the "additive" ASF does share the same merit. In particular, with the one-step backfitting method, a more efficient estimator of ASF is made possible when $\mathbb{H}_0$ is true.[14]

3.2 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 method that combines information from empirical quantile means (EQM).[15] For the purpose of illustration, consider the univariate continuous variable X for now, but generalizations to multivariate X is straightforward. Denote the empirical ASF difference by $D(X_i) \equiv g(X_i) - a(X_i)$, $\forall i = 1, \ldots, N$. Under $\mathbb{H}_0$, $D(X_i) = 0$ for each i almost surely. To proceed, evenly split the whole sample into P_N number of subsamples or quantile regions thereafter, over the support of X.[16] It can be postulated that for each quantile region, the average 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. Next, define the pth-quantile empirical mean difference as the following,

$$T_N^p \equiv N^{-1} \sum_{i=1}^{N} t_i^p D(X_i), \quad p = 1, \ldots, P_N \tag{3.7}$$

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

[15]The test idea is firstly mentioned in Klein (1993) in specification tests of parametric error distributions versus semiparametric single-index binary choice models.

[16]The number of quantiles can be any positive integer so long as $P_N/N = o(1)$ in theory. More discussion on P_N is at the end of this section.

where the quantile-trimming indicator is defined in Eq. (3.8),

$$t_i^p \equiv \mathbf{1}\left\{\min[c_L, q_X(p-1/P_N)] \le X_i < \max[q_X(p/P_N), c_U]\right\} \tag{3.8}$$

where $q_X(\cdot)$ is the quantile function of X, i.e. $q_X(\tau) = \inf\{x : F_X(x) \ge \tau\}$. 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 $N \to \infty$ by the standard central limit theorem. A Wald-type statistic in Eq. (3.9) could be thus constructed,

$$W_N \equiv N T_N' \Omega^{-1} T_N \tag{3.9}$$

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 4.2 for explicit expressions.

Note that dividing sample into subregions enables researchers to have a closer look across quantiles, discover anomalies hidden in the data and be informative about where the power of the test comes from. In addition, the test can be performed on specific regions of observables of policy interest rather than the whole population. A feasible test statistic is made possible by substituting unknown objects with estimators, i.e. $\widehat{W}_N = N\widehat{T}_N'\widehat{\Omega}_N^{-1}\widehat{T}_N$ where $\widehat{T}_N = (\widehat{T}_N^1, \ldots, \widehat{T}_N^P)'$ and $\widehat{\Omega}$ is the consistent estimator of Ω to be given later,

$$\widehat{T}_N^p = N^{-1}\sum_{i=1}^{N} \widehat{t}_i^p \widehat{D}(X_i), \quad p = 1, \ldots, P_N \tag{3.10}$$

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, i.e. $\widehat{t}_i^p \equiv \mathbf{1}\left\{\min[c_L, \widehat{q}_X(p-1/P_N)] \le X_i < \max[\widehat{q}_X(p/P_N), c_U]\right\}$ where the quantile function is defined as $\widehat{q}_X(\tau) = \inf\left\{x : N^{-1}\sum_{i=1}^{N} \mathbf{1}(X_i > x) \ge \tau\right\}$.

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 as in the following Monte Carlo studies, though the testing results are found to be relatively robust in Section 4.2.

4 Asymptotic and Finite Sample Properties

Notations. 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 defined. 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.

4.1 Asymptotic Results

To conserve space, asymptotic Assumptions A.1-A.6 are presented in Appendix B. Assumption A.1-A.4 are regularity conditions frequently employed in nonparametric estimation and testing. Assumption A.1 formally states the data generating process (DGP) and requires the boundedness of conditional variances. 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 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.[17] Assumption A.4 provides additional smoothness 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

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

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. For even larger dimensions, a semiparametric version of the test in Section 5 is suggested.

Theorem 4.1 below gives the asymptotic null distribution of pth-quantile average difference. Only the sketch of the proof is provided below whereas the complete proof and supporting lemmas are left in the supplemental materials.

Theorem 4.1. *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 $\Omega_p = E(\xi_i^p \xi_i^{p'})$ and the influence function, ξ_i^p, is defined in Eq. (4.1) with $\epsilon_i = Y_i - C(X_i, V_i)$.*

$$\xi_i^p \equiv [t_i^p + E(t^p|V_i)]f^*(X_i, V_i) - t_i^p]\epsilon_i + E(t^p)h(V_i) \tag{4.1}$$

Theorem 4.1 says the quantile average difference in Eq. (3.10) 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.

Step 1: I show that the estimated quantile (and trimming) indicator can be replaced by its true counterpart plus remaining terms converging faster than $\sqrt{N}$. Consider the pth-quantile sample average difference

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

It is trivial to see that $I_3^p = 0$ and $I_2^p = o_p(N^{-1/2})$ under $\mathbb{H}_0$ where $D(X_i) = 0$ uniformly.

Step 2: To deal with I_1^p, further decompose it 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. (3.6) 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 $h(\cdot)$, i.e. $\widehat{a}(X_i) = \widetilde{a}(X_i) - \widehat{E}(\Delta_i|X_i)$, 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. 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 as $\widetilde{T}_N^p \equiv D_N^g + D_N^a + D_N^h$ where by definition

$$D_N^g = N^{-1} \sum_{i=1}^{N} t_i^p (\widehat{g}(X_i) - g(X_i)) \tag{4.2}$$

$$D_N^a = -N^{-1} \sum_{i=1}^{N} t_i^p (\widetilde{a}(X_i) - a(X_i)) \tag{4.3}$$

$$D_N^h = N^{-1} \sum_{i=1}^{N} t_i^p \widehat{E}(\Delta(V_i)|X_i) \tag{4.4}$$

Step 3: Finally, 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.[18] Therefore, I 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

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

estimator to ensure that the asymptotic biases vanish faster than $\sqrt{N}$ so that the vector of quantile mean differences will recenter at 0.

Theorem 4.2 and Corollary 4.2.1 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.

Theorem 4.2 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* $\widehat{W}^0_N \xrightarrow{\mathrm{D}} \chi^2_P$ *where* $\widehat{W}^0_N = N\widehat{T}'_N\Omega^{-1}\widehat{T}_N$*, with* $\widehat{T}_N$ *being the consistent estimator of* T_N *and* $\Omega = E(\xi_i\xi_i')$*, where* $\xi_i \equiv (\xi_i^1, \xi_i^2, \cdots, \xi_i^P)'$*.*

Corollary 4.2.1 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* $\widehat{W}_N \xrightarrow{\mathrm{D}} \chi^2_P$*, where* $\widehat{W}_N = N\widehat{T}'_N\widehat{\Omega}_N^{-1}\widehat{T}_N$*, with* $\widehat{T}_N$ *and* $\widehat{\Omega}_N$ *being the respective consistent estimators.*

The consistent estimator of covariance matrix $\widehat{\Omega}_N$ in Theorem 4.1 is therefore obtained by plugging in consistent estimators of all building blocks.

4.2 Finite Sample Simulation Results

In this section, I investigate the finite sample properties using simulations under two data generating processes (DGPs). In DGP 1, the simple additive model is tested against nonseparable models in polynomials. In DGP 2, 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, e.g. $P \in \{4, 6, 8\}$, 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 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 $\kappa_2 = 0.025$. Moderate number of replications, $N_{mc} = 250$ and sample size $N \in \{250, 500\}$ are considered for computational manageability.

DGP1 The data is generated from the following DGP,

$$Y = (X + \varepsilon) + \delta\,(X\varepsilon)^2 \tag{4.5}$$

where δ represents the level of nonseparability and if $\delta = 0$, the model becomes completely additive, i.e. $Y = X + \varepsilon$.

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

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 $\mathbb{H}_0$ in column 3-5 and power analysis under $\mathbb{H}_1$ in column 6-11. The test statistics are likely to be undersized for small samples 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, 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 P over $\{4, 6, 8\}$. As P varies, the rejection rates are relatively stable and robust. To sum up, the empirical sizes produced by our test statistics look close to what theory predicts even under small samples. 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.

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 $\kappa_2 = 0.025$.

DGP2 DGP 2 incorporates multiple unobservables (ε, η), featured in the test. For simplicity, assume the true model is like (4.6),

$$Y = X\eta + \varepsilon + \delta \exp(X\varepsilon) \tag{4.6}$$

where $\eta \sim U[0.5, 1, 5]$ and (X, V, ε) are generated in the same way as DGP 1. Note that when $\delta = 0$, then $Y = X\eta + \varepsilon$, is a nonseparable model. But our test has no power against structural separability in theory. Because it occurs that the ASF of this nonseparable function of multiple unobservable exactly equates that of some additive models. And it is verified by the empirical power results shown in column 3-5 in Table 2. Results are relatively stable across tuning parameters.

To make it more interesting, column 6-11 presents the simulation results under $\delta > 0$ where our test should have power against the additional nonseparable portion. Again, our theoretical conjecture is confirmed. Reasonable powers are obtained for large δ and N. In the end, DGP2 is presented here to remind researchers to be cautious when drawing conclusions that only nonseparability can be declared for sure given $\mathbb{H}_0$ is rejected.

Finally, I summarize key results from all above DGPs. 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 P. This property gives us 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 2: Empirical Power Results of DGP 2

		$\delta = 0$			$\delta = 0.5$			$\delta = 1$		
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 $\kappa_2 = 0.025$.

5 Extensions

In this section, two extensions in practice are considered. First, I consider the semiparametric test where a single-index restriction is placed on the high-dimensional conditioning set of variables. Second, I show how to adapt the test in nonseparable triangular simultaneous equations models with "generated" control variates.

5.1 Semiparametric Models

For semiparametric single-index models, the test can be performed in two stages with the finite-dimensional parameters estimated firstly by weighted semiparametric least square (WSLS) Ichimura (1993). Given a linear index structure, e.g. $I_0 \equiv X'\beta_0$, where β_0 is a conformable vector of finite-dimensional true parameters, redefine the model (2.1) as the semiparametric single index nonseparable model in Eq. (5.1).

$$Y = m(X'\beta_0, \varepsilon) \tag{5.1}$$

where it is assumed that there exists at least one continuous variable in X for identification purpose. The new set of hypotheses are stated below.

$$\mathbb{H}_0 : g(x'\beta_0) = a(x'\beta_0), \text{ for each } x \in \mathcal{X}; \quad \mathbb{H}_1 : \text{Otherwise}$$

where the ASFs are defined analogously.

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. 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 3. It has been well recognized that β_0 is only identified up to location and scale in semiparametric models. Here I normalize $\beta_{10} = 1$, where β_{10} is the coefficient of some continuous variable. Note that when X and ε are not correlated of any sort, model (5.1) 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 $E(U|X) = 0$. Given point identification, consistent estimator can be obtained by minimizing weighted sum of squared residuals

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

where the bias-reducing conditional expectation estimator is defined in Eq. (3.2). $\widehat{W}_i(\widehat{\beta}^0) = 1/\widehat{E}(\hat{\epsilon}^2|X_i'\widehat{\beta}^0, V_i)$ 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 are nested by nonseparable models (5.1), $\widehat{\beta}$ is also consistent even if the true model is additive separable. $\sqrt{N}$-consistency of $\widehat{\beta}$ follows directly from Ichimura and Lee (1991); Ichimura (1993); Klein and Shen (2015).

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 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 4.1, Theorem 4.2 and Corollary 4.2.1 would immediately apply by substituting the single index estimators, $\widehat{\beta}$. Loosely speaking, under mild conditions, it is true that $\widehat{W}_N(\widehat{\beta}) \xrightarrow{\text{D}} \widehat{W}_N(\beta_0)$ under $\mathbb{H}_0$.

5.2 "Generated" Control Variates

In many cases, control variables V are not directly observable. For example, the endogenous regressors are determined by first stage equations given in (5.2)

$$X_k = h_k(Z, \eta_k), \quad \forall k = 1, \cdots, d_X \tag{5.2}$$

where $h_k(z, \cdot)$ is an unknown strictly monotonic function, for each z and each $k = 1, 2, \cdots, d_X$ and Z is a vector of exogenous variables satisfying $Z \perp (\varepsilon, \eta)$. Let $\eta \equiv (\eta_1, \cdots, \eta_{d_X})'$. In the triangular system like above, the first stage disturbances suffice to serve as control variables. Fortunately, the asymptotic properties of the test statistic are robust to the problem of such "generated" regressors. Imbens and Newey (2009) consider the above assumptions in nonparametric nonseparable triangular simultaneous equations models and show that $X \perp \varepsilon | V$, where $V = [F_{X_1|Z}(X_1, Z), \cdots, F_{X_{d_X}|Z}(X_{d_X}, Z)] = [F_{\eta_1}(\eta_1), \cdots, F_{\eta_{d_X}}(\eta_{d_X})]$. 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 this situation, an additional step is needed to estimate $F_{X|Z}(X, Z)$ first by the recursive conditional expectation estimator defined in Eq. (5.3)

$$\widehat{V}_k = \widehat{F}_{X_k|Z}(x, z) = \frac{\sum_{i=1}^{N} K_h(Z_i - z)]\{\mathbf{1}[X_{ki} \le x] - \widehat{\delta}_i(z)\}}{\sum_{i=1}^{N} K_h(Z_i - z)}, \ k = 1, 2, \cdots, d_X \tag{5.3}$$

Fortunately, asymptotic results of the test statistic are not influenced by the first stage estimation. Again, under $\mathbb{H}_0$ and relatively standard conditions, it can be shown that $\widehat{W}_N(\widehat{V}) \xrightarrow{\mathrm{D}} \widehat{W}_N$ where $\widehat{W}_N(\widehat{V})$ denote the test statistic defined earlier with all V replaced by $\widehat{V}$. So the test proposed is robust with respect to issues of "generated" regressors. Such result is a direct implication Mammen et al. (2012) who study nonparametric regression with nonparametrically "generated" covariates. In the example of estimating ASFs, 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.

6 Hicks-neutral Productivity

Understanding how inputs are related to outputs is a fundamental issue in empirical industrial organization and other fields of economics (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. However, when it comes to empirical strategies, substantial parametric assumptions are always taken for granted. Among those, the most fundamental one is the Hicks-neutral productivity. 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. Nevertheless, when it comes to firm-level, Hicks-neutrality unavoidably puts substantial restrictions on how unobserved firm heterogeneity in productivity is modeled. Despite the need for accuracy of estimation, Hicks-neutrality remains one of the most stubborn assumptions even in many of nonparametric settings Gandhi et al. (2013). In what follows, I show that the proposed test can be adapted into testing Hicks-neutrality and provide empirical evidences in support of the recent interest that calls for fully flexible estimation of production functions (Kasahara et al., 2015; Li and Sasaki, 2017; Hu et al., 2017, etc).

To be concrete, I define a firm's production in year t to be Hicks-neutral if its production function is multiplicative separable or log additive separable in unobservables[19]

$$Y = F_t(K, L, \omega, \varepsilon) = F_t^1(K, L)A_t(\omega, \varepsilon) \text{ or } y = f_t^1(K, L) + a_t(\omega, \varepsilon) \tag{6.1}$$

where following the notation of the literature, F_t denotes the nonparametric firm production function at year t.[20] 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. The "simultaneity bias" or endogeneity of input choices usually arises because the amount of variable inputs like labor is determined after private productivity shocks ω is realized.

The Hicks-neutral functional form has at least three implications. First, such restriction dictates that input substitution patterns be free of firm-specific unobserved technological shocks, as a result, eliminating an important channel of firm-level heterogeneity. For example, for firms that use the same amount of inputs, it implies that output-input elasticities are identical and degenerate and so are other firm-specific structural parameters such as elasticity of substitution, return-to-scale and even markups under imperfect competition, etc.[21] So commonly used labor or capital-augmented production functions are being ruled out.

Second, the wrongly imposed Hicks-neutrality would make most of the commonly employed identification strategies invalid. Those identification methods include, but are not

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

[20]It is usually assumed that both F_t^1 and A_t map into strictly positive sets in this context.

[21]I define the output-input elasticities as $\beta_{L,t} \equiv \frac{\partial Y_t}{\partial L_t}\frac{L_t}{Y_t} = \frac{\partial y_t}{\partial l_t}$ and $\beta_{K,t} \equiv \frac{\partial Y_t}{\partial K_t}\frac{K_t}{Y_t} = \frac{\partial y_t}{\partial k_t}$, respectively for labor and capital.

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. Since all the above methods begins with converting the productivity and error terms (ω, ε) in an additive way via log transformation, this step would unavoidably lead to inconsistent estimators of structural parameters provided that Hicks-neutrality does not hold.

Furthermore, the non-Hicks-neutrality poses serious threat to the usual interpretation of the "Solow" residual as a single measure of total factor productivity (TPF). On the one hand, TPF is not well-defined if productivity shocks come in several different dimensions, like labor-augmented, capital-augmented, etc. On the other hand, even for the single-dimensional shock, misleading distribution of productivity is likely to be produced if simply calculated as the "Solow" residuals. In supplemental materials, the results from a simple simulation indicate that serious distortions of productivity distributions do generally occur. Given all above, a test of Hicks-neutrality may have its own empirical appeal.

According to the log production function in Eq. (6.1), testing Hicks-neutrality amounts to testing the additive separability in unobservables so that our proposed methods directly apply with additional assumptions. 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 Functional forms. $Y_t = F_t(K_t, L_t, \omega_t, \varepsilon_t)$, where $\omega_t \in \mathbb{R}, \varepsilon_t \in \mathbb{R}^\infty$.

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 $(l, k) \in \mathcal{L} \times \mathcal{K}$.

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.[22] Assumption F.3 is standard in the proxy variable literature.[23] To single out the testing problem, I consider the static model and perform the test year by year, in order to separate out the sample selection and other dynamic issues.

Proposition 6.1 below 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 6.1 is given in Appendix A. The "generated" control variable can be estimated by any nonparametric estimators in principle. Here again I consider a bias-reduced local constant estimator similar to Eq. (5.3).[24]

Proposition 6.1 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.

7 Empirical Testing Results

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 from net sales deflated by industry-

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

[23] 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.

[24] The analytic variance of the test statistics under $\mathbb{H}_0$ is robust to the "generated" control variates problem as explained in Section 5.

level price index for shipments.[25] 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 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.[26] 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.[27] Table 3 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.

Table 3: 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 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.

$$\mathbb{H}_0 : F_t(L, K, \omega, \varepsilon) = F_t^1(L, K) + A_t(\omega, \varepsilon); \quad \mathbb{H}_1 : \text{Otherwise}$$

[25] 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.

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

[27] 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 4 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.

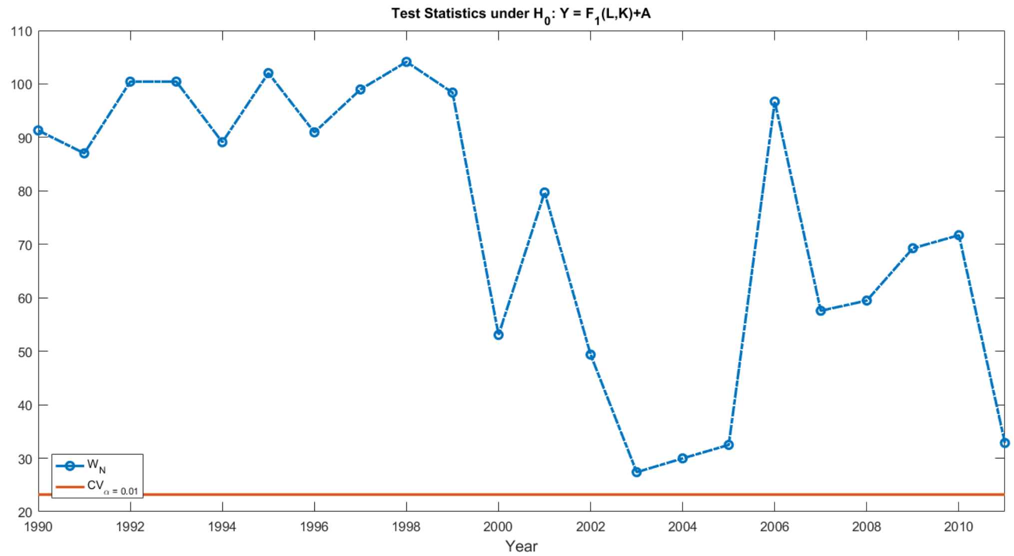
Table 4: Empirical Testing Results by Year 1990-2011

		(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 $\kappa_2 = 0.025$.

In Figure 2, I present the testing results for the Hicks-neutrality for each year for

Figure 1: Test Statistics by Year of Nonparametric Production Functions



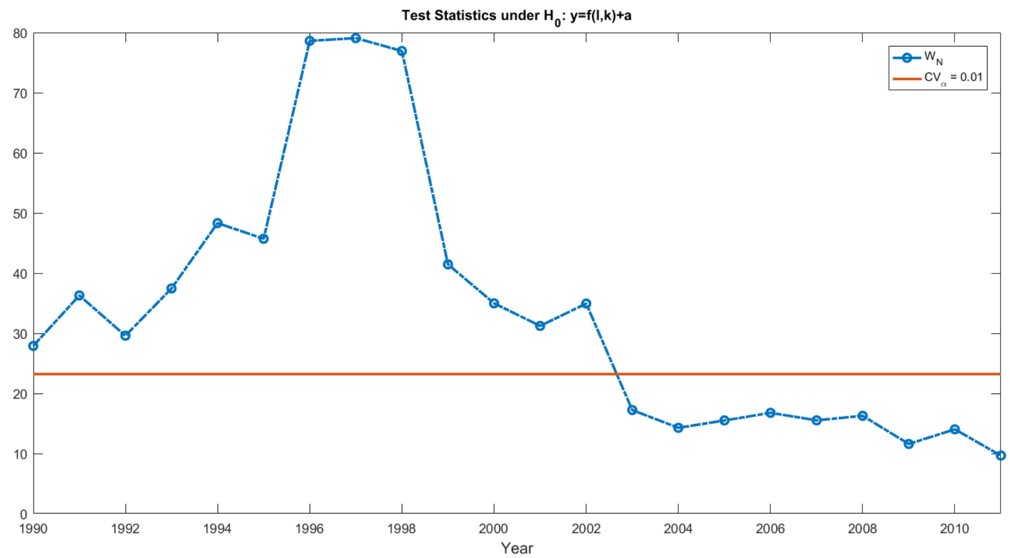
Note: The horizontal line stands for critical values at 1% significant level. Circled markers are test statistics at each year.

the specification (2) in Table 4. As shown before, it is equivalent to testing the additive separability of the log-transformed production function,

$$\mathbb{H}_0 : f_t(L, K, \omega, \varepsilon) = f_t^1(L, K) + a_t(\omega, \varepsilon); \quad \mathbb{H}_1 : \text{Otherwise}$$

The results show that non-Hicks neutral production occured during 1990-2002 and thereafter became Hicks-neutral. 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 differ greatly 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. Future empirical evidences are needed to confirm such conjecture.

Figure 2: Test Statistics by Year of Log Transformed Models



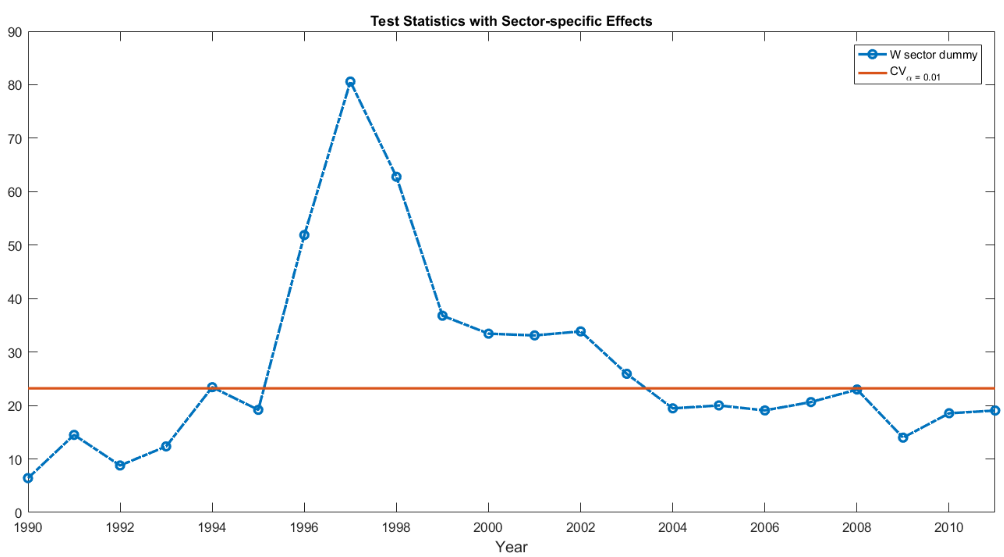
Note: The horizontal line stands for critical values at 1% significant level. Circled markers are test statistics at each year.

However, it is also likely 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. For example, it would make no sense to expect that a labor union strike to affect machinery and food product equally as their input substitution pattens are quite different. When there are sufficient amount of observations in each sector, one might mitigate this problem by placing sector dummy variables to nonparametrically 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, some of which only have a few observations in some years. Therefore, the variances of the test statistic could be extremely large and renders the test of very low power. The limitation of data forces us to compromise on the full nonparametric sector-specific effects. Instead, I test with linear sector dummies.

$$\mathbb{H}_0 : f_t(L, K, \omega, \varepsilon) + S_t = f_t^1(L, K) + S_t + a_t(\omega, \varepsilon); \;\; \mathbb{H}_1 : \text{Otherwise}$$

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 substitution patterns. The results are presented in the specification (3) in Table 4 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.

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.

The empirical results show that there are indeed periods of non-Hicks-neutral productions in the U.S. manufacturing industries, which may coincide with a period of rapid adoption of computer technology 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.

8 Conclusions

In this paper, I propose an easy-to-implement test for structural separability of fully nonparametric models, explicitly allowing for unrestricted 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 make a compromise on the equivalence of testing hypotheses due to the non-identification problem. 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 both Monte Carlo studies and real data application.

To apply the proposed test, this paper also provides empirical methods to test for the Hicks-neutral technological shocks in firm-level production function estimation. As argued, not only does such assumption significantly restrict input substitution patterns but also make many identifying strategies questionable once it fails to hold. With a firm-level dataset on U.S. manufacturing industry over 22 years, the test has successfully detected a period of strong non-Hicks-neutral productions in the late 90s. It may be interesting to investigate its correlation with the mass adoption of computer technology in future works.

Appendix A Proof of Identification Results

Proof of Proposition 2.3. a) is obvious from Proposition 2.2 once $c_h = 0$ by normalization since $a(x) = m_1(x) = E(Y - h(V)|X = x)$. To show b), given $X = x$,

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

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

Proof of Proposition 6.1. In the first step, I 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 Assumption F.2 that $(K, L) \perp (\omega, \varepsilon)|V$. □

Appendix B Asymptotic Assumptions

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). $\mathrm{Var}(Y|U) < \infty$.

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

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, Yuan Liao, Yonghong An, Yuya Sasaki, Xiye Yang, John Landon-Lane and others. All errors are purely the author's.

References

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

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.

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.

Amit Gandhi, Salvador Navarro, and David A Rivers. On the identification of production functions: How heterogeneous is productivity? 2013.

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.

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.

Yingyao Hu, Guofang Huang, et al. Estimating production functions with robustness against errors in the proxy variables. 2017.

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.

Hiroyuki Kasahara, Paul Schrimpf, and Michio Suzuki. Identification and estimation of production function with unobserved heterogeneity. Technical report, Working Paper, 2015.

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.

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.

Tong Li and Yuya Sasaki. Constructive identification of heterogeneous elasticities in the cobb-douglas production function. *arXiv preprint arXiv:1711.10031*, 2017.

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, Oliver Linton, J Nielsen, et al. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, 27(5): 1443–1490, 1999.

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

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.

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.

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

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.